抗微生物肽机器学习预测算法综述

刘明友[1,2]，刘红美[1]，张招方[3]，朱映雪[1]，黄 健[2*]

(1. 贵州医科大学生物与工程学院 贵阳 550025；2. 电子科技大学生命科学与技术学院 成都 610041；
3. 泰禾云工程咨询有限公司 贵阳 550081)

【摘要】传统抗微生物肽识别分析主要通过实验手段进行，效率低，耗费较多人力物力。最新的抗微生物肽识别方法是将计算机技术和生物信息学相结合，通过机器学习方法进行大数据挖掘分析，从大量的多肽序列数据里面预测抗微生物肽，从而加快抗微生物肽的识别。收集并分类整理了近 10 年来计算机辅助抗微生物肽识别的研究文献，从中梳理出抗微生物肽的主要数据资源、抗微生物肽识别的特征工程、抗微生物肽的机器学习预测算法和抗微生物肽的回归分析方法。同时，进一步对机器学习算法的模型性能评估方法进行综述，总结其中存在的不足并展望了未来的发展方向。

关 键 词 抗微生物肽； 生物信息学； 生物医学大数据； 机器学习
中图分类号 TP399; R318 **文献标志码** A doi:10.12178/1001-0548.2022188

Review of Machine Learning Prediction Algorithms for Antimicrobial Peptides

LIU Mingyou[1,2], LIU Hongmei[1], ZHANG Zhaofang[3], ZHU Yingxue[1], and HUANG Jian[2*]

(1. School of Biology and Engineering, Guizhou Medical University Guiyang 550025;
2. School of Life Science and Technology, University of Electronic Science and Technology of China Chengdu 610041;
3. Taiheyun Engineering Consulting Co., Ltd. Guiyang 550081)

Abstract The traditional methods for the identification of antimicrobial peptides are experimental means, which is inefficient and consumes a lot of manpower and material resources. The latest ways to identify antimicrobial peptides combine computer technology, bioinformatics, and machine learning methods together. Based on big data mining and analysis, antimicrobial peptides can be predicted from a large amount of peptide sequence data. The identification of antimicrobial peptides thereby could be accelerated. This paper classifies and analyzes the main literatures of the computer-aided antimicrobial peptide recognition in the recent 10 years, sorts out the main antimicrobial peptides data resources, the characteristic engineering of antimicrobial peptide recognitions, the machine learning prediction algorithms of antimicrobial peptides, the regression analysis methods of antimicrobial peptides. In the meanwhile, this paper reviews the model performance evaluation methods of machine learning algorithms, summarizes the existing shortcomings, and prospects the future development directions.

Key words antimicrobial peptides; bioinformatics; biomedical big data; machine learning

随着新冠病毒的大流行，微生物感染造成的伤害越来越严重。世界卫生组织在 2017 年估计，仅流感每年就造成多达几十万人死亡，而新冠病毒大流行也已导致数百万人死亡。病毒、耐药细菌、真菌等微生物感染已成为人类面临的严重健康威胁。传统药物在治疗微生物感染性疾病时，会出现诸多问题，包括产生耐药性、毒副作用等，急需开发安全高效的新型抗微生物感染药物。抗微生物肽 (antimicrobial peptides, AP) 是能抵抗微生物感染的多肽[1]，包括抗菌肽、抗病毒肽、抗真菌肽等，具有高效、低毒、广谱的抗微生物活性的优点且基本无耐药性问题[2]。以抗菌肽为例，与抗生素相比，抗菌肽能快速杀死细菌，还有免疫调节作用等优点[3]。抗菌肽对高等动物的正常细胞基本没有毒性作用，

收稿日期：2022－06－05；修回日期：2022－08－25
基金项目：国家自然科学基金 (62071099)；贵州省卫生健康委科学技术基金 (gzwkj2023-590)
作者简介：刘明友 (1985－)，男，主要从事生物信息学及大数据方面的研究.
*通信作者：黄健，E-mail：hj@uestc.edu.cn

还能抑制某些靶肿瘤细胞的生长。因此，抗菌肽已成为人类与动物医学研究的热点[4-5]。传统识别抗微生物肽的方法通过生物实验来进行，随着高通量测序技术的发展和测序成本的持续降低，产生了海量的测序数据。用传统方法从高通量序列中识别抗微生物肽工作量大、效率低、耗时费力、成本高昂。抗微生物肽预测方法风生水起，这类方法通过对已有抗微生物肽数据的分析来挖掘出序列特征和抗微生物活性之间的关联，从而做出定性或定量的推断[6-7]。由于不依赖于生物实验，其计算方法具有高效快捷、成本低廉等特点[8]，非常适合大规模抗微生物肽的数据预测[9]。本文将从数据资源、数据处理方法、预测算法、性能评估等几个模块对抗微生物肽预测研究进行综述。

1　数据资源

随着生物医学实验与生物信息学的发展，相关学者已构建了一批抗微生物肽数据库。这些数据资源的积累，为后续抗微生物肽预测算法研究提供了必不可少的数据支撑。抗微生物肽的详细数据资源如表 1 所示。

表 1　抗微生物肽数据资源详细列表

数据类型	数据库名称	URL	简介
抗微生物肽综合数据库	APD	http://aps.unmc.edu/AP/	抗菌肽数据为主，包含抗病毒肽、抗癌肽等数据
	DRAMP	http://dramp.cpu-bioinfor.org/	包含有专利的抗菌肽和临床验证的抗菌肽等数据
抗病毒肽数据库	ACovPepDB	http://i.uestc.edu.cn/ACovPepDB/	收录了518条抗冠状病毒肽序列数据
	AVPdb	http://crdd.osdd.net/servers/avpdb	收录了经过医学验证的多种抗病毒肽数据
	AVPpred	http://crdd.osdd.net/servers/avppred	收集了大量经过实验验证的抗病毒肽数据
	HIPdb	http://crdd.osdd.net/servers/hipdb	通过人工管理的经过实验验证的抗病毒肽数据库
抗菌肽数据库	CAMP	http://www.camp3.bicnirrh.res.in	包含抗菌肽的序列、结构及其家族特异性等信息
	DBAASP	http://dbaasp.org	收集了大量经过实验验证的抗菌肽及其靶标信息
	dbAMP	http://awi.cuhk.edu.cn/dbAMP	收录了大量经过实验验证的抗菌肽数据
抗真菌肽数据库	PlantAFP	http://bioinformatics.cimap.res.in/sharma/PlantAFP/	收集了经过实验验证的植物源抗真菌肽数据

