基于深度学习的多肽预测方法研究

梁 潇[1,2],吴 昊[1],刘全中[1,2*]
(1. 西北农林科技大学 信息工程学院,陕西 杨凌 712100;
2. 陕西省农业信息感知与智能服务重点实验室,陕西 杨凌 712100)

摘 要:多肽,也可简称为肽,是 α-氨基酸通过肽键连接在一起而形成的一类化合物,也是蛋白质水解的产物。它对人体的生长、发育、代谢有着重要的影响,部分多肽具有抗癌、抗菌、抗病毒、穿透细胞等特性,对于相应疾病的治疗具有重大意义。因此研究识别具有治疗特性的多肽方法至关重要,然而传统生物实验方法鉴定多肽耗时且昂贵,不适合处理高通量的序列数据。现有的基于机器学习的预测模型虽然大大提高了多肽的识别效率,但存在识别性能不足,泛化能力不够,以及一种模型只能有效识别特定的一种多肽等问题。针对以上问题,该文提出了一种通用深度学习模型 DeepPEPred,该模型能有效预测多种不同的肽。在抗癌肽、抗菌肽、细胞穿透肽和结合肽四种不同肽数据集上进行十折交叉验证和独立测试,实验结果表明:与目前最新的方法 PEPred-Suit 相比,DeepPEPred 在抗癌肽数据集上准确度提升了 29.6%,MCC 提升了 59.7%;在抗菌肽、细胞穿透肽和结合肽三种数据集上准确度均提升了 1.2%,MCC 分别提升了 2.3%、2.5% 和 2.4%,AUC 分别提升了 0.8%、0.3% 和 1.2%。

关键词:多肽;深度学习;预测模型;识别方法;特征提取
中图分类号:TP181 **文献标识码**:A **文章编号**:1673-629X(2021)07-0140-07
doi:10.3969/j.issn.1673-629X.2021.07.023

Deep PEPred:A Deep Learning-based Approach for Predicting Peptides

LIANG Xiao[1,2],WU Hao[1],LIU Quan-zhong[1,2*]
(1. School of Information Engineering,Northwest A&F University,Yangling 712100,China;
2. Shaanxi Key Laboratory of Agricultural Information Perception and Intelligent Service,Yangling 712100,China)

Abstract:Polypeptides,also known as peptides,are a type of compounds that are formed by linking α-amino acids together via peptide bonds,which are also the products of protein hydrolysis. It has an important influence on the growth,development and metabolism of human body. Some polypeptides have the properties of anticancer,antibacterial,antiviral and penetrating cells,so that they are quite important for the treatment of corresponding diseases. Therefore,it is vital to identify peptides with therapeutic properties. However,the experimental methods are time-consuming and expensive,and are not practically suitable for high-throughput sequence data. Although the existing machine learning-based models greatly improve the efficiency of peptide recognition,they are still limited in respect of performance and generalization ability. Moreover,most models are peptide-specific models that can only effectively identify a specific therapeutic peptide. To address these problems,we propose DeepPEPred (deep learning based method for PEptide prediction),a general deep learning-based computational model for peptide prediction. That is,it can effectively predict a variety of different peptides. Ten-fold cross-validation test and independent test were conducted on anticancer peptides (ACPs),anti-bacterial peptides (ABPs),cell penetrating peptides (CPPs) and surface-binding peptides (SBPs) datasets. Compared with PEPred-Suit,the latest predictive method of polypeptides,for ACPs,DeepPEPred improved the accuracy and MCC by 29.6% and 59.7%,respectively. For ABPs,CPPs and SBPs,DeepPEPred improved the accuracy by 1.2%,and MCC by 2.3%,2.5% and 2.4%,respectively,and AUC by 0.8%,0.3% and 1.2%,respectively.

Key words:polypeptides;deep learning;prediction model;recognition method;feature extraction

收稿日期:2020-08-17 **修回日期**:2020-12-23
基金项目:国家自然科学基金面上项目(61972322);教育部人文社科交叉项目(18YJCZH190);基本科研业务费前沿与交叉科学研究项目(2452019180);中央高校基本科研业务费(2452017342);博士科研启动经费(2452017019)
作者简介:梁 潇(1996-),女,硕士研究生,研究方向为数据挖掘、计算生物学;吴 昊,副教授,博士,CCF 会员(73301M),研究方向为数据挖掘、计算生物学;刘全中,副教授,博士,通信作者,研究方向为数据挖掘、计算生物学。

0　引　言

生物体内广泛分布着上万种多肽,近年来,随着科学研究的发展和对生命活动规律的深入探索,越来越多的功能性多肽分子被不断发现,部分多肽具有抗癌、抗菌、抗炎、抗病毒、穿透细胞等特性,这些特性为疾病治疗提供了重要依据[1]。

