

深度学习与分子模拟相结合：高效筛选宏基因组数据中的抗菌肽

田源¹, 韩爱萍¹, 徐春明^{1,*}

(1.北京工商大学 轻工科学技术学院, 北京市 海淀区 100048)

摘要: 抗菌肽是一种可以通过与细菌细胞膜或细胞内生物分子相互作用, 破坏细菌生理过程, 最终导致细菌死亡, 发挥抗菌功能的一种多肽。鉴于食品安全和质量正在面临的挑战, 开发新型食品抗菌剂以提高食品安全性, 已成为当前食品科学研究的关键方向。通过构建全新的深度学习模型, 从土壤宏基因组数据中筛选抗菌肽, 并使用分子对接和分子动力学模拟等技术对筛选出的肽进行验证。模型的精准度为 98.7%、准确率为 96.5%、召回率为 91.9%、F1-score 为 95.2%、特异性为 99.2%, 该模型展现了出色的性能和强大的泛化能力, 同时在效率、可解释性和实际应用价值方面也体现了显著的优势。经过训练后, 模型成功地识别出了若干极具抗菌潜力的短肽, 并选择了部分样本进一步研究。结果表明, 筛选出的短肽 Gly-Thr-Ala-Trp-Arg-Trp-His-Tyr-Arg-Ala-Arg-Ser 能有效地附着在细菌的转录调节因子 MrkH 蛋白上, 对肺炎克雷伯氏菌、大肠杆菌和金黄色葡萄球菌产生抑菌作用。研究旨在结合深度学习与分子模拟技术等技术, 为开发食品行业中新型抗菌剂的应用提供一定的理论依据。

关键词: 机器学习; 分子对接; 分子动力学模拟; 抗菌肽; 抗菌活性验证

中图分类号: TS201.6

文献标识码:

文章编号:

Screening Antimicrobial Peptides from Metagenomes Based on Deep Learning and Molecular Simulation

TIAN Yuan¹, HAN Aiping¹, XU Chunming^{1,*}

(1. School of Light Industry, Beijing Technology and Business University Beijing 100048, China)

Abstract: Antimicrobial peptides are a type of peptide capable of exerting antibacterial functions by interacting with bacterial cell membranes or intracellular biomolecules, thereby disrupting bacterial physiological processes and ultimately leading to bacterial death. Given the challenges posed to food safety and quality, the development of novel food antimicrobials to enhance food safety has become a key direction in current food science research. A novel deep learning model was constructed to screen for antimicrobial peptides from soil metagenomic data, with the screened peptides being validated using techniques such as molecular docking and molecular dynamics simulations. The model demonstrated an outstanding performance with a precision of 98.7%, an accuracy of 96.5%, a recall rate of 91.9%, an F1-score of 95.2%, and a specificity of 99.2%, showcasing excellent efficiency, interpretability, and practical application value alongside robust generalization capabilities. Post-training, the model successfully identified several peptides with significant antimicrobial potential, with a subset chosen for further investigation. The findings revealed that the screened peptide Gly-Thr-Ala-Trp-Arg-Trp-His-Tyr-Arg-Ala-Arg-Ser could effectively attach to the bacterial transcription regulator protein MrkH, exhibiting inhibitory effects on *Klebsiella pneumoniae*, *Escherichia coli*, and *Staphylococcus aureus*. This study aims to provide a theoretical basis for the application of new antimicrobials in the food industry by integrating deep learning with molecular simulation technologies.

Keywords: Machine Learning; Molecular dynamics; Molecular docking; Bioactive Peptides; Antimicrobial activity

收稿日期: 2023-11-21

基金项目: 国家自然科学基金资助项目(30471225); 国家重点研发计划项目(2023YFD2100202)。

Foundation: National Natural Science Foundation of China (30471225); National Key Research and Development Program of China (2023YFD2100202).