1.1　抗菌肽数据

APD (antimicrobial peptide database) 是一个以抗菌肽数据为主的抗微生物肽数据库[10]，收集并存储了 2 169 条抗菌肽序列及其功能活性等特征信息。此外，该数据库还收录了 172 条抗病毒肽[11]、80 条抗寄生虫肽和 185 条抗癌肽。CAMP 抗菌肽数据库[12] 收录了抗菌肽的序列、结构及家族特异性等方面的信息。升级版的 CAMPR3 目前拥有 10 247 条序列，为研究抗菌肽的结构和功能信息提供了资源。DBAASPv3 抗菌肽数据库[13] 存储了大量通过实验验证的抗菌肽及其靶标信息，该数据库包含超过 15 700 条序列记录，包括超过 14 500 条单体及近 400 条同源和异源多聚体抗菌肽。 在单体抗菌肽 (monomeric antimicrobial peptides, AMP) 中，超过 12 000 条是合成的，约 2 700 条是核糖体合成的，约 170 条是非核糖体合成的。DRAMP 抗菌肽数据库[14] 收录了普通抗菌肽、有专利的抗菌肽和经过临床验证的抗菌肽。DRAMP2.0 版本共收录条目 19 899 条 (新增条目 2 550 条)，其中一般条目 5 084 条、专利条目 14 739 条、临床验证的条目 76 条，与 APD 和 CAMP 相比，DRAMP包含 14 040 条非冗余序列，DRAMP 已经更新到 3.0 版本[15]，包含 22 259 条抗菌肽记录 (新增 2 360 条)，其中一般条目 5 891 条、专利条目 16 110 条、临床验证的条目 77 条、stapled 抗菌肽 181 条。dbAMP 抗菌肽数据库[16] 包括 4 271 条经过实验验证的抗菌肽和 8 118 条根据其功能活性推测的抗菌肽。升级后的 dbAMP2.0[17] 数据库包含了来自 3 044 个物种的 26 447 条抗菌肽和 2 262 条抗菌蛋白。

1.2　抗病毒肽数据

AVPdb 抗病毒肽数据库[18] 提供了 60 余种经过医学验证了的能够抵抗如流感病毒[19]、丙型肝炎病毒[20]、单疱疹病毒[21]、呼吸道合胞病毒[22]、乙型肝炎病毒[23]、登革热病毒[24]、SARS 病毒[25] 等感染的多肽序列。HIPdb 抗病毒肽数据库[26] 是一个手工管理的数据库，收录经过实验验证的 981 条抗病毒肽，包含抗病毒肽的序列、长度、来源、靶标、细胞系等各方面的信息。AVPpred 数据库[27] 收集了 1 245 条经过实验验证的能够抵抗如流感、HIV、HCV 和 SARS 等重要人类病毒的抗病毒肽记录，同时还提供抗病毒肽预测服务。也有针对特定病毒的抗病毒肽数据库，如最新的抗冠状病毒肽数据

库 ACovPepDB[28]。该数据库收集了大量的抗冠状病毒肽数据资源，主要来自于 1972～2021 年间的 2 199 篇已发表论文，还有部分抗冠状病毒肽数据从 AVPpred[27] 和 DPL[29] 等数据库收集而来。该数据库共收录了 518 条抗冠状病毒肽序列，其中 214 条为非冗余序列，包括抗冠状病毒肽的名称、长度、来源、靶标等信息。该数据库的构建为后期抗冠状病毒肽的预测分析研究提供了资源。

1.3 抗真菌肽数据

在前述抗菌肽数据库中，有的也收录了不少抗真菌肽数据[30]，如 APD 数据库除了收集大量抗菌肽数据外，还收录了 959 条抗真菌肽记录，升级到 APD3 后，抗真菌肽增加到了 1 133 条。DRAMP 数据库中也有 1 802 条抗真菌肽。PlantAFP[31] 是一个植物源抗真菌肽数据库，收集了经过实验验证的植物源抗真菌肽数据，该数据库的当前版本包含 2 585 条肽条目，每个条目都包含肽的综合信息，包括肽序列、肽名称、肽类别、肽长度、分子量、抗真菌活性和肽来源，并以 SMILES 格式存储肽序列。为了方便用户使用，该数据库中集成了许多检索工具，包括 BLAST 搜索、肽搜索、SMILES 搜索，且还包含肽图。Antifp_main[32] 也收集了大量抗真菌肽的研究数据，共计1 168 条抗真菌肽记录。文献 [33] 在抗真菌肽研究上做了大量工作，通过计算机辅助方法，从大量多肽序列中，定量预测了 5 000 多条抗真菌肽数据。

2 抗微生物肽分析方法

2.1 传统实验方法

为了确定多肽的抗微生物活性，需要做许多实验验证的工作，如文献 [34] 通过实验确定了家蝇的防卫素 (Phormicin) 多肽[35] 对金黄色葡萄球菌和耐甲氧西林金黄色葡萄球菌 (Methicillin-resistant Staphylococcus aureus, MRSA)[36] 的体内外抗感染作用。在小鼠烫伤模型中，经防卫素处理后，MRSA 细菌载量明显下降。在黑水虻幼虫实验中，该防卫素破坏了金黄色葡萄球菌和 MRSA 生物膜的形成，表明家蝇防卫素通过影响生物膜和相关基因网络，帮助宿主抑制 MRSA 感染。

通过实验能够确定多肽的抗微生物活性，但这样的实验需要耗费较多人力，所需时间和开销也很大。面对高通量多肽数据，更适合的策略是开发快速高效的预测方法进行初筛，再对最为可能的候选抗微生物肽进行实验验证，这就需要对抗微生物肽数据进行特征提取。

2.2 计算机辅助特征工程

多肽序列由 20 种氨基酸残基序列组成，这 20 种氨基酸分别为：甘氨酸 (Gly 缩写 G) 、丙氨酸 (Ala 缩写 A) 、缬氨酸 (Val 缩写 V) 、亮氨酸 (Leu 缩写 L) 、异亮氨酸 (Ile 缩写 I) 、甲硫氨酸 (Met 缩写 M) 、脯氨酸 (Pro 缩写 P) 、色氨酸 (Trp 缩写 W) 、丝氨酸 (Ser 缩写 S) 、酪氨酸 (Tyr 缩写 Y) 、半胱氨酸 (Cys 缩写 C) 、苯丙氨酸 (Phe 缩写 F) 、天冬酰胺 (Asn 缩写 N) 、谷氨酰胺 (Gln 缩写 Q) 、苏氨酸 (Thr 缩写 T) 、天冬氨酸 (Asp 缩写 D) 、谷氨酸 (Glu 缩写 E) 、赖氨酸 (Lys 缩写 K) 、精氨酸 (Arg 缩写 R) 和组氨酸 (His 缩写 H)，这 20 种氨基酸是组成生命体中蛋白质的主要单元[37]。要进行肽功能的识别，首先需要提取多肽序列中的特征信息。

2.2.1 AAC 计算法

氨基酸组分 (amino acid composition, AAC) [38] 是指在给定序列中，20 种天然氨基酸各自出现的频率，然后计算每一种氨基酸在整个肽序列中的组分属性，具体如下：

$$\mathrm{AAC}=\frac{X(i)}{\sum_{i=1}^{20}X(i)} \tag{1}$$

式中，i 表示任意一种氨基酸；$X(i)$代表整个序列中第 i 种氨基酸出现的次数，如氨基酸序列 ‘KTCENLADTFRGPCFATSNC’，其中氨基酸 A 出现了 2 次，则其 AAC 值为 2/20=0.1，通过 AAC 的分析提取肽序列中每种氨基酸的特征信息。