抗癌肽(anticancer peptides,ACPs)能破坏肿瘤细胞膜结构,对癌细胞增殖和迁移具有抑制作用,而对正常的体细胞基本无损伤,因此抗癌肽检测有助于抗肿瘤药物的研究[1];抗菌肽(anti-bacterial peptides,ABPs)对部分细菌、真菌、病毒有杀伤作用,其潜在的价值也受到医学界的广泛关注[2];细胞穿透肽(cell penetrating peptides,CPPs)被广泛用作药物进入细胞的运输载体[3];结合肽(surface-binding peptides,SBPs)有助于在噬菌体展示实验中建立高效的 ELISA(enzyme linked immunosrbent assay)系统[4]。

具有治疗特性的多肽目前已经越来越广泛地应用于临床诊断和治疗中,因此识别这些多肽对于发现新的、高效的疾病治疗方法具有重要的现实意义[2]。传统的生物实验方法识别多肽耗时、耗力且成本高,随着高通量测序技术的发展和测序成本的持续降低,研究界和医学界不断产生海量的测序序列,然而传统方法从高通量序列中识别多肽效率低下。为了提高多肽的识别效率,基于机器学习的多肽识别方法越来越受到研究界的青睐[5]。近年来,研究界已提出了许多基于机器学习的治疗肽的预测模型,根据其使用算法进行分类,分为基于传统的机器学习肽预测模型与基于深度学习的肽预测模型。

基于传统的机器学习肽预测模型主要使用不同的序列特征把肽序列表示为特征向量,构造二分类样本集,使用不同的分类模型进行训练,然后预测新的肽序列。主要工作如下:2007 年 7 月,Lata 等人利用抗菌肽中 N 端和 C 端残基的特异性分别建立了基于神经网络、QM(quantitative matrices)和支持向量机的 ABP 预测模型[5];2017 年 5 月,Wei 等人整合了基于序列的特征描述符 PC-PseAAC(parallel correlation pseudo-amino-acid composition)、SC-PseAAC(series correlation pseudo-amino-acid composition)、ASDC(adaptive skip dipeptide composition)、PPs(physicochemical properties),构建了基于随机森林算法的两层 CPP 预测框架 CPPred-RF[6];2017 年 7 月,Li 等人使用 OAAC(optimized amino acid composition)和 ODPC(optimized dipeptide composition)两种特征开发了基于支持向量机的 SBP 预测器 PSBinder,它可以快速有效地排除假阳性肽,更准确地获得 SBP[7];2018 年 6 月,Wei 等人提出了一个基于支持向量机的 ACP 预测器 ACPred-FL[8],使用了 BPF(binary profile features)、GDC(G-gap dipeptide composition)、OPF(overlapping property features)、CTD(composition-transition-distribution)4 种序列特征表示样本,通过最大相关-最小冗余和顺序前向搜索特征选择方法剔除冗余特征,提高了预测器的预测性能。以上预测方法都是针对识别特定的肽而构造的模型,2019 年 4 月,Wei 等人提出了基于随机森林的多肽预测模型 PEPred-Suit,该模型引入了一种自适应特征表示策略,可以学习不同肽类型的最具代表性的特征,能有效识别多种不同类型肽[9]。

深度学习主要使用卷积神经网络和循环神经网络自动抽取出抽象特征,其中循环神经网络主要用于处理文本和序列数据。肽是一种序列数据,因此循环神经网络更适合肽的预测研究。针对基于深度学习的肽预测模型,2019 年 9 月,Yi 等人使用两种序列特征 K-mer 稀疏矩阵和 BPF(binary profile features),构建了基于长期短期记忆 LSTM(long short-term memory)循环神经网络的 ACP 预测模型 ACP-DL,实现了一个 DeepLSTM 模型来自动学习如何识别抗癌肽和非抗癌肽。在基准数据集五折交叉验证实验结果表明,ACP-DL 具有较高的识别性能[10]。

已有的基于机器学习肽预测方法促进了肽的研究,但分类器的识别性能仍有待提高,而且除了 PEPred-Suit 模型外,其他模型都只能识别某一种特定肽。针对以上问题,该文提出了一种通用的基于 GRU 循环神经网络的多肽预测模型 DeepPEPred,能有效识别多种类型的肽。DeepPEPred 用如下四种特征作为输入序列的编码:氨基酸组成(amino acid composition,AAC)、K-spaced 氨基酸对的组成(composition of k-spaced amino acid pairs,CKSAAP)、构成/变迁/分布(composition/transition/distribution,CTD)、伪氨基酸组成(pseudo-amino acid composition,PAAC)能够有效预测不同的肽段,其中 AAC 在 ACPred-FL 模型被使用预测抗癌肽,CTD 在 PEPred-Suit 模型被使用预测各种类型的肽。通过初步的实验验证:这四种特征使得 DeepPEPred 模型能够获得较好的总体性能。为了验证 DeepPEPred 的性能,该研究在抗癌肽、抗菌肽、细胞穿透肽和结合肽四种不同肽数据集上进行实验。经过十折交叉验证和独立测试结果表明,与现有的肽预测模型相比,DeepPEPred 模型具有更强的识别性能。