第一作者: 田源, 男, 硕士研究生, 研究方向为生物信息学。

*通信作者: 徐春明, 男, 副教授, 博士, 主要从事生物过程方面的研究。

抗菌肽 (antimicrobial peptides, AMP) 是自然界中广泛存在且具有抗菌功能的多肽, 它们不仅在生命科学和医学研究领域发挥着不可或缺的作用, 同时在食品保质和安全领域也显示出巨大潜力^[1]。随着抗生素耐药性问题的日益严峻, 传统抗生素在保障食品安全方面的有效性正面临挑战^[2-3]。特别地, 细菌生物膜的形成大大增强了细菌对抗生素的抵抗能力^[4-5]。C2E (cyclic-di-GMP) 是一种广泛存在于革兰氏阴性菌中, 可以调节多种细胞过程包括毒力因子和基因表达的信号分子^[6-8]。在结构上, C2E 是一个仅含有 2 个鸟嘌呤碱基并且通过核糖和磷酸连接而成的环。在细菌中, 某些信号通过合成或降解环二鸟苷单磷酸来传递。MrkH 蛋白是一种依赖于 C2E 调控的经典转录激活剂, 它同时激活 III 型菌毛的基因簇和 *mrkHI* 操纵子。在这个正反馈循环中, MrkH 激活自身的产生, 在有利环境下确保细菌 MrkH 蛋白快速产生、III 型菌毛的表达及生物膜的形成^[9]。这种调控机制为从宏基因组数据中筛选能够抑制该通路的多肽提供了理论依据。

本研究旨在利用先进的计算方法和机器学习技术, 构建预测模型, 筛选出具有抑制 MrkH-C2E 细菌转录关键蛋白或信号通路的抗菌肽, 以干扰和控制生物膜的形成, 实现从基因组数据中高通量筛选极具潜力的食品抗菌剂, 为食品安全和抗生素耐药性问题提供新的解决方案。模型拟选取氨基酸序列和理化性质作为特征指标, 结合注意力 (attention, ATT) 模型、长短期记忆 (long short term memory, LSTM) 模型和卷积神经网络 (convolutional neural networks, CNN) 3 种深度学习方法对输入数据进行筛选。此外, 为了获得筛选结果在土壤微生物菌群中的含量, 我们将分析筛选结果的丰度信息, 以提供输出结果在细菌中具体的含量。并通过抑菌实验验证筛选结果对克雷伯氏细菌菌株和食源性致病菌的抗菌活性, 同时使用在线预测工具, 检测多肽的细胞毒性及其溶血性。鉴于在计算和实验评估中得到结果, 使用本研究提出的模型预测出的抗菌肽可以安全应用于食品, 为食品抗菌提供新的策略^[10]。

1 材料与方法

1.1 材料与试剂

实验中所采用的肺炎克雷伯氏菌菌株由中国医学细菌菌种保藏管理中心提供。金黄色葡萄球菌与大肠杆菌均由本实验室维持培养, 并已用于多项其他的研究项目。LB (Luria-Bertani) 固体培养基, 北京奥博星生物技术有限责任公司。基因组提取试剂盒, 天根生化科技 (北京) 有限公司。N,N'-二异氰酸亚胺、1-羟基苯并三唑、哌啶、二甲基亚砜、二甲基甲酰胺, 北京化工厂。

1.2 仪器与设备

全基因组可视化分析平台 (操作系统: Ubuntu 22.04, CPU: Xeon Scalable Processors, GPU: NVIDIA DGX A100), Dell 公司。酶标仪, 香港 Thermo scientific 公司。高通量测序平台 Illumina NextSeq 550, 美国 Illumina 公司。

1.3 实验方法

1.3.1 数据收集与清洗

AMP 数据: 来自公开可用的 AMP 数据库, 如 APD (Antimicrobial Peptide Database)、CAMP (Collection of Anti-Microbial Peptides) 和 DBAASP (Database of Antimicrobial Activity and Structure of Peptides) 等数据库以及先前研究提供的 AMP 数据。AMP 数据中涵盖了不同来源的各种 AMP 序列^[11-17], 确保了 AMP 数据的多样性。使用 CD-HIT 程序去除数据冗余后, 得到 1 个包含 3600 条 AMP 序列的阳性数据集。这些序列的长度从 13~154 个氨基酸不等。排除了长度超过 100 个氨基酸的序列, 最终得到 3167 个非冗余的 AMP 序列, 序列氨基酸长度具有特定的分布。随机选取 1500 个 AMP 数据组成训练数据集, 剩余数据组成独立的测试数据集。