2.2.2 DPC 计算法

DPC (dipeptide composition) 二肽组分是 AAC 计算法的扩展，它统计氨基酸对出现的次数特征信息[39]，如序列‘LFRLIKSLIKRLVSAFK’中 LI 出现了 2 次，则其计算特征为 2/(17−1)=0.125。同理，可计算三肽出现的特征信息。

2.2.3 CKSAAP

CKSAAP 描述符[40] 是 DPC 的进一步扩展。当 k=0 时，该特征方法就是 DPC，通过计算两个氨基酸对之间间隔的氨基酸数来提取特征；当 k=3 时，表示一对氨基酸之间间隔 3 个氨基酸残基， 该方法作为一个有效的特征描述符来表示短肽序列的特征信息，其计算公式为：

$$\left(\frac{N_{\mathrm{AA}}}{N_{\mathrm{total}}},\frac{N_{\mathrm{AC}}}{N_{\mathrm{total}}},\frac{N_{\mathrm{AD}}}{N_{\mathrm{total}}},\cdots,\frac{N_{\mathrm{YY}}}{N_{\mathrm{tatol}}}\right)_{400} \tag{2}$$

为了更细化地进行分析，CKSAAGP[41] 将氨基酸按照其物理化学性质分成 5 类，分别为：脂肪族氨基酸 (g1)，芳香族氨基酸 (g2)，带正电荷氨基酸 (g3)，带负电荷氨基酸 (g4) 和不带电氨基酸 (g5)。如“aliphatic.*X*.*X*.aromatic”，其中“*X*”表示任何出现在“脂肪族和芳香族”两种氨基酸之间的氨基酸对，对于长度为 *L* 的氨基酸肽，如果 *k* 间隔残基对在肽中出现了 *n* 次，则特征计算为 $n/(L-(k+1))$，从而提取到氨基酸的 CKSAAGP 特征信息。

类似这种氨基酸提取方法还有很多，这些特征选择算法大部分都集成到 iFeature 中[42-43]，它能够计算和提取包含 53 种不同类型的特征描述符。iFeature 还集成了 12 种不同类型的常用的特征聚类、选择和降维算法，为后期的机器学习预测算法提供强大的支撑。

2.3　其他特征工程

随着机器学习的演进，特征提取算法也出现了许多新变化，新的特征工程为特定的预测分析方法提供了更多选择。

2.3.1　PEPred-Suite

PEPred-Suite[44] 利用 10 种常用的特征编码方法对特征进行编码，包括：AAC、DPC、GGAP[45]、ASDC[46]、组成−转换−分布 (CTD)[47]、20 位特征 (BIT20)、21 位特征 (BIT21)[48]、重叠属性特征、信息论特征和物理化学特征 (188D) [49]。这些特征编码信息作为特征数据集存储在特征池中，然后运用随机森林算法对特征进行学习，从而产生新的特征向量，将特征向量输入最小冗余最大相关算法中进行特征排序[50]，然后将排序靠前的特征作为后期随机森林分类的学习特征。

2.3.2　Meta-iAVP

Meta-iAVP[51] 进一步提出了新的特征选择方法，该模型先利用前面的数学模型特征提取算法把多肽序列特征提取出来，然后将这些特征输入随机森林、支持向量机、KNN 等机器学习算法模型中进行进一步特征学习提取，将学习提取到的特征作为后期随机森林分类算法的特征输入。

2.3.3　iAMP-CA2L

iAMP-CA2L[52] 使用了新的特征提取方法，首先将氨基酸按照二进制编码进行转化，20 种氨基酸按照 5 位二进制编码方法转换成 20 种二进制编码，然后利用 CAL 机制将二进制编码转换成图片[53]。该方法的特点是图片可以通过裸眼的方式区分不同多肽序列的差异之处，该方法转换生成的图片如图 1 所示。

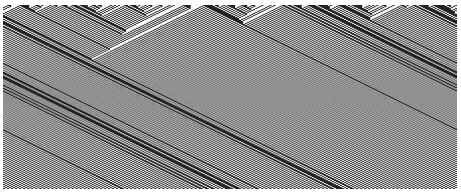

图 1　CAL 转换后的肽序列图片

有了多肽序列图片数据，就可以通过深度学习或神经网络来进行图像特征学习，iAMP-CA2L 采用卷积神经网络 (convolutional neural network, CNN) 来进行图像特征学习，CNN 是一种前馈神经网络，通过卷积运算提取特征，然后使用池化层学习数据的局部特征。它不需要对输入数据进行大量预处理，可以学习到真实反映数据类型的内在相关性的特征信息[54]。然后将 CNN 学习到的特征输入到 BiLSTM 模型中进行上下文特征提取，从而获取到最终的多肽序列特征信息，将这些信息最终输入支持向量机进行分类分析。BiLSTM 由前向 LSTM 和后向 LSTM 组成，BiLSTM 在自然语言处理任务中常用于处理上下文信息，使用 LSTM 模型可以更好地捕捉长距离依赖，因为 LSTM 在学习过程中可以学习到哪些信息要记住，哪些信息要忘记。BiLSTM 能够有效地捕获到前向和后向特征之间的关系，从而更好地学习到多肽序列特征之间的特征信息[55]。

3　抗微生物肽预测算法

近年来，随着机器学习算法的日新月异，许多优秀的机器学习算法不断运用到生物医药大数据的分析当中，加快了抗微生物肽的预测识别进程。

3.1　机器学习预测方法

3.1.1　随机森林

最典型的机器学习预测方法就是随机森林算法 (random forest, RF) [56]，RF 算法是一个包含多个决策树的分类器。RF 模型基于许多弱分类和回归树 (classification and regression tree, CART) 而成，其中每个分类器是使用独立于输入向量采样的随机向量生成的，以提高 CART 的预测性能[57]。

RF 已被广泛用于模拟各种生物学问题当中[58]。文献 [59] 首次运用 RF 进行抗病毒肽的预测，该模型提取抗病毒肽序列的物理化学属性作为特征选

项，然后运用 RF 算法进行分类学习，其准确率为 90%，马修斯相关系数 (Matthews correlation coefficient, MCC)[60] 为 0.79。另一个运用随机森林进行抗病毒肽预测分析的模型是 2019 年文献 [61] 提出的 AntiVPP，其准确率为 93.00%，MCC 为 0.87。同样 2019 年基于 RF 的 PEPred-Suite[44]，其抗病毒肽预测性能为 86.4%，MCC 系数为 0.725。基于 RF 的 AMPfun[62] 作为一个抗微生物肽分类模型，在独立数据集上测试，其抗病毒肽预测准确率为 86.13%，MCC 值为 0.71，抗真菌肽预测准确率为 74.58%，MCC 值为 0.52。2021 年发表的 PreAntiCoV[63] 用于抗冠状病毒肽的预测，考虑到数据的不平衡性，该模型引入非平衡随机森林技术预测抗冠状病毒肽的性能指标，其 MCC 值为 0.57。同样，基于随机森林的 AVPIden[64] 用来预测分析抗病毒肽，其准确率为 91.50%。