1　数据集

该文旨在构建一个通用的深度学习模型预测具有不同治疗特性的肽,使用 ACP、ABP、CPP 和 SBP 四种

肽数据集评估提出的模型,每种肽数据集包括一个训练集和一个独立测试集,训练集和独立测试集都由正例样本和负例样本组成,正例样本是经过实验验证的治疗性多肽(如抗癌活性),负例样本通常是没有相关特性(如非抗癌活性)或随机序列的多肽[9]。

该研究使用的 ACP 数据包括文献[9-13]提供的数据集和数据库 CancerPPD[14] 中最新的 ACP 数据,为了避免整合后序列中含有重复序列,该研究使用 CD-HIT 软件[15] 去除同源性超过 90% 的序列。最后得到的 ACP 训练集中包括 422 个经实验验证的 ACP 序列以及 1 688 个非 ACP 序列;ACP 独立测试集中包括 97 个经实验验证的 ACP 序列以及 97 个非 ACP 序列。该文使用了 Lata 等人[5] 提供的 ABP 数据集、Wei 等人[6] 提供的 CPP 数据集以及 Li 等人[7] 提供的 SBP 数据集。四种肽数据集的详细信息如表 1 所示。

表 1 四种肽数据集

数据集	训练集样本数量		独立测试集样本数量	
	正例	负例	正例	负例
ACP	422	1 688	97	97
ABP	800	800	199	199
CPP	370	370	92	92
SBP	80	80	24	24

2 特征提取

该研究通过 iLearn[16] 选取了四种特征表示肽序列,分别是:氨基酸组成(AAC)、K-spaced 氨基酸对的组成(CKSAAP)、构成/变迁/分布(CTD)、伪氨基酸组成(PAAC)。

2.1 氨基酸组成(AAC)

氨基酸组成(AAC)[17] 是计算肽序列中每种氨基酸的出现频率,AAC 特征编码的维度为 20,序列中每种氨基酸出现的频率可由公式(1)计算:

$$F = \frac{R(i)}{L}, i \in \{1,2,\cdots,20\} \tag{1}$$

其中, $R(i)$ 是肽序列中名称为 i 的氨基酸出现的次数, L 是肽序列的长度。最终可以得到 20 种氨基酸在肽序列中的出现频率。

2.2 K-spaced 氨基酸对的组成(CKSAAP)

K-spaced 氨基酸对的组成(CKSAAP)特征编码是计算任意 k 个残基($k=0,1,2\cdots$)分隔的氨基酸对的频率[17]。以 $k=0$ 为例,有 400 个间隔为 0 的残基对(AA,AC,AD,…,YY),那么特征向量可以定义为:$(\frac{N(AA)}{Ntotal},\frac{N(AC)}{Ntotal},\frac{N(AD)}{Ntotal},\cdots,\frac{N(YY)}{Ntotal})400$,其中每个描述符的值表示对应的残基对在蛋白质或肽序列中的组成[17]。例如,如果残基对 AA 在蛋白质中出现了 n 次,那么残基对 AA 的组成就等于 n 除以蛋白质中间隔为 0 的残基对总数(Ntotal)。对于 $k=0,1,2\cdots$ 时,长度为 L 的肽序列 Ntotal 值分别为 $L-1$、$L-2$、$L-3$ …。因此,CKSAAP 把肽序列编码为 $20\times20\times(k+1)$ 维向量。该研究经过实验测试所有可能的 k 值,当 $k=4$ 时模型预测性能最优。

2.3 构成/过渡/分布(CTD)

CTD 使用组成(C)、过渡(T)和分布(D)三个描述符描述蛋白质序列中[18] 的每个基团中各性质的氨基酸分布,CTD 采用七种物理化学性质表示蛋白质或肽序列,它们包括疏水性、标准化范德华体积、极性、极化度、电荷、二级结构和溶剂可及性,ilearn 包[16] 中将疏水性又分为七个不同性质,加上其他六种性质,共有 13 种性质。基于主要的氨基酸指数,针对每一种性质,将 20 种氨基酸分为三类。本研究只使用描述符 D 来编码肽序列,D 统计三类氨基酸中每类氨基酸含量为 0%,25%,50%,75%,100% 时相对于整条肽序列的分布情况,即每类有五个描述符值,因此每种性质使用 $3\times5=15$ 个描述符表示。因此,CTD 将一个肽序列编码成一个由 $13\times15=195$ 个描述符值组成的向量。

