基于 BERT 和集成学习的抗菌肽预测

高皖陵，赵 俊，岳振宇*

（安徽农业大学 信息与人工智能学院，合肥 230036）

摘要 利用计算方法准确识别抗菌肽是近年来生物信息学领域研究的重点问题。传统的机器学习方法需要自主从序列信息中提取和选择特征，导致抗菌肽识别准确率低。为此提出基于 BERT 的深度学习预测方法，从预训练策略、词向量嵌入、预测性能等方面比较了 4 种现有基于 BERT 的抗菌肽预测模型，并基于集成学习思想提出了一个新的抗菌肽预测工具。实验结果表明，该模型在多个性能评价指标上都有所提升。

关键词 抗菌肽；评估；BERT；集成学习；预训练模型

中图分类号 TP399 **文献标志码** A DOI 10.12178/1001-0548.2023295

Antimicrobial Peptides Prediction Based on BERT and Ensemble Learning

GAO Wanling, ZHAO Jun, and YUE Zhenyu*

(School of Information and Artificial Intelligence, Anhui Agricultural University, Hefei 230036, China)

Abstract As the best substitute for antibiotics, antimicrobial peptides (AMPs) have important research significance. How to accurately identify AMPs using computational methods has been a key issue in the field of bioinformatics in recent years. However, traditional machine learning methods require autonomous extraction and selection of features from sequence information, resulting in low AMPs identification accuracy. Faced with the above challenges, a deep learning prediction methods based on Bidirectional Encoder Representation from Transformers (BERT) is proposed. In order to conduct a comprehensive evaluation of existing BERT-based AMP tools and further improve the performance of AMP calculation methods, four existing BERT-based AMP prediction tools in terms of pre-training strategies, word vector embeddings, and prediction performance are compared, and thus a novel AMP prediction tool based on the idea of ensemble learning is proposed. The experimental results show that the proposed model has been improved on several performance evaluation indexes.

Key words antimicrobial peptides; assessment; BERT; ensemble learning; pre-trained model

由于抗生素的滥用，致病菌的耐药性问题日益严重，已经对人类健康造成巨大的威胁。寻找新的抗生素原料是保证人类生命安全的有效途径。抗菌肽（Antibacterial Peptides, AMPs）是生物先天免疫系统的关键组成部分，具有广泛的医学功能，因此被认为是抗生素的最佳替代品，具有重要的生物学研究意义[1]。

近年来，研究人员建立了许多抗菌肽公共数据库，其中包括综合数据库和专业数据库。综合数据库整合不同来源和类型的 AMPs，如 APD3[2] 和 CAMPR4[3]。专业数据库包括特定类别或来源的 AMPs，如 DRAMP[4] 和 dbAMP[5]。虽然相较于综合数据库，专业数据库收集的数据相对较少，但在研究特定类型的抗菌肽时，使用专业数据库能够获得更详细的分析和描述，促进对 AMPs 领域的研究。

目前对 AMPs 的识别和鉴定主要分为湿实验和计算机辅助识别两种方法。湿实验的设计复杂、操作困难、耗时，并且需要大量的人力物力来满足大规模批量预测的需求。越来越多的研究都使用计算机辅助识别 AMPs。计算机辅助识别方法可以分为基于经验分析的方法和基于机器学习的方法。基于经验分析的方法主要利用已知的经验规则对肽链性

收稿日期：2023−11−28；修回日期：2024−01−08
基金项目：国家自然科学基金 (62102004)
作者简介：高皖陵，主要从事生物信息计算方面的研究。
*通信作者 E-mail: zhenyuyue@ahau.edu.cn

质与抗菌活性之间的关系进行统计分析，然后建立预测模型。建模的方法主要包括主成分分析和偏最小二乘法等。这种方法本质上是识别待测试序列是否具有训练集上的某些特定特征，其依赖于训练集现有的语义模式，很难迁移到其他类型的 AMPs 识别任务上。近年来，基于机器学习方法的研究也不断涌现[6]，包括 CAMPr3[7]，Geu-AMP50[8]，AmPEP[9]，AntiBP2[10] 和 amPEPpy[11] 等。现有的机器学习方法通常利用多种物理化学性质作为输入特征，其中包括氨基酸组成（AAC）、伪氨基酸组成（PseAAC）、电荷、等电点、疏水性、极性和二级结构等。结合序列特征和理化性质，机器学习算法如随机森林（RF）和支持向量机（SVM）被应用于 AMPs 的二元分类任务。这些方法能够从不同的维度捕捉 AMPs 的结构和性质，为下游分类任务提供丰富的信息。