非 AMP 数据: 在 Uniprot 数据库中设置“亚细胞定位”过滤器为细胞质, 并移除包括抗菌、抗生素、抗病毒、抗真菌、效应物或分泌关键词匹配的条目。选择长度小于 100 个氨基酸的肽段, 去除数据冗余。去除冗余序列后, 与 AMP 数据集进行比较, 删除两者之间的公共序列。最终, 保留了 3167 个非抗菌肽序列。随机选择 1500 个序列组成非 AMP 的训练数据集, 剩余数据组成阴性测试数据集。

宏基因组数据: 基因组数据来源于云南省温带森林区域次表层的土壤样本。这是一种肥沃的土壤类

型，具有良好的排水与保水能力，该地区周围有一定的草本植物和灌木覆盖。使用 Illumina NextSeq 550 进行测序，得到 1 个包含 123,092 个序列的基因组文件。对于基因组数据，使用蛋白质模拟消化软件，使用胰蛋白酶、胃蛋白酶和蛋白酶 K 模拟酶消化。确保小型开放阅读框 (small open reading frame, sORF) 中的数据是具有活性的蛋白质数据。经过训练的模型通过分析 sORF 中的数据，从而筛选出潜在的抗菌肽。软件参数设置为最小碎片质量 750 Da，最大碎片质量 3750 Da。

1.3.2 氨基酸的特征提取

每个氨基酸残基由 1 个 21 位二进制向量表示。这种编码方法将氨基酸的序列信息转换为数字形式，使计算机能够对其进行处理和分析。编码中的每个字节表示氨基酸的不同属性或位置信息，从而使每个氨基酸具有唯一的二进制编码表示。“10000000000000000000”可以被模型读取为丙氨酸，在氨基酸序列编码的基础上，同时编码氨基酸的物理化学性质，包括亲水性、疏水性和电荷。(-1, 0, +1) 代表电荷，(-3.5, 1.8, -4.5) 表示氨基酸的两亲性，最终用这些数值构成了一个特征向量。并将这些向量组合成 BLOSUM62 (Blocks Substitution Matrix 62) 矩阵。

1.3.3 预测模型组建

通过设计卷积神经网络分类器实现更精确的抗菌肽预测。该分类器由输入层、输出层、池化层、2 个卷积层、3 个密集的全连接层、1 个展开层和 1 个丢弃层组成。卷积层使用 3×3 矩阵进行操作，并采用 2×2 的最大池化层以降低维度。为了减轻过拟合，使用了丢弃率设置为 0.5 的丢弃层。在模型中，第一个卷积层的设置如下：“filters” 参数设置为 64，表示这一层有 64 个卷积核。每个核负责检测输入数据中的不同特征。“kernel_size” 参数设置为 3，意味着每个卷积核的大小为 3。这表明每次卷积操作考虑输入序列中连续 3 个时间步长的窗口。“padding” 参数设置为 “same”，表示使用填充来保持输入和输出大小一致。

第 2 个卷积层配置了 32 个大小为 3 的卷积核，并使用填充以确保输出大小与输入匹配。第 3 个卷积层由 64 个大小为 3 的卷积核组成，也采用了填充。每个卷积核在输入序列上滑动并执行卷积操作，从而捕捉序列中的各种模式和特征。

3 个卷积层堆叠起来，从输入序列中提取局部特征。每个卷积核在输入序列上滑动并进行卷积操作以捕捉不同的模式和特征。卷积层的层次结构使模型能够自动从数据中学习抽象特征，这些特征在随后的全连接层中进一步处理，以支持分类任务。通过这种方式，卷积神经网络擅长处理序列数据。若将模型的输入样本设为 x ，那么卷积层中的操作见式 (1)~式 (3) 如下：

$$con(x) = \sum_{p=0}^{p-1} \sum_{q=0}^{Q-1} (w_{pq}^k) x_{i+pq} \quad ; \quad (1)$$

$$Relu(x) = \max(0, x) \quad ; \quad (2)$$

$$LeakyRelu(x) = \begin{cases} ax, & x < 0 \\ x, & x \geq 0 \end{cases} \quad \circ \quad (3)$$