2.4 伪氨基酸组成(PAAC)

传统的氨基酸组成只考虑蛋白质序列中 20 个氨基酸出现的频率,这会丢失蛋白质链的序列信息。PAAC 将 20 个氨基酸的序列顺序信息和频率整合在一起进行编码[19]。一个蛋白质序列编码成一个 $20+A$ 维向量,向量的前 20 个分量表示 20 个氨基酸的出现频率,最后的 A 个分量表示序列顺序信息。PAAC 被证明是一种有效的特征编码方案,并被广泛应用于蛋白质序列或者肽序列相关领域的研究[20]。输入肽序列的 PAAC 计算由 ilearn 包提供。经实验验证,当 $A=4$ 时,模型预测性能最优,因此 PAAC 将一个肽序列编码成一个 24 维特征向量。

3 特征标准化

不同特征向量往往具有不同的量度,这将影响到模型预测性能,因此需要对原始特征组合进行标准化使得每个特征处于同一数量级,有利于预测模型的建立[21]。

该研究使用的 Z-score 方法是基于原始特征的均值(mean)和标准差(standard deviation)进行数据的标准化,该方法适用于数据属性值的最大值和最小值未知的情况,或有超出取值范围的离群数据的情况[21]。Z-score 标准化可由公式(2)计算:

$$Z = \frac{X-\mu}{\sigma} \tag{2}$$

其中, Z 为标准化后的变量值, X 为实际变量值, μ 为

各变量(特征)的算术平均值(数学期望),σ 为标准差。

4　基于深度学习的肽识别方法

深度学习(deep learning,DL)作为机器学习的新兴技术[22],近年来已被广泛应用于生物信息学中[23]。深度学习模型包括卷积神经网络模型、堆栈自编码网络模型、长短期记忆网络模型(long short-term memory,LSTM)[24-25]等。

该研究选用门递归单元(gated recurrent unit,GRU),GRU 是 LSTM 的一个简化的变体,旨在缓解梯度消失问题。GRU 基本可以达到与 LSTM 网络相同的效果,并且 GRU 的参数更少,能减少过拟合的风险[26]。GRU 将遗忘门和输入门组合成一个更新门,并有一个附加的重置门[27],更新门控制上一时刻的时间步的记忆被带入到当前时刻中的量,更新门的值越大说明上一时刻的时间步的记忆带入越多,重置门控制上一时刻有多少记忆被写入到当前的候选集上,重置门越小,上一时刻的记忆被写入的越少[27]。

4.1　模型的整体框架

提出的基于深度学习的多肽识框架如图 1 所示,主要包含以下几个步骤。

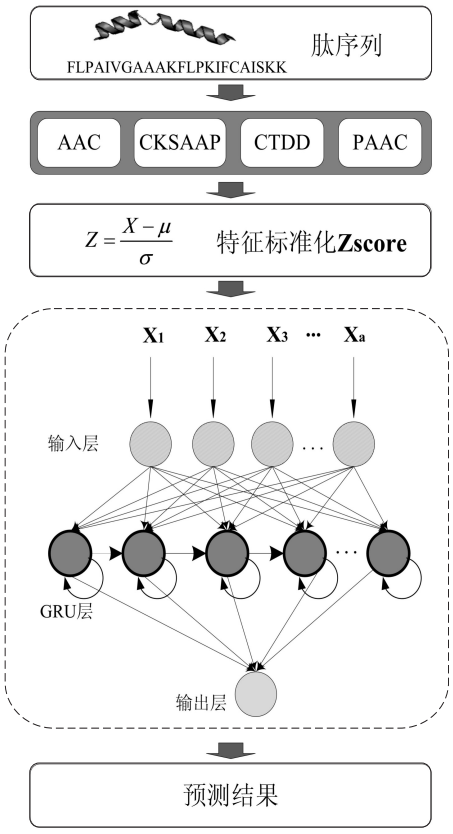

图 1　DeepPEPred 预测方法流程

步骤一:肽序列数据集构造。该研究收集了 ACP、ABP、CPP 和 SBP 四种肽数据集,四种数据集分别包含一个训练集和一个独立测试集,每种数据的训练集和独立测试集见表 1。

步骤二:肽序列样本集表示。该研究通过对多种肽序列特征进行性能评估,筛选出四种对于 ACP、ABP、CPP、SBP 序列有较强识别能力的特征,四种特征分别是 AAC、CKSAAP、CTDD、PAAC,它们编码维数分别是 20、2 000、195、24,每一个肽序列样本被编码的维度为 2 239,得到四种肽序列的二分类样本集。