3.1.2 支持向量机

支持向量机 (support vector machine, SVM) 模型[65] 可以通过将输入样本映射到更高维空间，然后搜索用于构造分类器的超平面来解决由于使用小型训练数据集而引起的过拟合问题。为了对高维样本进行线性分离，SVM 采用许多核函数将输入从具有 p 维特征向量的样本空间转换为具有 n 维特征向量的特征空间，其中 $p < n$。2012 年提出的 AVPpred[66] 模型运用支持向量机来进行抗病毒肽的预测，并构建了基准数据集，其抗病毒肽预测最高准确率为 85.00%，MCC 为 0.70。另外一个使用支持向量机构建的模型是 2020 年的 FIRM-AVP[67] 模型，其预测抗病毒肽准确率为 92.40%，MCC 值为 0.84。2017 年发表的 iAMPpred[68] 也构建在支持向量机基础之上，其抗菌肽预测准确率最高为 94.69%，MCC 相关系数为 0.89，其抗病毒肽预测准确率最高为 90.08%，MCC 相关系数为 0.80，其抗真菌肽预测最高准确率为 93.35%，MCC 相关系数为 0.87。

3.1.3 神经网络

基于神经网络的算法分为两种，一种对氨基酸进行二进制编码，然后将编码信息输入神经网络进行特征学习，将学习到的抗微生物肽特征信息输入分类算法进行识别。iAMP-CA2L[52] 就是其中的典型算法之一，该算法用于识别抗菌肽的最高准确率为 94.13%，然后在识别出的抗菌肽中分辨抗病毒功能，其分辨抗病毒肽准确率为 80.57%。另一种是将传统数学模型提取的特征信息输入到神经网络当中，通过神经网络来进行学习，从而完成抗病毒肽的识别，典型代表是 ENNAVIA[69]。该算法专门用来进行抗病毒肽和抗冠状病毒肽的预测，算法模型基于深度神经网络构建而成，其抗病毒肽最高准确率为 93.90%，MCC 值为 0.87。Deep-AntiFP[70] 基于深度神经网络技术，用来对抗真菌肽进行预测分析，在独立数据集上进行测试，其准确率为 89.08%，MCC 值为 0.78。

3.1.4 其他分类模型

iAMP-2L[71] 基于模糊 k 最邻近 (fuzzy k-nearest neighbor, FKNN) 网络进行抗菌肽分类，该算法进行二阶段分类识别，第一阶段识别肽序列是否为抗菌肽，第二阶段对识别到的抗菌肽进行功能区分，其抗菌肽识别最高准确率为 92.23%，MCC 值为 0.84。另一个抗菌肽预测模型是基于谷歌公司推出的 BERT 模型[72]，该模型是一个自然语言处理模型，能够实现自然语言的上下文识别，该模型的抗菌肽识别最高准确率为 95.94%，MCC 值为 0.91。该模型第一阶段利用网络公开蛋白质数据集对模型进行预训练，然后再用预训练模型对抗菌肽进行识别。iAFPs-EnC-GA[73] 是基于多个分类模型构建的抗真菌肽集成分类器，首先对氨基酸进行特征编码，然后将抗真菌肽的编码特征信息输入 FKNN、随机森林 (RF) 模型、K 近邻 (KNN) 模型以及 SVM 模型进行分类，其抗真菌肽预测最高准确率为 93.92%。

4 抗微生物肽回归分析

文献 [33] 通过支持向量机回归模型 (support vector regression, SVR) 验证了抗真菌肽对念珠菌属等真菌的有效性，回归分析中，其相关性系数 R 均大于 0.90，进一步证实了该模型在抗真菌活性预测分析方面的准确性。文献 [74] 通过深度学习方法构建分类模型，然后对识别后的抗菌肽进行实验验证，在对小鼠进行体内感染实验之前，评估了 11 种 c_AMP 对真核细胞的毒性，最终选择了 c_-AMP1043、c_-AMP593 和 c_-AMP575 进行体内分析，使用感染肺炎克雷伯菌的小鼠模型，监测体重恢复数据情况。结果表明，3 种抗菌肽对肺部感染具有抗菌活性，对宿主无明显不良影响。iAFPs-EnC-GA 通过本地模型无关局部解释技术 (local interpretable model-agnostic explanations, LIME) 分析解释了单个特征对整体预测的贡献，同时引入另一个黑盒模型事后归因解析算法 (shapley additive exPlana-

tion, SHAP) 来衡量每个特征在建议模型中的贡献。SHAP 是基于最佳 Shapley 值聚合的全局解释方法，具有提供可解释的预测的能力，还涵盖了由于缺乏特征的方向性而发生的限制[73]。如果 SHAP 值为正，这意味着该特征推动了对抗真菌肽的预测并产生了积极的影响。如果 SHAP 值为负，则该特征会推动对非抗菌肽的预测并产生负面影响，LIME 分析测量单个特征，而 SHAP 分析测量整个模型特征。抗微生物肽预测模型与工具如表 2 所示。

表 2　抗微生物肽预测模型与工具列表

类型	名称	URL	简介
抗微生物肽分类预测模型与工具	AMPfun	http://fdblab.csie.ncu.edu.tw/AMPfun/index.html	基于随机森林的抗微生物肽预测模型
	AntiVPP 1.0	https://github.com/bio-coding/AntiVPP	基于随机森林的抗病毒肽预测模型
	AVPIden	http://awi.cuhk.edu.cn/AVPIden/	基于随机森林抗病毒肽预测模型
	AVPpred	http://crdd.osdd.net/servers/avppred	基于支持向量机的分类预测模型
	Deep-AntiFP	https://github.com/shahidawkum/Deep-AntiFP	基于深度神经网络的抗真菌肽预测模型
	ENNAVIA	https://research.timmons.eu/ennavia	基于深度神经网络的抗病毒肽预测模型
	FIRM-AVP	https://msc-viz.emsl.pnnl.gov/AVPR	基于支持向量机的抗病毒肽预测模型
	iAMP-2L	http://www.jci-bioinfo.cn/iAMP-2L	基于FKNN的多分类预测模型
	iAMP-CA2L	http://www.jci-bioinfo.cn/	基于神经网络的抗微生物肽预测模型
	iAMPpred	http://cabgrid.res.in:8080/amppred/	基于支持向量机的抗微生物肽预测模型
	PEPred-Suite	http://server.malab.cn/PEPred-Suite	基于随机森林的治疗性预测模型
	PreAntiCoV	https://github.com/poncey/PreAntiCoV	基于随机森林的抗冠状病毒肽预测模型
抗微生物肽回归预测模型与工具	APD	https://www.chemoinfolab.com/antifungal/	基于SVR构建的抗真菌肽预测模型
	c_AMP	https://github.com/mayuefine/c_AMPs-prediction	基于深度神经网络的分类模型
	iAFPs-EnC-GA	https://github.com/farmanit335/iAFPs-EnC-GA	基于集成模型的抗真菌肽分类模型