式 (1)~式 (3) 中， W 表示卷积核的大小； $W^k = (w_{pq}^k)_{p \times q}$ 表示第 k 个卷积核的权矩阵； $P \times Q$ 表示矩阵的大小。

LSTM 模型包含 1 个输入层、1 个包含 16 个 LSTM 单元的 LSTM 层、1 个展开层、1 个密集层和 1 个输出层。模型从输入层接收序列特征数据，并处理这些数据以提取重要信息。在 LSTM 层之后，使用展开层将 LSTM 层的输出展平为一维向量。在展平层之后是包含 10 个神经元的密集层。密集层接收展平层的输出作为输入，并执行线性变换和激活函数以生成特征向量。最后，输出层包含 2 个神经元，使用 Sigmoid 激活函数。输出层用于二元分类任务，提供 2 个类别的概率分布。注意力模型只是将 LSTM 模型中的 LSTM 层替换为注意力层。

1.3.4 模型性能评估

通过结合长短期记忆模型、注意力模型和卷积神经网络来构建一个用于预测抗菌肽的集成深度学习框架。该框架的 CNN 层将被设计用来有效捕获输入数据的局部特征，而 LSTM 层则将处理序列的全局

信息。这种组合将使得模型能够同时考虑本地和全局信息，从而更全面地理解输入数据。注意力机制的加入使模型能够在序列处理过程中更多地关注与抗菌肽相关的关键特征，从而提高效率和性能。同时，引入准确率、召回率、特异性等[式(4)~式(9)]参数对模型的性能进行验证。

$$\text{准确率} = \frac{(T_n + T_p)}{(T_n + T_p + F_n + F_p)} \times 100\% \quad ; \quad (4)$$

$$\text{召回率} = \frac{T_p}{(T_p + F_n)} \times 100\% \quad ; \quad (5)$$

$$\text{特异性} = \frac{T_n}{(T_n + F_p)} \times 100\% \quad ; \quad (6)$$

$$\text{精确度} = \frac{T_p}{(T_p + F_p)} \times 100\% \quad ; \quad (7)$$

$$F1_{\text{score}} = 2 \times \frac{(\text{精确度} \times \text{召回率})}{\text{精确度} + \text{召回率}} \times 100\% \quad ; \quad (8)$$

$$\text{马修斯相关系数} = \frac{T_p \times T_n - F_p \times F_n}{\sqrt{(T_p + F_p)(T_p + F_n)(T_n + F_p)(T_n + F_n)}} \times 100\% \quad (9)$$

式(4)~式(9)中， T_n 为真阴性； T_p 为真阳性； F_n 为假阴性； F_p 为假阳性。

1.3.5 土壤中抗菌肽和微生物群的丰度分析

利用 Kraken 2 软件将宏基因组数据与细菌数据库进行比对，确定宏基因组数据中基因数据的来源及相应的细菌丰度。通过将 sORF 数据和预测结果与原始基因进行比对，最终确定能产生特定肽段的细菌及其丰度。

1.3.6 分子对接设计

利用 Autodock Vina 软件进行分子对接。在 Vina 可视化操作平台中，对蛋白质进行脱水和加氢处理，并将结果保存为 pdbqt 格式。箱体盒子被放置在蛋白质的活性口袋上，每个肽经过 8 次对接，最终选择结合能绝对值最大的构象作为最终结果。随后利用 Discovery Studio 获得配合物的三维对接结构，包括氢键、电荷分布和疏水相互作用。