步骤三:深度学习模型构建。输入层特征维度为 2 239 个,将输入层神经元输入隐藏层,隐藏层的第一层为 GRU 层,输出维度为 59,GRU 层后面增加一个 Dropout 层,设置为 0.465,防止模型过拟合;输出层空间维度为 1,使用 sigmoid 激活函数。在模型训练过程中,使用 early-stop 早停机制,防止模型过拟合;损失函数使用交叉熵损失函数,优化器使用 Adam,迭代次数(epoch)为 100 次。

步骤四:模型训练。该研究先使用 ACP 数据集训练一个初步的预测模型,由于 ACP 数据集中负例样本数是正例样本数的四倍,样本集严重不平衡,将影响模型的性能。该研究借鉴 BootStrapping[28] 方法来解决数据集中正负例样本不平衡问题,BootStrapping 方法是指对数据集进行有放回的抽样,将每次抽取的数据作为一个新样本,重复多次,形成多个新样本。该研究对负例样本集采取不放回抽样方法,该策略的示意图如图 2 所示。假设 P 和 N 分别表示正例样本集(ACP 序列)和负例样本集(非 ACP 序列),TP 和 TN 表示正例样本和负例样本的数量,以大小为 TP 的窗口循环遍历负例样本集,循环 n = TN/TP 次,每次循环抽取的 TP 个负例样本作为一个负子集,与正例样本集结合生成一个正负例数目相同训练集,并用这个训练集进行模型训练,保留每次循环训练的模型,最终预测结果取 n 次模型预测结果的均值。

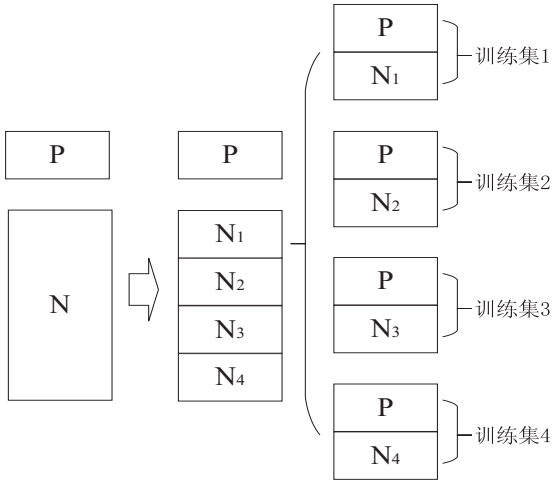

图 2　ACP 数据集划分图

步骤五:模型优化。首先在 ACP 数据集上经过贝

叶斯参数调优[29]进行参数寻优,得到一个最优参数的框架,然后用最优参数框架训练 ABP、CPP、SBP 数据集。贝叶斯优化方法首先采用高斯过程不断地更新目标函数的后验分布,然后在预先设置的参数范围内自动搜索最好的参数。在参数优化时,设置 GRU 层输出维度的初始范围为[8,128],优化后的最优值为 59;设置 Dropout 的初始范围为[0.1,0.6],优化后的最优值为 0.465。经过上述操作确定了最优参数,并构建了一个适用于四种治疗肽的最优模型。

步骤六:模型评估。该研究使用十折交叉验证和独立测试方法对模型进行评估,并与现有模型进行预测性能比较。

4.2 评价指标

为了评估 DeepPEPred 模型的预测性能,该研究使用了五种常用指标来评价模型的性能,包括 AUC(area under the ROC curve)值、准确度(accuracy,Acc)、特异性(specificity,Sp)、敏感性(sensitivity,Sn) 和马修斯相关系数(Matthews correlation coefficient,MCC)。其中 AUC 表示 ROC(receiver operating characteristic)曲线下的面积,ROC 曲线是指按顺序逐个对样本进行预测,每次计算出真阳性率(TPR)与假阳性率(FPR)分别以它们作为纵、横坐标进行绘制而生成的曲线。较大的 AUC 值表示该模型实现了更好和更强大的预测性能。这五种评价指标的定义如下:

$$Acc = \frac{TP + TN}{TP + FP + TN + FN} \times 100\% \tag{3}$$

$$Sp = \frac{TN}{TN + FP} \times 100\% \tag{4}$$

$$Sn = \frac{TP}{TP + TN} \times 100\% \tag{5}$$

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TN + FN) \times (TP + FP) \times (TN + FP) \times (TP + FN)}} \tag{6}$$