5　模型性能评估方法

为评价模型的性能，引入众多机器学习模型评价指标，进行模型之间的相互比较。列举部分常用模型评估方法：1) true positive (TP)：将正例预测为正例的个数；2) true negative (TN)：将负例预测为负例的个数；3) false positive (FP)：将负例预测为正例的个数，即误报；4) false negative (FN)：将正例预测为负例的个数，即漏报。通过评估混淆矩阵，就可以完成对模型性能的定量评估[75]。

5.1　分类模型性能评估

5.1.1　准确率

准确率 (accuracy, ACC) 表示识别准确的正例和负例占总体样本的比例。通常而言，准确率越高，模型也越好。具体计算如下：

$$\mathrm{ACC}=\frac{\mathrm{TN}+\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}+\mathrm{TN}+\mathrm{FP}} \tag{3}$$

5.1.2　敏感性

敏感性 (sensitivity, Sn)，用来表示正例预测为正占所有样例预测为正的比例，用来测试将抗菌肽、抗病毒肽正确分类的能力，其计算公式如下：

$$\mathrm{Sn}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{4}$$

在生物医学数据分析中，敏感性是测试所有患者中成功定位了多少患者，敏感性越高，正确识别出患有疾病的患者判别能力越强。

5.1.3　特异性

特异性 (specificity, Sp) 用来表示将负例预测为负例的个数，测试正确区分非抗菌肽和非抗病毒肽的能力。具体计算如下：

$$\mathrm{Sp}=\frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} \tag{5}$$

在生物医学方面，特异性是测试所有健康人中有多少健康人被检测为阴性，测试对没有疾病的健康人群的识别能力。

5.1.4　马修相关系数

马修相关系数 MCC 在机器学习中被用来衡量二分类和多分类的标准，它考虑了真值、假值、阳性和阴性等情况，通常被认为是一种平衡的度量措施，即使类别的数目大小不同，数据不平衡，也可以用来进行评估。MCC 本质上是−1～ +1 之间的相关系数值， +1 表示完美预测，0 表示平均随机预测，−1 表示逆预测，其计算公式如下：

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

除此之外，还有很多评价指标，可以根据实际需求进行选择，从而实现模型性能的进一步量化评估。

5.1.5 ROC 与 AUC

受试者工作特征 (receiver operating characteristic, ROC) 通过绘制一条曲线，显示各种截断点的假阳性率 (x 轴) 和真阳性率 (y 轴) 之间的权衡。假阳性率和真阳性率的计算公式为：假阳性率 (FPR) = FP/(TN+FP)，真阳性率 (TPR) = TP/(TP+FN)，ROC 曲线的形状提供了对模型性能的洞察，曲线越凸出，模型性能越好。曲线下面积 (area under the curve, AUC) 是度量 ROC 曲线下面积的指标，AUC 接近 1.0 表示预测接近完美，AUC 为 0.5 则表示随机猜测[76]。

5.2 回归模型性能评估

回归是估计一个因变量与一个或多个自变量之间关系的过程，通过比较预测结果与实际结果之间的差异率来评估回归模型的性能。

5.2.1 平均绝对误差

平均绝对误差 (mean absolute error, MAE)[77] 也叫平均绝对离差，这个指标先对真实值与预测值的距离求和，再取平均值。具体计算如下：

$$\mathrm{MAE}=\frac{1}{m}\sum_{i=1}^{m}|f(x_i)-y_i| \tag{7}$$

式中，$f(x_i)$ 为预测值；y_i为真实值；m 为数据量，平均绝对误差可以准确地反映实际预测误差的大小，但 MAE 的缺点是不能显示回归模型拟合是优还是劣。

5.2.2 均方根误差

均方根误差 RMSE 也称标准误差[77]，是在均方误差的基础上进行开方运算，常用于衡量观测值与真实值间的偏差，可以消除样本数量对评价指标的影响，使得评估指标的大小不会太依赖于样本数量，而是更多地反映模型误差，具体计算如下：

$$\mathrm{RMSE}=\sqrt{\frac{1}{m}\sum_{i=1}^{m}(f(x_i)-y_i)^2} \tag{8}$$

5.2.3 决定系数R^2

决定系数 R^2由 3 个指标组成[78]，分别为 SSR (sum of squares of the regression)， SST(total sum of squares) 和 SSE(sum of squares for error)，具体表达式为：

$$\mathrm{SSR}=\sum_{i=1}^{m}(f(x_i)-\bar{y})^2 \tag{9}$$

$$\mathrm{SST}=\sum_{i=1}^{m}(y_i-\bar{y})^2 \tag{10}$$

$$\mathrm{SSE}=\sum_{i=1}^{m}(f(x_i)-y_i)^2 \tag{11}$$

决定系数R^2 通过计算 SSR 与 SST 的比值，反应因变量 y 的全部变异能通过回归模型被自变量 x 解释的比例，如R^2 为 0.9，则表示回归关系可以解释因变量 90% 的变异。具体表达式为：

$$R^2=\frac{\mathrm{SSR}}{\mathrm{SST}} \tag{12}$$

决定系数R^2越高，越接近 1，模型的拟合效果就越好，反之，决定系数R^2 越接近 0，则回归直线拟合效果越差。

6 结 束 语

抗微生物肽是具有抗菌、抗病毒或者抗真菌功能的多肽[79-81]，相较于传统的抗生素或抗病毒药物，抗微生物肽尚无耐药等问题，可以作为传统抗生素等药物的替代品[82]。

目前，基于计算机辅助的抗微生物肽的预测方法已较为成熟，能提供较为可靠的预测结果，能减少大量的人力投入，成本更低、效率更高。

尽管机器学习预测抗微生物肽已经取得很大进步，但该领域依然存在诸多挑战。1) 当前机器学习预测算法针对抗菌肽的预测较多[83-84]，但是专门针对抗病毒肽的预测方法依然较少[85-86]。未来研究可以考虑将抗菌肽的预测分析方法迁移到抗病毒肽、抗真菌肽的研究当中。2) 抗微生物肽的机器学习预测算法研究虽然较多，但直接结合生物学意义进行可解释性分析以及随后开展生物学实验验证的研究较少[87-88]。目前，LIME、SHAP 等算法[89] 能够根据机器学习提取的特征进行 t 特征重要性排序，找出影响最终结果最多的氨基酸特征，从而为下一步的多肽功能的生物学特征研究提供参考。3) 当前抗菌肽数据库较多，但是专门的抗病毒肽、抗真菌肽数据库较少，如表 1 所示，现存的抗病毒肽、抗真菌肽数据大多来源于论文附带的数据。同时，不同的数据库数据格式各异，导致通用性不高，需要专门的标准化的抗微生物肽数据库[90]。4) 抗微生物肽的预测准确率有待进一步提高。最新的抗菌肽预