除机器学习方法外，深度学习的方法也开始广泛应用于 AMPs 的预测，如 AMPscanner[12]，ACEP[13] 和 APIN[14]。与传统的机器学习方法相比，深度学习方法能够自动提取特征，使其在处理原始输入序列数据时表现更为出色。由于 AMPs 的氨基酸序列和自然语言具有相似之处，自然语言处理（NLP）领域的深度学习方法也被引入 AMP 的识别和预测中，如文献 [12] 提出结合卷积层和长短期记忆网络（LSTM）的 AMPs 预测工具，文献 [15] 采用的双向 LSTM 端到端网络。这些自然语言处理方法极大地提高了 AMPs 识别的速度和准确性，但仍然存在一定的提升空间。

近年来，BERT 预训练模型[16] 在许多自然语言处理任务中表现优异。它将以自注意力机制为核心的预训练策略和下游微调任务相结合，提高了 AMPs 的分类预测效果。Bert-protein[17] 从 UniProt 中下载蛋白质序列用于 BERT 模型的预训练，同时采用 3 种不同的分词方法，在 6 个 AMPs 数据集上进行微调，显著提高了模型在不同数据集上的通用性。LM_pred[18] 采用 BERT 模型生成上下文嵌入来表示肽序列中的氨基酸，并选择卷积神经网络（CNN）作为下游任务分类器，效果分类。AMP-BERT[19] 利用 BERT 架构从输入的氨基酸序列中提取结构或功能信息。此外，AMP-BERT 还利用自注意力机制，实现可解释性的特征分析，帮助模型确定已知 AMPs 中具有抗菌功能的特定残基，提高模型的分类性能。cAMPs-pred[20] 将氨基酸序列视为文本信息，每个氨基酸都是一个单词代码，在 BERT 模型的末尾添加了一个线性层，将维度降至二维，并使用交叉熵损失函数训练模型。

本文从预训练策略、词向量嵌入、预测性能等方面全面比较了 4 种现有的基于 BERT 的抗菌肽预测模型，包括 Bert-protein、LM_pred、AMP-BERT、cAMPs-pred。基于集成学习的方法是传统分类问题中最广泛使用的技术，集成方法的优势在于它们与基学习器相比，在预测性能上表现的更好。本文基于集成学习的思想，选择 SVM 和 XGBoost 两种机器学习集成算法提出了一个新的 AMPs 预测工具。实验结果表明，此方法提高了模型的预测性能。通过结合多个深度学习模型的预测结果，能够有效提高模型的分类能力，相比较于单分类器更具优势。集成学习可以减小模型的方差，减少单个模型的过拟合风险，提高模型的鲁棒性。

1 基于 BERT 的抗菌肽预测模型

1.1 Bert-protein

Bert-protein 是一种基于 BERT 构建的 AMPs 预测模型。Bert-protein 将预训练策略应用于 AMPs 分类器的模型训练中，并提出了一种新的识别算法。通过使用 3 种肽链分词方法（k-mer，k = 1，2，3）在 6 个不同的 AMPs 数据集上对模型进行微调和测试，证明了预训练的优势和平衡正负样本的作用。

预训练过程包括遮蔽语言模型（Masked Language Model, MLM）和下一句预测（Next Sequential Prefetching, NSP）两个任务。MLM 随机覆盖序列中 15% 的单词，并通过最大化似然函数来预测这些掩蔽令牌。在 NSP 任务中，数据被随机分为两部分。在 50% 的数据中，句子对在上下文中是连续的，而其余的一半则不是。在该模型中，每个标记的输入向量由标记嵌入、段嵌入和位置嵌入 3 部分组成，充分的训练使 Bert-protein 能够全面了解蛋白质的长期依赖性，然后修改输出层的结构，使模型能够完成特定的任务。