1.3.7 分子动力学模拟

采用 GROMACS 2020.3 软件进行分子动力学模拟。利用 amber99sbildn 力场和通用琥珀力场(GAFF)分别生成蛋白质和配体的参数和拓扑结构。将模拟盒子的尺寸优化，使蛋白质的每个原子与盒子的距离大于 1.0 nm。然后，以密度为 1 的水分子填充盒子。为保证模拟系统为电中性，用 Cl^- 和 Na^+ 离子取代水分子。采用最陡下降法，对 5.0×10^4 步骤进行能量优化，使整个系统的能量消耗最小化，最终减少整个系统中不合理的接触或原子重叠。在能量最小化之后，使用 NVT 组件在 300 K 下进行 100 ps 的第一相平衡，以稳定系统的温度。在 1 bar 和 100 ps 的条件下，用 NPT 系统模拟第二相平衡。模拟的主要目标是优化目标蛋白与溶剂和离子之间的相互作用，使模拟系统完全预平衡。所有的 MD 模拟都是在温度为 300 K，压力为 1 个大气压的等温和等静力集合下进行的，速度为 30000 ps。温度和压力分别采用 V-rescale 和 Parrinello-Rahman 方法控制，温度和压力耦合常数分别为 0.1、0.5 ps。采用 Lennard-Jones 函数计算范德华力，非键截断距离设为 1.4 nm。所有原子的键长由 LINCS 算法约束。采用粒子网格-埃瓦尔德法计算了远距离静电相互作用，傅里叶间距为 0.16 nm。在经过 100 ns 的模拟后，对结果进行均方根偏差、旋转半径、均方根波动和溶剂可及表面积计算。

1.3.8 固相合成实验

采用 Fmoc (Fmoc chemistry strategy) 化学策略对筛选结果进行固相合成^[18]。使用装有 Fmoc 的固相树脂对 C-末端氨基酸进行保护，在体积分数为 20% 的哌啶溶液中去保护。 $\text{N,N}'$ -二异氰酸亚胺与 1-羟基苯并三唑 (DIC-HOBT) 作为活化试剂促使氨基酸之间完成耦合，耦合反应均持续 1h，完成后使用二甲亚砜 (DMSO) 和二甲基甲酰胺 (DMF) 交替洗涤。去保护和耦合反应结束后对固相载体进行洗涤，去除未反应的氨基酸和副产物。重复耦合-洗涤过程直至多肽合成完毕。

1.3.9 抗菌活性验证

选取肺炎克雷伯氏菌、大肠杆菌和金黄色葡萄球菌作为验证对象。3 种菌株均被复苏于 LB 固体培养基平板上,并在 37℃ 条件下孵育 16~18h,以确保充分生长。选取单个健康菌落,转接至液体 LB 培养基中,并在 37℃、220r/min 的摇床上振荡培养至对数生长期^[19]。

采用双缩脲法测定饱和肽液质量浓度后^[20],使用去离子水将肽液分别稀释至 125.0、90.5、70.5、50.5、31.5、0 μg/ mL 的溶液并编号为 1~6 组。使用 0.22μm 的滤膜过滤除菌后取 1mL 加入到液体 LB 培养基中,与菌液共同孵育。使用酶标仪测定吸光度数值,波长设置为 600nm,每 4h 测定 1 次菌液的浑浊程度并记录 OD_{600} 。通过绘制光密度曲线,从而测定肽液对于 3 种细菌的最小抑菌浓度 (minimal inhibitory concentration, MIC)。通过计算菌落总数 (CFU/mL) [式 (10)],从而研究不同多肽质量浓度对细菌生长的影响。所有实验均在无菌条件下进行,避免交叉污染。实验采用的菌株培养条件及操作细节均依照国际标准化操作程序执行。

$$\text{菌落形成单位} = \frac{(\text{菌落数} \times \text{总稀释因子})}{(\text{接种培养基的体积})} \quad (10)$$

2 结果与分析

2.1 多模型对比分析

图 1 展示了不同模型在相同独立测试数据集上的曲线下面积 (receiver operating characteristic curve, ROC)。ROC 通过描绘真正例率 (true positive rate, TPR) 和假正例率 (false positive rate, FPR) 之间的关系来展示分类器在不同阈值下的表现。结果表明,本研究提出的 DMLAM 模型取得了较好的预测结果 AUC 为 0.99。与以往的深度学习模型相比,如 CAMP-ANN、AMPlify 和 iAMP。DLMAMP 在数据预处理和特征提取阶段采用了新颖的组合方式,提高了模型对于复杂数据结构的理解能力,从而提升了整体的预测精度。通过优化网络结构和训练策略,模型在计算上更为高效,降低了对硬件的依赖,使得模型在不同的环境下都能够稳定运行。针对过拟合问题,正则化技术的引入有效提升了模型的泛化能力。与传统机器学习模型,如 ClassAMP 和 CAMP 相比,DMLAMP 在处理大规模数据集和识别复杂模式方面具有明显的优势。传统模型往往依赖于手动提取特征,而 DMLAMP 能够自动学习和提取高级特征,这提高了模型处理复杂数据的能力,这在处理高维度数据时尤为重要,能够显著提升预测的准确性和模型的鲁棒性。通过独立测试集,进行 10 倍交叉测试,测试 9 种经典抗菌肽预测方法的性能 (表 1) 发现,本研究提出的模型达到了 96.5% 的准确率、98.7% 的精度和 99.2% 的特异性,均优于其他预测模型。这些结果表明,本模型是完全可行和有效的。