测准确率达到 95.94%，抗病毒肽预测准确率最高为 93.90%，抗真菌肽预测准确率最高为 93.35%。5) 目前有许多用于抗微生物肽预测的算法，大多基于传统机器学习分类方法，如随机森林、KNN、SVM 等，人工智能最新的预测分析算法还未完全引入抗微生物肽的预测当中。最新深度学习已经在生物信息学上进行了大量应用[91-92]，对抗神经网络也应用于最新分类研究[93]，图神经网络[94-95]、自然语言处理模型也开始应用到生物学数据处理上[96]，这些算法在抗微生物肽等生物医学数据分析的潜力有待进一步挖掘运用。

综上，基于对抗微生物肽的预测算法研究，目前仍需要进一步研究的 3 个方向：1) 开发专门的标准化的抗微生物肽数据库，收集大量分散的抗微生物肽序列，并归类整理、动态更新，为将来的研究分析提供支撑；2) 开发通用的针对抗微生物肽的机器学习预测算法；3) 开发更多能够解读抗微生物肽机器学习预测分析结果的算法，为生物学家下一阶段的实验验证提供理论依据。抗微生物肽的分析不仅是生物科学家的工作，更需要计算机、数学等相关行业专家积极参与、多方协作，才能完成抗微生物肽的分析、验证、推广、应用等一系列完整的生态系统研究。

参　考　文　献

[1] GOMES B, AUGUSTO M T, FELÍCIO M R, et al. Designing improved active peptides for therapeutic approaches against infectious diseases[J]. Biotechnology Advances, 2018, 36(2): 415-429.

[2] 曹隽喆, 顾宏. 基于计算方法的抗菌肽预测[J]. 计算机学报, 2017, 40(12): 2777.
CAO J Z, GU H. A review on prediction of antimicrobial peptides based on computational methods[J]. Chinese Journal of Computers, 2017, 40(12): 2777.

[3] PFALZGRAFF A, BRANDENBURG K, WEINDL G. Antimicrobial peptides and their therapeutic potential for bacterial skin infections and wounds[J]. Frontiers in Pharmacology, 2018, 9: 281.

[4] O'BRIEN-SIMPSON N M, HOFFMANN R, CHIA C S, et al. Antimicrobial and anticancer peptides[J]. Frontiers in Chemistry, 2018, 6: 13.

[5] LV Z, CUI F, ZOU Q, et al. Anticancer peptides prediction with deep representation learning features[J]. Briefings in Bioinformatics, 2021, 22(5): bbab008.

[6] SHARMA R, SHRIVASTAVA S, KUMAR S S, et al. Deep-AFPpred: Identifying novel antifungal peptides using pretrained embeddings from seq2vec with 1DCNN-BiLSTM[J]. Briefings in Bioinformatics, 2022, 23(1): bbab422.

[7] MIKUT R. Computer-based analysis, visualization, and interpretation of antimicrobial peptide activities[J]. Methods Mol Biol, 2010, 618: 287-299.

[8] KIESLICH C A, ALIMIRZAEI F, SONG H, et al. Data-driven prediction of antiviral peptides based on periodicities of amino acid properties[M]. Istanbul: Elsevier, 2021.

[9] LEE H T, LEE C C, YANG J R, et al. A large-scale structural classification of antimicrobial peptides[J]. Biomed Res Int, 2015, 2015: 475062.

[10] WANG G, LI X, WANG Z. APD3: The antimicrobial peptide database as a tool for research and education[J]. Nucleic Acids Research, 2016, 44(D1): 1087-1093.

[11] MIZUGUCHI T, OHASHI N, NOMURA W, et al. Anti-HIV screening for cell-penetrating peptides using chloroquine and identification of anti-HIV peptides derived from matrix proteins[J]. Bioorganic & Medicinal Chemistry, 2015, 23(15): 4423-4427.

[12] 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): 1094-1097.

[13] PIRTSKHALAVA M, AMSTRONG A A, GRIGOLAVA M, et al. DBAASP v3: Database of antimicrobial/cytotoxic activity and structure of peptides as a resource for development of new therapeutics[J]. Nucleic Acids Research, 2021, 49(D1): 288-297.

[14] KANG X, DONG F, SHI C, et al. DRAMP 2.0, an updated data repository of antimicrobial peptides[J]. Scientific Data, 2019, 6(1): 1-10.

[15] SHI G, KANG X, DONG F, et al. DRAMP 3.0: An enhanced comprehensive data repository of antimicrobial peptides[J]. Nucleic Acids Research, 2022, 50(D1): 488-496.

[16] 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): 285-297.

[17] JHONG J H, YAO L, PANG Y, et al. dbAMP 2.0: Updated resource for antimicrobial peptides with an enhanced scanning method for genomic and proteomic data[J]. Nucleic Acids Research, 2022, 50(D1): 460-470.

[18] QURESHI A, THAKUR N, TANDON H, et al. AVPdb: A database of experimentally validated antiviral peptides targeting medically important viruses[J]. Nucleic Acids Research, 2014, 42(D1): 1147-1153.

[19] JONES J C, SETTLES E W, BRANDT C R, et al. Identification of the minimal active sequence of an anti-influenza virus peptide[J]. Antimicrobial Agents and Chemotherapy, 2011, 55(4): 1810-1813.

[20] CHEN S L, MORGAN T R. The natural history of hepatitis C virus (HCV) infection[J]. International Journal of Medical Sciences, 2006, 3(2): 47.

[21] TAYLOR T J, BROCKMAN M A, MCNAMEE E E, et al. Herpes simplex virus[J]. Frontiers in Bioscience-Landmark, 2002, 7(4): 752-764.

[22] WELLIVER R C. Review of epidemiology and clinical risk factors for severe respiratory syncytial virus (RSV)