1.2 LM_pred

LM_pred 是一种基于预训练语言模型和深度学习的 AMPs 预测模型。该模型利用在大型蛋白质数据库上进行预训练的语言表示模型，生成上下文嵌入。并通过卷积神经网络（CNN）作为分类器预测 AMPs。卷积层的应用使滤波器可以解释由上下文语言模型嵌入表示的氨基酸之间的空间和时间依赖性。

LM_pred 采用预训练模型生成上下文化的嵌入的方法，克服了现有的词向量嵌入方法不能传达由每个氨基酸的位置所编码的复杂的上下文信息的缺陷。在 BERT 模型中，创建词向量嵌入时会产生 [CLS] 和 [SEP] 两种特殊的标记。其中 [CLS] 作为一个智能的平均一维向量，总结了完整的二维嵌入，通常用作 NLP 分类任务的输入；而 [SEP] 则用于将任何特殊标记与嵌入分开。由于并非所有的语言模型都产生 [CLS] 标记，LM_pred 使用了完全嵌入，确保了结果之间更大的可比性。此外，采用完整嵌入可以确保有价值的信息不会丢失。

1.3　AMP-BERT

AMP-BERT 是一种具有来自 BERT 架构的微调双向编码器表示的深度学习模型。该模型可以从输入肽序列中提取结构或功能信息；此外，还利用注意力机制实现可解释性的特征分析，帮助确定已知抗菌肽序列中有助于肽结构和抗菌功能的特定残基。

AMP-BERT 在两个方面对现有的 AMPs 分类模型进行了改进：1）利用外部数据做出更准确的类预测。2）凸显了重要的肽残基，有助于其分类过程。AMP-BERT 首先将每个氨基酸序列标记为单个残基使用正弦函数进行位置编码，再通过下游的全连接层（FC Layer）和 sigmoid 函数对来自预训练模型 ProtBERT-BFD 的 BERT 编码器进行微调。AMP-BERT 能够充分利用 BERT 的注意机制，提高预测的准确性，并通过随后的自我注意分析，捕捉 AMPs 的重要结构特征。

1.4　cAMPs-pred

cAMPs-pred 是一个统一的 AMPs 识别管道。在宏基因组数据中识别开放阅读框架（sORFs；长度为 5~50 个氨基酸）非常耗时并占用计算资源。结合改进的 BERT 模型和大规模的人类微生物组数据资源可以发现具有高抗菌能力的 AMPs。

cAMPs-pred 将氨基酸视为文本信息，每个氨基酸都是一个单词代码。在训练过程中，氨基酸用间隙分隔，序列起始和结束位置使用 [CLS] 和 [SEP] 标签标记。在 BERT 模型的最后添加了一个线性层，将维数降至二维。并使用交叉熵损失函数，通过 ADAM 优化器对 BERT 进行微调。为了防止过拟合，cAMPs-pred 采用了早期停止策略进行训练，一旦模型的性能开始下降，训练就会停止并保存。cAMPs-pred 提取了 BERT 模型的最后一个隐藏层，反映其在序列分类中的个体属性。对于每个序列，获得大小为 $1\times N$ 的输出向量，N 是在 BERT 模型构建过程中确定的，大小为 768，且不添加任何处理过的信息。

2　集成实验

2.1　实验方案

为了客观比较 4 种现有的基于 BERT 的抗菌肽预测工具，并基于集成学习的思想提出一种新的抗菌肽预测模型，本文构建了一个 AMP 综合数据集。对于正样本数据，整合来自 CAMP、APD、DRAMP 和 dbAMP 这 4 个 AMPs 公共数据库中的所有 AMPs。排除数据集中氨基酸序列长度大于 100 或小于 10 的样本，最终获得 1 916 条正样本数据。为了构建负样本数据集，从 UniProt[21] 数据库中检索肽序列，随后排除所有包含“抗菌”相关关键词的序列，并去除长度大于 100 或小于 10 的样本。为了平衡正负样本的数量，从中随机挑选出 1 916 条肽序列作为负样本数据。