其中,TP、TN、FP 和 FN 分别表示真阳性、真阴性、假阳性和假阴性的样本数量。

5 实验结果

该研究对比的 ACP 识别方法在同样的数据集上采用独立测试,其他三种肽的识别方法在相同的数据集上采用十折交叉验证方法,为了公平比较,该研究分别采用同样的策略。

5.1 十折交叉验证

图 3 表示 DeepPEPred 与现有模型在 ABP、CPP、SBP 三种肽数据集上十折交叉验证结果的比较。由于现有的模型仅仅通过 AUC 值进行评价,为了公平对比,该研究也仅仅提供了每种数据集的 AUC 值。

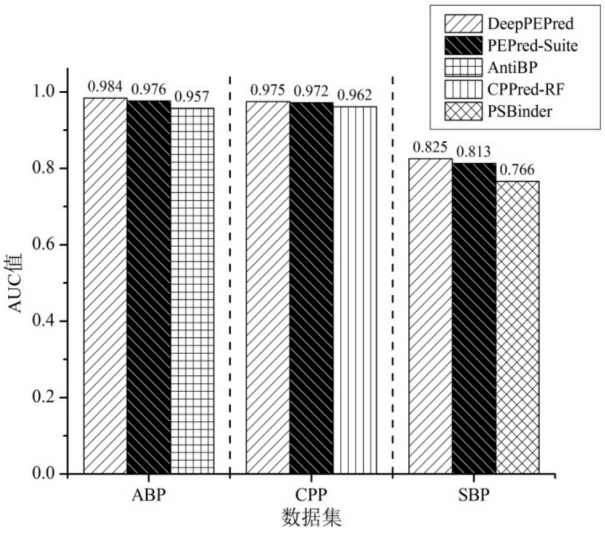

图 3 DeepPEPred 和现有预测器在 ABP、CPP 和 SBP 数据集上的性能对比

从图 3 的结果可知:在相同数据集上 DeepPEPred 预测模型在 AUC 方面取得了比其他预测方法更好的性能。在三个数据集(ABP、CPP 和 SBP)上比目前最新模型 PEPred-Suite 的 AUC 值分别高 0.8%、0.3% 和 1.2%,比其他预测同类型肽模型(AntiBP、CPPred-RF 和 PSBinder)的 AUC 值分别高出 2.7%、1.3% 和 5.9%。

在表 2 分别给出了 DeepPEPred 和 PEPred-Suite 在三个数据集上十折交叉验证的其余指标(Acc、Sn、Sp、MCC)的值,在 ABP 数据集上实验结果表明:DeepPEPred 比 PEPred-Suite 的 MCC 和 Acc 分别高出 2.3% 和 1.2%;在 CPP 数据集上实验结果表明:DeepPEPred 比 PEPred-Suite 模型的 MCC 和 Acc 分别高出 2.5% 和 1.2%;在 SBP 数据集上实验结果表明:DeepPEPred 比 PEPred-Suite 模型的 MCC 和 Acc 分别高出 2.4% 和 1.2%。

表 2 ABP、CPP、SBP 数据集上十折交叉验证结果

数据集	模型	MCC	Acc	Sn	Sp
ABP	DeepPEPred	0.890	0.945	0.945	0.945
	PEPred-Suite	0.867	0.933	0.912	0.954
CPP	DeepPEPred	0.849	0.924	0.900	0.949
	PEPred-Suite	0.824	0.912	0.903	0.922
SBP	DeepPEPred	0.462	0.731	0.725	0.738
	PEPred-Suite	0.438	0.719	0.725	0.713

5.2 独立测试

为了验证 DeepPEPred 的泛化能力,该研究在 ABP、CPP 和 SBP 数据集上进行了独立测试,并与现有方法进行了性能比较,结果如图 4 所示。从图 4 中结果可知:在三个数据集上与 PEPred-Suite 预测模型相比,AUC 值分别提升了 0.7%、1.5% 和 1.0%。在

ABP 数据集上,DeepPEPred 与同类型肽预测模型 AntiBP 相比 AUC 值分别提升了 0.7%;在 CPP 数据集上,DeepPEPred 与在同类型肽预测模型 CPPred-RF 相比 AUC 值提升了 2.6%;在 SBP 数据集上,DeepPEPred 与 PSBinder 的 AUC 值相等。