infection[J]. The Journal of Pediatrics, 2003, 143(5): 112-117.
[23] FERRARI C. Hbv and the immune response[J]. Liver International, 2015, 35: 121-128.
[24] MURPHY B R, WHITEHEAD S S. Immune response to dengue virus and prospects for a vaccine[J]. Annual Review of Immunology, 2011, 29: 587-619.
[25] WEC A Z, WRAPP D, HERBERT A S, et al. Broad neutralization of SARS-related viruses by human monoclonal antibodies[J]. Science, 2020, 369(6504): 731-736.
[26] QURESHI A, THAKUR N, KUMAR M. HIPdb: A database of experimentally validated HIV inhibiting peptides[J]. Plos One, 2013, 8(1): e54908.
[27] POLANCO C, SAMANIEGO J L, CASTAÑÓN-GONZÁLEZ J A, et al. Polar profile of antiviral peptides from AVPpred Database[J]. Cell Biochemistry and Biophysics, 2014, 70(2): 1469-1477.
[28] ZHANG Q, CHEN X, LI B, et al. A database of anti-coronavirus peptides[J]. Scientific Data, 2022, 9(1): 1-9.
[29] WANG F, LI N, WANG C, et al. DPL: A comprehensive database on sequences, structures, sources and functions of peptide ligands[J]. Database (Oxford), 2020, 2020: baaa089.
[30] LI T, LI L, DU F, et al. Activity and mechanism of action of antifungal peptides from microorganisms: A review[J]. Molecules, 2021, 26(11): 3438.
[31] TYAGI A, PANKAJ V, SINGH S, et al. PlantAFP: A curated database of plant-origin antifungal peptides[J]. Amino Acids, 2019, 51(10): 1561-1568.
[32] AGRAWAL P, RAGHAVA G. Prediction of antimicrobial potential of a chemically modified peptide from its tertiary structure[J]. Frontiers in Microbiology, 2018, 9: 2551.
[33] ZHANG J, YANG L, TIAN Z, et al. Large-scale screening of antifungal peptides based on quantitative structure-activity relationship[J]. ACS Medicinal Chemistry Letters, 2021, 13(1): 99-104.
[34] WANG B, YAO Y, WEI P W, et al. Housefly phormicin inhibits staphylococcus aureus and MRSA by disrupting biofilm formation and altering gene expression in vitro and in vivo[J]. International Journal of Biological Macromolecules, 2021, 167: 1424-1434.
[35] WANG B, WEI P W, YAO Y, et al. Functional and expression characteristics identification of Phormicins, novel AMPs from Musca domestica with anti-MRSA biofilm activity, in response to different stimuli[J]. International Journal of Biological Macromolecules, 2022, 209: 299-314.
[36] BRUMFITT W, HAMILTON-MILLER J. Methicillin-resistant staphylococcus aureus[J]. New England Journal of Medicine, 1989, 320(18): 1188-1196.
[37] NARITA M, NARITA M, ITSUNO Y, et al. Protein folding structures: Formation of folding structures based on probability theory[J]. ACS Omega, 2016, 1(6): 1355-1366.
[38] YANG S, HUANG J, HE B. CASPredict: A web service for identifying cas proteins[J]. PeerJ, 2021, 9: e11887.
[39] ZHOU Y, XIE S, YANG Y, et al. SSH2.0: A better tool for predicting the hydrophobic interaction risk of monoclonal antibody[J]. Frontiers in Genetics, 2022, 13: 842127.
[40] ZHOU Y W, YUE P, HUANG J. CISI2.0: A better tool for predicting cross-interaction or self-interaction of antibodies based on sequences[J]. Journal of University of Electronic Science and Technology of China, 2021, 50(5): 659-666.
[41] CHEN Y M, ZU X P, LI D. Identification of proteins of tobacco mosaic virus by using a method of feature extraction[J]. Frontiers in Genetics, 2020, 11: 569100.
[42] NASEER S, ALI R F, MUNEER A, et al. IAmideV-deep: Valine amidation site prediction in proteins using deep learning and pseudo amino acid compositions[J]. Symmetry, 2021, 13(4): 560.
[43] 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.
[44] 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.
[45] QIAN L, WEN Y, HAN G. Identification of cancerlectins using support vector machines with fusion of G-gap dipeptide[J]. Frontiers in Genetics, 2020, 11: 275.
[46] QIANG X, ZHOU C, YE X, et al. CPPred-FL: A sequence-based predictor for large-scale identification of cell-penetrating peptides by feature representation learning[J]. Briefings in Bioinformatics, 2020, 21(1): 11-23.
[47] HOU R, WU J, XU L, et al. Computational prediction of protein arginine methylation based on composition-transition-distribution features[J]. ACS Omega, 2020, 5(42): 27470-27479.
[48] KOU Z, FAN X, LI J, et al. Using amino acid features to identify the pathogenicity of influenza B virus[J]. Infectious Diseases of Poverty, 2022, 11(1): 1-13.
[49] ZOU Q, WAN S, JU Y, et al. Pretata: Predicting TATA binding proteins with novel features and dimensionality reduction strategy[J]. BMC Systems Biology, 2016, 10(4): 401-412.
[50] SAKAR C O, KURSUN O, GURGEN F. A feature selection method based on kernel canonical correlation analysis and the minimum redundancy-maximum relevance filter method[J]. Expert Systems with Applications, 2012, 39(3): 3432-3437.
[51] SCHADUANGRAT N, NANTASENAMAT C, PRACHAYASITTIKUL V, et al. Meta-iAVP: A sequence-based meta-predictor for improving the prediction of antiviral peptides using effective feature representation[J]. International Journal of Molecular Sciences, 2019, 20(22): 5743.
[52] XIAO X, SHAO Y T, CHENG X, et al. iAMP-CA2L: A new CNN-BiLSTM-SVM classifier based on cellular