数据集构建完成后，将样本数据分别输入到 4 个基于 BERT 的 AMPs 预测模型中，以获得每一个基模型的 AMPs 预测结果。选择 SVM 和 XGBoost 作为本文的集成模型，并将每个基模型的样本预测概率作为特征输入到集成模型 SVM 和 XGBoost 中进行分类。此外，使用五折交叉验证策略[22] 对模型进行性能评估。将原始样本数据随机分为 5 个部分，每次选择其中 4 个部分作为训练集，剩余的部分作为测试集。交叉验证重复 5 次，以 5 次实验结果的平均值作为模型的性能度量标准。相关数据与代码可以在如下网址下载：https://github.com/WanlingGao/AMPpred-BERT-ensemble，整体的实验流程如图 1 所示。

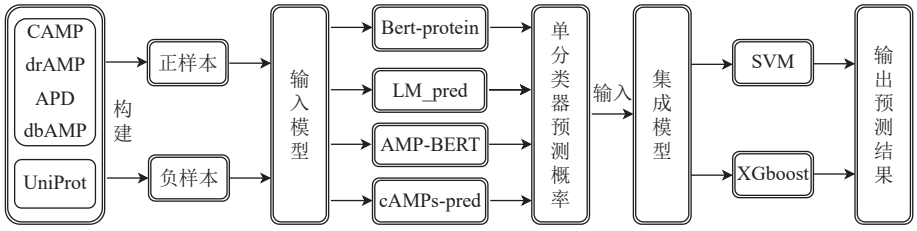


图 1　实验方案流程图

在模型性能比较方面，对每个模型进行了准确性（Accuracy）测试，在测试集上，准确性被认为是目标度量，因为正确地识别正样本和负样本同样重要。此外，还计算了灵敏度（Sensitivity）、特异性（Specificity）、精确率（Precision）、F1 分数（F1_Score）以及 Matthew 相关系数（MCC）来评估模型的性能。在抗菌肽预测模型评估方面，本文还使用了 ROC 曲线下的面积（AUC）和 PR 曲线下的面积（AUPR）作为额外的评价指标。常见的分类评价指标计算公式如下：

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

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP}+\text{FN}} \tag{2}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN}+\text{FP}} \tag{3}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} \tag{4}$$

$$\text{F1\_Score} = \frac{2\text{TP}}{2\text{TP}+\text{FP}+\text{FN}} \tag{5}$$

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

2.2 集成算法

2.2.1 SVM

支持向量机（Support Vector Machine, SVM）是一种用于分类和回归的机器学习算法。其基本思想是找到一个能够将数据点分隔开的超平面，使得间隔（margin）最大化，同时限制分类错误。超平面可用线性方程$\boldsymbol{\omega}^{\mathrm{T}}x+\boldsymbol{b}=0$表示。$\boldsymbol{\omega}$表示法向量，$\boldsymbol{b}$表示位移项。对于新的数据点，可以使用超平面来进行分类决策。

为了最大化间隔，SVM 的优化目标是找到$\boldsymbol{\omega}$和$\boldsymbol{b}$，以最大化$\|\boldsymbol{\omega}\|$，同时满足以下约束条件。

1）对于每个训练样本(x_i,y_i)，都满足：

$$y_i(\boldsymbol{\omega}^{\mathrm{T}}x_i+\boldsymbol{b})\geqslant 1 \tag{7}$$

2）$\|\boldsymbol{\omega}\|$需要最小化。

SVM 是一种高效的学习器，使用集成技术可以进一步增强 SVM 的性能。文献 [23] 指出偏差–方差分解是 SVM 集成的理论基础，并提出了两个发展 SVM 集成的方向：选择低偏差支持向量机的套袋集成和支持向量机的异构集成。文献 [24] 表明支持向量机集合是单一支持向量机的一种交叉验证优化，因此比其他模型具有更稳定的分类性能。

2.2.2 XGBoost

XGBoost（Extreme Gradient Boosting）是一种强大的集成学习算法，用于解决分类和回归问题。它通过集成多个弱学习器（通常是决策树），逐步提升模型的性能。

XGBoost 的总优化目标是最小化数据损失和正则化项的和，再加上一个常数项以控制树的复杂度。这可以表示为：

$$\text{Obj}(W)=\sum_{i=1}^{n}L(y_i,\widehat{y_i})+\sum_{j=1}^{T}\lambda\|\boldsymbol{\omega}\|+\sum_{j=1}^{T}\gamma+C \tag{8}$$