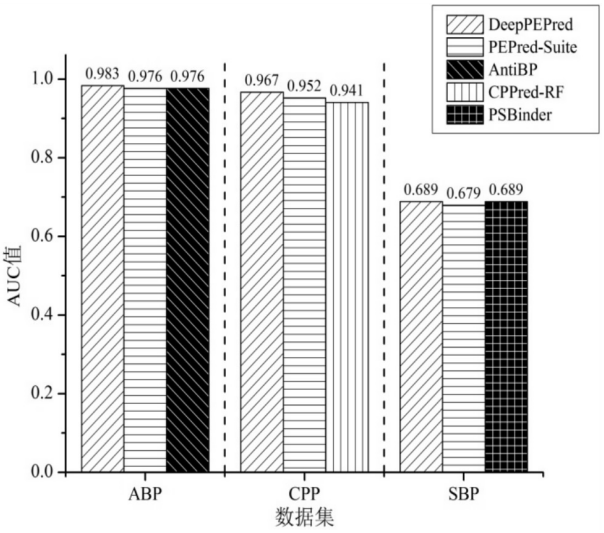

图 4　DeepPEPred 和现有预测器独立测试的性能对比

DeepPEPred 模型在 ABP、CPP 和 SBP 数据集上独立测试的 MCC、Acc、Sn、Sp 评价结果如表 3 所示。即使 DeepPEPred 是基于 ACP 数据集构建及调优的,从图 4 和表 3 结果可知,DeepPEPred 对于 ABP、CPP 和 SBP 三种肽预测也是有效的,说明 DeepPEPred 具有较强的迁移能力。

表 3　ABP、CPP、SBP 独立测试详细结果

数据集	MCC	Acc	Sn	Sp
ABP	0.889	0.945	0.940	0.950
CPP	0.805	0.902	0.891	0.913
SBP	0.293	0.646	0.625	0.667

由于 ACP 数据集正负例样本不平衡,该研究首先将负例样本分成四份,每份与正例样本相结合生成四个正负均衡的训练集;然后对模型进行四次训练,每次训练得到的模型进行独立测试,测试结果如表 4 所示,最终结果为四次结果的均值,其 AUC、MCC 和 Acc 最终值分别为 0.875、0.631 和 0.811。

表 4　ACP 数据集独立测试结果

	AUC	MCC	Acc	Sn	Sp
1	0.813	0.554	0.768	0.639	0.897
2	0.889	0.603	0.794	0.680	0.907
3	0.891	0.675	0.835	0.773	0.897
4	0.908	0.693	0.845	0.804	0.887
平均值	0.875	0.631	0.811	0.724	0.897

为了进一步验证 DeepPEPred 模型预测 ACP 的性能,该研究比较了 DeepPEPred 与 PEPred-Suite、ACPred[30] 两个 ACP 预测模型,独立测试结果如表 5 所示。需要强调的是,PEPred-Suite 和 ACPred 独立测试结果是使用对应文献中提供的在线预测平台测试获得的。从表 5 的结果可知:DeepPEPred 相对于 PEPred-Suite 和 ACPred,在 Acc、MCC、Sp 值方面都有较为显著的提升,其中 Acc 值分别提升了 29.6% 和 4.3%,MCC 值分别提升了 59.7% 和 9.4%,Sp 分别提升了 17.5% 和 10.3%,Sn 相比 PEPred-Suite 提升了 41.5%。这说明了该研究提出的模型对于 ACP 预测是有效的。

表 5　不同模型预测 ACP 的性能对比

模型	AUC	MCC	Acc	Sn	Sp
DeepPEPred	0.875	0.631	0.811	0.724	0.897
PEPred-Suite	/	0.034	0.515	0.309	0.722
ACPred	/	0.537	0.768	0.742	0.794

6　结束语

提出了一种基于深度学习的多肽预测方法 DeepPEPred。该方法利用四种特征对输入序列进行编码,将标准化的编码作为模型输入,经过贝叶斯参数调优,构建出一个最优的多种肽预测模型。

该方法的主要贡献是构造一个通用的模型,能有效预测多种肽。DeepPEPred 模型对不同的多肽表现出一致的鲁棒性,说明它具有很强的泛化能力。在四种肽数据集上与现有的方法进行了对比,实验结果表明:DeepPEPred 模型在 AUC、Acc 和 MCC 三个综合性评价指标上比现有的预测方法更好。

参考文献:

[1] 岳硕豪,田　弛,胡元昭,等. 抗癌肽研究进展[J]. 生物技术通报,2017,33(11):41-47.

[2] 尹昆仑,王嘉榕,孙红宾. 抗菌肽的研究进展及应用前景[J]. 中国生化药物杂志,2015(5):181-185.

[3] 张萌萌,姜　宁,张爱忠,等. 细胞穿透肽的转导机制及应用现状[J]. 基因组学与应用生物学,2019,38(6):2546-2550.

[4] FENG Bo, DAI Youzhi, WANG Lu, et al. A novel affinity ligand for polystyrene surface from a phage display random library and its application in anti-HIV-1 ELISA system[J]. Biologicals, 2009, 37(1): 48-54.