automata image for identifying antimicrobial peptides and their functional types[J]. Briefings in Bioinformatics, 2021, 22(6): bbab209.
[53] XIAO X, LIN W Z, CHOU K C. Recent advances in predicting protein classification and their applications to drug development[J]. Current Topics in Medicinal Chemistry, 2013, 13(14): 1622-1635.
[54] SHIN H C, ROTH H R, GAO M, et al. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning[J]. IEEE Transactions on Medical Imaging, 2016, 35(5): 1285-1298.
[55] SIAMI-NAMINI S, TAVAKOLI N, NAMIN A S. The performance of LSTM and BiLSTM in forecasting time series[C]//2019 IEEE International Conference on Big Data (Big Data). Los Angeles: IEEE, 2019: 3285-3292.
[56] BIAU G. Analysis of a random forests model[J]. The Journal of Machine Learning Research, 2012, 13(1): 1063-1095.
[57] LOH W Y. Classification and regression trees[J]. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2011, 1(1): 14-23.
[58] BASITH S, MANAVALAN B, SHIN T H, et al. SDM6A: A web-based integrative machine-learning framework for predicting 6mA sites in the rice genome[J]. Molecular Therapy-Nucleic Acids, 2019, 18: 131-141.
[59] CHANG K Y, YANG J R. Analysis and prediction of highly effective antiviral peptides based on random forests[J]. PloS One, 2013, 8(8): e70166.
[60] CHICCO D, JURMAN G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation[J]. BMC Genomics, 2020, 21(1): 1-13.
[61] LISSABET J F B, BELÉN L H, FARIAS J G. AntiVPP 1.0: A portable tool for prediction of antiviral peptides[J]. Computers in Biology and Medicine, 2019, 107: 127-130.
[62] CHUNG C R, KUO T R, WU L C, et al. Characterization and identification of antimicrobial peptides with different functional activities[J]. Briefings in Bioinformatics, 2020, 21(3): 1098-1114.
[63] PANG Y, WANG Z, JHONG J H, et al. Identifying anti-coronavirus peptides by incorporating different negative datasets and imbalanced learning strategies[J]. Brief Bioinform, 2021, 22(2): 1085-1095.
[64] PANG Y, YAO L, JHONG J H, et al. AVPIden: A new scheme for identification and functional prediction of antiviral peptides based on machine learning approaches[J]. Briefings in Bioinformatics, 2021, 22(6): bbab263.
[65] 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.
[66] THAKUR N, QURESHI A, KUMAR M. AVPpred: Collection and prediction of highly effective antiviral peptides[J]. Nucleic Acids Research, 2012, 40(W1): W199-W204.
[67] CHOWDHURY A S, REEHL S M, KEHN-HALL K, et al. Better understanding and prediction of antiviral peptides through primary and secondary structure feature importance[J]. Scientific Reports, 2020, 10(1): 1-8.
[68] MEHER P K, SAHU T K, SAINI V, et al. Predicting antimicrobial peptides with improved accuracy by incorporating the compositional, physico-chemical and structural features into Chou 's general PseAAC[J]. Scientific Reports, 2017, 7(1): 1-12.
[69] TIMMONS P B, HEWAGE C M. ENNAVIA is a novel method which employs neural networks for antiviral and anti-coronavirus activity prediction for therapeutic peptides[J]. Briefings in Bioinformatics, 2021, 22(6): bbab258.
[70] AHMAD A, AKBAR S, KHAN S, et al. Deep-AntiFP: Prediction of antifungal peptides using distanct multi-informative features incorporating with deep neural networks[J]. Chemometrics and Intelligent Laboratory Systems, 2021, 208: 104214.
[71] XIAO X, WANG P, LIN W Z, et al. iAMP-2L: A two-level multi-label classifier for identifying antimicrobial peptides and their functional types[J]. Analytical Biochemistry, 2013, 436(2): 168-177.
[72] ZHANG Y, LIN J, ZHAO L, et al. A novel antibacterial peptide recognition algorithm based on BERT[J]. Briefings in Bioinformatics, 2021, 22(6): bbab200.
[73] AHMAD A, AKBAR S, TAHIR M, et al. iAFPs-EnC-GA: Identifying antifungal peptides using sequential and evolutionary descriptors based multi-information fusion and ensemble learning approach[J]. Chemometrics and Intelligent Laboratory Systems, 2022, 222: 104516.
[74] MA Y, GUO Z, XIA B, et al. Identification of antimicrobial peptides from the human gut microbiome using deep learning[J]. Nat Biotechnol, 2022, 40(6): 921-931.
[75] PARIKH R, MATHAI A, PARIKH S, et al. Understanding and using sensitivity, specificity and predictive values[J]. Indian Journal of Ophthalmology, 2008, 56(1): 45.
[76] DZISOO A M, HE B, KARIKARI R, et al. CISI: A tool for predicting cross-interaction or self-interaction of monoclonal antibodies using sequences[J]. Interdisciplinary Sciences: Computational Life Sciences, 2019, 11(4): 691-697.
[77] CHAI T, DRAXLER R R. Root mean square error (RMSE) or mean absolute error (MAE)[J]. Geoscientific Model Development Discussions, 2014, 7(1): 1525-1534.
[78] CHICCO D, WARRENS M J, JURMAN G. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation[J]. PeerJ Computer Science, 2021, 7: e623.
[79] VILASBOAS L C P, CAMPOS M L, BERLANDA R L A, et al. Antiviral peptides as promising therapeutic drugs[J]. Cellular and Molecular Life Sciences, 2019, 76(18): 3525-3542.

[80] EL-BITAR A M H, SARHAN M, ABDEL-RAHMAN M A, et al. Smp76, a scorpine-like peptide isolated from the venom of the scorpion Scorpio maurus palmatus, with a potent antiviral activity against hepatitis C virus and dengue virus[J]. International Journal of Peptide Research and Therapeutics, 2020, 26(2): 811-821.
[81] LI Q, ZHAO Z, ZHOU D, et al. Virucidal activity of a scorpion venom peptide variant mucroporin-M1 against measles, SARS-CoV and influenza H5N1 viruses[J]. Peptides, 2011, 32(7): 1518-1525.
[82] RIDER T H, ZOOK C E, BOETTCHER T L, et al. Broad-spectrum antiviral therapeutics[J]. PloS One, 2011, 6(7): e22572.
[83] MEHER P K, SAHU T K, SAINI V, et al. Predicting antimicrobial peptides with improved accuracy by incorporating the compositional, physico-chemical and structural features into Chou's general PseAAC[J]. Scientific Reports, 2017, 7: 42362.
[84] DONG G F, ZHENG L, HUANG S H, et al. Amino acid reduction can help to improve the identification of antimicrobial peptides and their functional activities[J]. Frontiers in Genetics, 2020, 21: 669328.
[85] ARONICA P G A, REID L M, DESAI N, et al. Computational methods and tools in antimicrobial peptide research[J]. Journal of Chemical Information and Modeling, 2021, 61(7): 3172-3196.
[86] BHADRA P, YAN J, LI J, et al. AmPEP: Sequence-based prediction of antimicrobial peptides using distribution patterns of amino acid properties and random forest[J]. Scientific Reports, 2018, 8(1): 1-10.
[87] MANAVALAN B, BASITH S, LEE G. Comparative analysis of machine learning-based approaches for identifying therapeutic peptides targeting SARS-CoV-2[J]. Briefings in Bioinformatics, 2022, 23(1): bbab412.
[88] QURESHI A, TANDON H, KUMAR M. AVP-IC50Pred: Multiple machine learning techniques-based prediction of peptide antiviral activity in terms of half maximal inhibitory concentration (IC50)[J]. Peptide Science, 2015, 104(6): 753-763.
[89] SUNDARARAJAN M, NAJMI A. The many Shapley values for model explanation[C]// International Conference on Machine Learning. [S.l.]: PMLR, 2020: 9269-9278.
[90] RAMAZI S, MOHAMMADI N, ALLAHVERDI A, et al. A review on antimicrobial peptides databases and the computational tools[J]. Database, 2022, 2022: baac011.
[91] ZHANG Y, YAN J, CHEN S, et al. Review of the applications of deep learning in bioinformatics[J]. Current Bioinformatics, 2020, 15(8): 898-911.
[92] LI J, PU Y, TANG J, et al. DeepAVP: A dual-channel deep neural network for identifying variable-length antiviral peptides[J]. IEEE Journal of Biomedical and Health Informatics, 2020, 24(10): 3012-3019.
[93] SUH S, LEE H, LUKOWICZ P, et al. CEGAN: Classification enhancement generative adversarial networks for unraveling data imbalance problems[J]. Neural Networks, 2021, 133: 69-86.
[94] ZHOU J, CUI G, HU S, et al. Graph neural networks: A review of methods and applications[J]. AI Open, 2020, 1: 57-81.
[95] WEI L, YE X, XUE Y, et al. ATSE: A peptide toxicity predictor by exploiting structural and evolutionary information based on graph neural network and attention mechanism[J]. Briefings in Bioinformatics, 2021, 22(5): bbab041.
[96] SALEM M, KESHAVARZI A A, YUAN J S. AMPDeep: Hemolytic activity prediction of antimicrobial peptides using transfer learning[J]. BMC Bioinformatics, 2022, 23(1): 1-17.

编 辑 刘飞阳