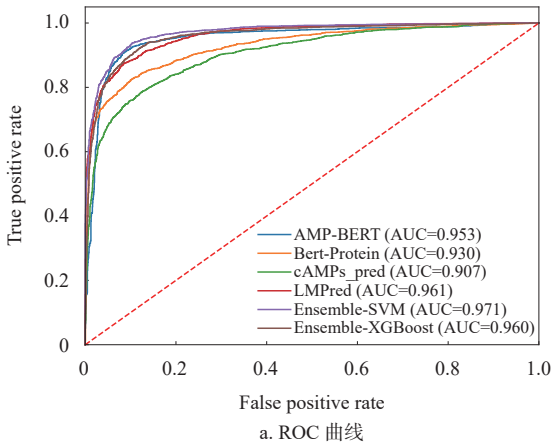
式中，n是训练样本的数量；T表示树的数量；λ是正则化强度的超参数；$\|\boldsymbol{\omega}\|$表示模型参数 $\boldsymbol{\omega}$ 的范数；γ是控制叶节点数量的惩罚参数；C是常数项。

XGBoost 基于树增强机器学习算法有助于处理更平滑的“偏–方差”权衡。它是集成学习方法的一种实现，通过梯度提升框架提供了一些额外的功能和优化。与其他集成学习算法相比，该方法在泛化性能、速度和精度方面优势卓越[25]。

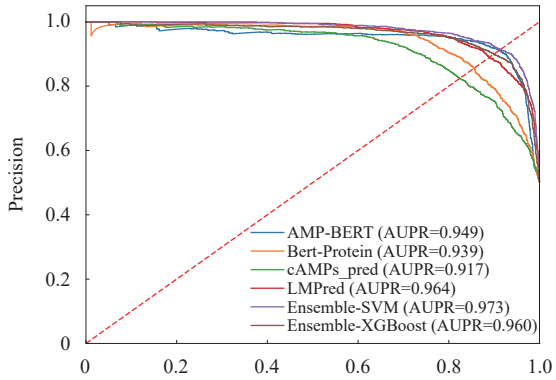
3 结果与讨论

3.1 基于 BERT 的抗菌肽预测模型评估

为了对现有的基于 BERT 的 AMPs 预测模型（Bert-protein、LM_pred、AMP-BERT 和 cAMPs-pred）进行全面的评估，研究它们在 AMPs 预测方面的性能。本文采用五折交叉验证策略，选择 AUC 和 AUPR 两个评价指标来衡量模型的预测性能，并绘制了对应的 ROC 曲线和 PR 曲线，如图 2 所示。同时还选择了 Sn、Sp、Pr、Acc、F1-Score、MCC 和 AUC 6 个常用的分类评价指标来更加全面地评估这些工具的预测性能，实验结果见表 1。



a. ROC 曲线



b. PR 曲线

图 2　基于 BERT 的抗菌肽预测模型和集成模型的性能评估

实验结果表明，AMP-BERT 在所有的评价指标上预测性能都最好。分析模型结构，可以得出利用自注意力机制实现可解释性的特征分析，帮助模型捕获抗菌肽的重要功能特征是 AMP-BERT 优于其他预测工具的重要原因。

3.2　集成模型预测性能

本文选择 SVM[26] 和 XGBoost[27] 两类机器学习算法对现有的 4 种基于 BERT 的抗菌肽预测模型进行集成，提出了一个新的抗菌肽预测工具。将 4 种预训练模型输出的抗菌肽预测概率作为特征输入到集成模型中，得到数据样本的最终分类结果。为验证集成模型在预测性能上的优势，在构建的抗菌肽综合数据集上，采用五折交叉验证策略，从包括 Sn、Sp、Pr、Acc、F1-Score、MCC 和 AUC 在内的多个分类评价指标上比较了基于 SVM 和基于 XGBoost 的集成模型与其他基模型在预测性能上的差异，实验结果见表 1 和图 2。