[5] LATA S, SHARMA B K, RAGHAVA G. Analysis and prediction of antibacterial peptides[J]. BMC Bioinformatics, 2007, 8(1): 263.

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

[7] LI N, KANG J, JIANG L, et al. PSBinder: a web service for predicting polystyrene surface-binding peptides[J]. BioMed Research International, 2017, 2017(6): 5761517.

[8] WEI L, ZHOU C, CHEN H, et al. ACPred-FL: a sequence-based predictor using effective feature representation to improve the prediction of anti-cancer peptides[J]. Bioinformatics, 2018, 34(23): 4007-4016.

[9] WEI L, ZHOU C, SU R, et al. PEPred-suite: improved and robust prediction of therapeutic peptides using adaptive feature representation learning[J]. Bioinformatics, 2019, 35(21): 4272-4280.

[10] YI H, YOU Z, ZHOU X, et al. ACP-DL: a deep learning long short-term memory model to predict anticancer peptides using high-efficiency feature representation[J]. Molecular Therapy Nucleic Acids, 2019, 17: 1-9.

[11] BOOPATHI V, SUBRAMANIYAM S, MALIK A, et al. mACPpred: a support vector machine-based meta-predictor for identification of anticancer peptides[J]. International Journal of Molecular Sciences, 2019, 20(8): 1964.

[12] RAO B, ZHOU C, ZHANG G, et al. ACPred-fuse: fusing multi-view information improves the prediction of anticancer peptides[J]. Briefings in Bioinformatics, 2019, 21(5): 1846-1855.

[13] TYAGI A, KAPOOR P, KUMAR R, et al. In silico models for designing and discovering novel anticancer peptides[J]. Scientific Reports, 2013, 3(1): 2984.

[14] TYAGI A, TUKNAIT A, ANAND P, et al. CancerPPD: a database of anticancer peptides and proteins[J]. Nucleic Acids Research, 2015, 43(Database issue): D837-D843.

[15] LI W, GODZIK A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences[J]. Bioinformatics, 2006, 22(13): 1658-1659.

[16] CHEN Z, ZHAO P, LI F, et al. iLearn: an integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data[J]. Briefings in Bioinformatics, 2019(1): 1-11.

[17] CHEN Z, ZHAO P, LI F, et al. iFeature: a Python package and web server for features extraction and selection from protein and peptide sequences[J]. Bioinformatics, 2018, 34(14): 2499-2502.

[18] RAO H B, ZHU F, YANG G B, et al. Update of PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence[J]. Nucleic Acids Research, 2006, 39: 385-390.

[19] CHOU K. Prediction of protein cellular attributes using pseudo-amino acid composition[J]. Proteins, 2001, 43(3): 246-255.

[20] SHEN H, CHOU K. PseAAC: a flexible web server for generating various kinds of protein pseudo amino acid composition[J]. Analytical Biochemistry, 2008, 373(2): 386-388.

[21] SHALABI L A, SHAABAN Z, KASASBEH B. Data mining: a preprocessing engine[J]. Journal of Computer Science, 2006, 2(9): 735-739.

[22] LECUN Y, BENGIO Y, HINTON G E. Deep learning[J]. Nature, 2015, 521(7553): 436-444.

[23] 辛月振,孙贝贝,夏盛瑜. 数据挖掘方法在生物实验数据上的应用[J]. 计算机技术与发展,2018,28(9):143-146.

[24] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[25] 司 阳,肖秦琨. 基于长短时记忆和动态贝叶斯网络的序列预测[J]. 计算机技术与发展,2018,28(9):59-63.

[26] 潘伟靖,陈德旺. 基于 GRU-SVR 的短时交通流量预测研究[J]. 计算机技术与发展,2019,29(10):11-14.

[27] CHO K, VAN MERRIENBOER B, GULCEHRE C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[C]//Conference on empirical methods in natural language processing, vol. 3: conference on empirical methods in natural language processing (EMNLP 2014). Doha, Qatar: [s. n.], 2014: 1724-1734.

[28] CHEON J H, HAN K, KIM A, et al. Bootstrapping for approximate homomorphic encryption[M]//Advances in cryptology - EUROCRYPT 2018. Tel Aviv, Israel: Springer, 2018: 360-384.

[29] SNOEK J, LAROCHELLE H, ADAMS R P. Practical Bayesian optimization of machine learning algorithms[C]//Advances in neural information processing systems. Lake Tahoe, Nevada, USA: [s. n.], 2012: 2951-2959.

[30] SCHADUANGRAT N, NANTASENAMAT C, PRACHAYASITTIKUL V, et al. ACPred: a computational tool for the prediction and analysis of anticancer peptides[J]. Molecules, 2019, 24(10): 1973.