表 1　基于 BERT 的抗菌肽预测模型和集成分类模型的性能比较

Model	Sensitivity（Sn）	Specificity（Sp）	Precision（Pr）	Accuracy（Acc）	F1_Score	MCC	AUC
Ensemble-SVM	**0.9181**	0.9092	**0.9178**	**0.9136**	**0.9132**	**0.8279**	**0.9721**
	(±0.020)	(±0.025)	(±0.018)	(±0.010)	(±0.011)	(±0.021)	(±0.004)
Ensemble-XGBoost	0.9061	0.8930	0.9054	0.8995	0.8987	0.8000	0.9605
	(±0.024)	(±0.034)	(±0.020)	(±0.015)	(±0.016)	(±0.029)	(±0.010)
AMP-BERT	0.9113	**0.9118**	0.9121	0.9115	0.9115	0.8236	0.9631
	(±0.024)	(±0.020)	(±0.018)	(±0.013)	(±0.013)	(±0.026)	(±0.005)
Bert-Protein	0.8106	0.9045	0.8950	0.8575	0.8501	0.7191	0.9316
	(±0.036)	(±0.028)	(±0.041)	(±0.027)	(±0.029)	(±0.053)	(±0.008)
cAMPs_pred	0.8231	0.8241	0.8239	0.8236	0.8235	0.6479	0.9088
	(±0.028)	(±0.031)	(±0.021)	(±0.015)	(±0.016)	(±0.029)	(±0.010)
LM_pred	0.8773	0.9092	0.9067	0.8933	0.8923	0.7895	0.9617
	(±0.032)	(±0.022)	(±0.043)	(±0.018)	(±0.025)	(±0.036)	(±0.006)

由于 SVM 自身具有较好的分类性能和泛化能力，Ensemble-SVM 在多个评价指标上预测性能都最高，特别是 AUC 值得到显著提升。由于只有 AUC 是与预测阈值无关的最客观的指标，进一步验证了 Ensemble-SVM 模型的优势。同时 Ensemble-XGBoost 的预测性能却低于 AMP-BERT，这是因为 SVM 相对于 XGBoost 更适合较小的数据集且分类效果更稳定。综上，对比实验结果表明集成学习方法可以提高抗菌肽的预测性能。

4　结束语

本文全面评估了现有的基于 BERT 的抗菌肽预测方法，为获得更好的预测性能，利用集成学习思想结合 4 种预测工具，包括 Bert-protein、LM_pred、AMP-BERT 和 cAMPs-pred，同时选择 SVM 和 XGBoost 两种集成学习算法进行对比，并将性能最好的集成方法作为本文提出的新的抗菌肽预测工具。实验结果表明，在所有的基分类器中，AMP-BERT 的表现最为突出，SVM 集成算法的预测性能优于 XGBoost。

目前由于模型仍存在一定的黑盒性，我们将在后续的研究中增加对预测结果的可解释性分析。并考虑将 BERT 模型与附加的模态信息相结合，以全面捕捉抗菌肽的多层次特征，提高模型的预测性能。

参 考 文 献

[1] SMITH W P J, WUCHER B R, NADELL C D, et al. Bacterial defences: Mechanisms, evolution and antimicrobial resistance[J]. Nature Reviews Microbiology, 2023, 21: 519-534.

[2] WANG G S, LI X, WANG Z. APD3: The antimicrobial peptide database as a tool for research and education[J]. Nucleic Acids Research, 2016, 44(D1): D1087-D1093.

[3] GAWDE U, CHAKRABORTY S, WAGHU F H, et al. CAMPR4: A database of natural and synthetic

antimicrobial peptides[J]. Nucleic Acids Research, 2023, 51(D1): D377-D383.
[4] KANG X Y, DONG F Y, SHI C, et al. DRAMP 2.0, an updated data repository of antimicrobial peptides[J]. Scientific Data, 2019, 6: 148.
[5] JHONG J H, CHI Y H, LI W C, et al. dbAMP: An integrated resource for exploring antimicrobial peptides with functional activities and physicochemical properties on transcriptome and proteome data[J]. Nucleic Acids Research, 2019, 47(D1): D285-D297.
[6] 刘明友, 刘红美, 张招方, 等. 抗微生物肽机器学习预测算法综述[J]. 电子科技大学学报, 2022, 51(6): 830-840.
LIU M Y, LIU H M, ZHANG Z F, et al. Review of machine learning prediction algorithms for antimicrobial peptides[J]. Journal of University of Electronic Science and Technology of China, 2022, 51(6): 830-840.
[7] WAGHU F H, BARAI R S, GURUNG P, et al. CAMPR3: A database on sequences, structures and signatures of antimicrobial peptides[J]. Nucleic Acids Research, 2016, 44(D1): D1094-D1097.
[8] SACHIN P, MADHU T, VIVEKANAND K, et al. GEU-AMP50: Enhanced antimicrobial peptide prediction using a machine learning approach[J]. Materials Today: Proceedings, 2023, 73(P1): 81-87.
[9] BHADRA P, YAN J L, LI J Y, et al. AmPEP: Sequence-based prediction of antimicrobial peptides using distribution patterns of amino acid properties and random forest[J]. Scientific Reports, 2018, 8: 1697.
[10] LATA S, MISHRA N K, RAGHAVA G P. AntiBP2: Improved version of antibacterial peptide prediction[J]. BMC Bioinformatics, 2010, 11(Suppl 1): S19.
[11] LAWRENCE T J, CARPER D L, SPANGLER M K, et al. amPEPpy 1.0: A portable and accurate antimicrobial peptide prediction tool[J]. Bioinformatics, 2021, 37(14): 2058-2060.
[12] VELTRI D, KAMATH U, SHEHU A. Deep learning improves antimicrobial peptide recognition[J]. Bioinformatics, 2018, 34(16): 2740-2747.
[13] FU H, CAO Z, LI M, et al. ACEP: Improving antimicrobial peptides recognition through automatic feature fusion and amino acid embedding[J]. BMC Genomics, 2020, 21(1): 597.
[14] SU X, XU J, YIN Y, et al. Antimicrobial peptide identification using multi-scale convolutional network[J]. BMC Bioinformatics, 2019, 20(1): 730.
[15] YOUMANS M, SPAINHOUR C, QIU P. Long short-term memory recurrent neural networks for antibacterial peptide identification[C]//Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine. New York: IEEE, 2017: 498-502.
[16] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[EB/OL]. [2023-04-25]. https://arxiv.org/pdf/1810.04805.pdf.
[17] ZHANG Y, LIN J Y, ZHAO L M, et al. A novel antibacterial peptide recognition algorithm based on BERT[J]. Briefings in Bioinformatics, 2021, 22(6): bbab200.
[18] DEE W. LMPred: Predicting antimicrobial peptides using pre-trained language models and deep learning[EB/OL]. [2023-04-25]. https://www.xueshufan.com/publication/3215887418.
[19] LEE H, LEE S, LEE I, et al. AMP-BERT: Prediction of antimicrobial peptide function based on a BERT model[J]. Protein Sci, 2023, 32(1): e4529.
[20] MA Y, GUO Z Y, XIA B B, et al. Identification of antimicrobial peptides from the human gut microbiome using deep learning[J]. Nature Biotechnology, 2022, 40: 921-931.
[21] CONSORTIUM T U. UniProt: A worldwide hub of protein knowledge[J]. Nucleic Acids Research, 2019, 47(D1): D506-D515.
[22] XU J, LI F Y, LEIER A, et al. Comprehensive assessment of machine learning-based methods for predicting antimicrobial peptides[J]. Briefings in Bioinformatics, 2021, 22(5): bbab083.
[23] VALENTINI G, DIETTERICH T G. Bias-variance analysis of support vector machines for the development of SVM-based ensemble methods[J]. Journal of Machine Learning Research, 2004, 5: 725-775.
[24] VALENTINI G. An experimental bias-variance analysis of SVM ensembles based on resampling techniques[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 2005, 35(6): 1252-1271.
[25] NOBRE J, NEVES R F. Combining principal component analysis, discrete wavelet transform and XGBoost to trade in the financial markets[J]. Expert Systems with Applications, 2019, 125: 181-194.
[26] YAO G, HU X J, WANG G X. A novel ensemble feature selection method by integrating multiple ranking information combined with an SVM ensemble model for enterprise credit risk prediction in the supply chain[J]. Expert Systems with Applications, 2022, 200: 117002.
[27] MAHESH T R, VINOTH KUMAR V, MUTHUKUMARAN V, et al. Performance analysis of XGBoost ensemble methods for survivability with the classification of breast cancer[J]. Journal of Sensors, 2022, 2022: 4649510.

编 辑 税 红