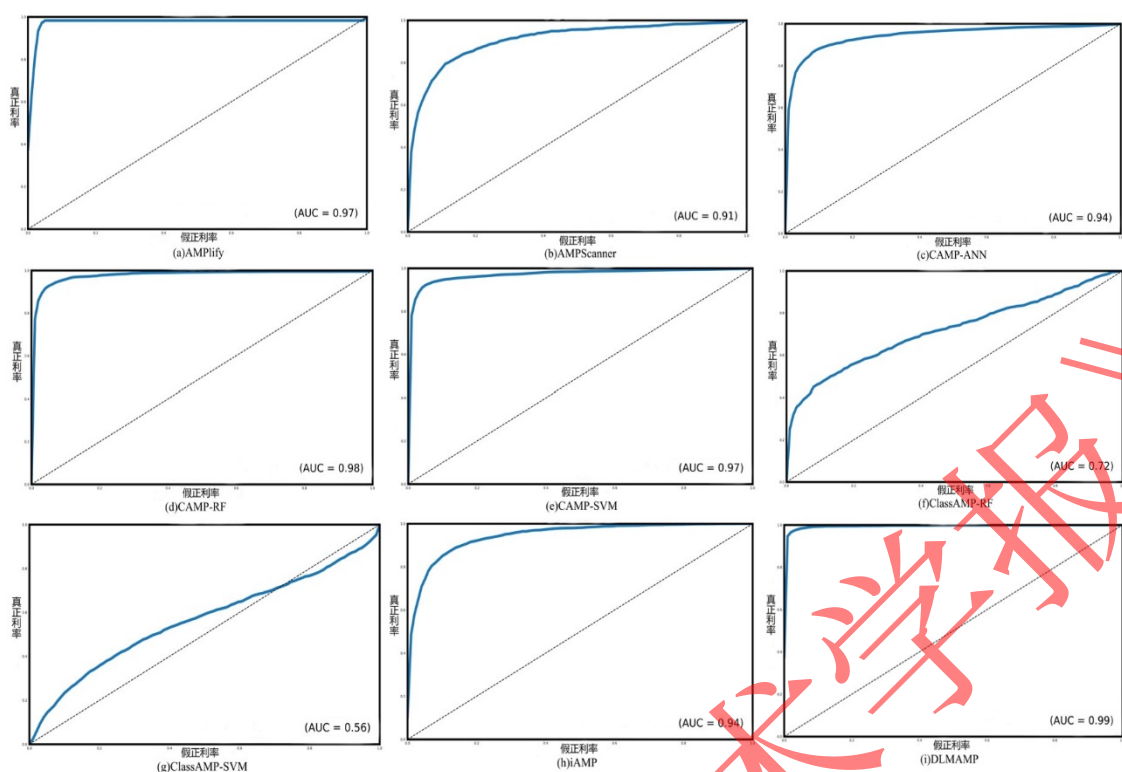


图 1 同一检验数据集下 9 个模型的 ROC 曲线
Fig.1 ROC curves of 9 models under same test data set

表 1 9 种经典抗菌肽预测方法的性能比较
Tab. 1 Comparison of performance of nine classical antimicrobial peptide prediction methods

模型	准确率	精准度	召回率	F1-score	特异性	马修斯相关系数	%
ClassAMP-SVM	36.8	37.1	98.3	53.8	0.1	9.9	
ClassAMP-RF	81.6	82.8	97.9	89.7	5.7	8.6	
CAMP-RF	93.8	93.7	93.9	93.8	93.7	78.2	
CAMP-SVM	93.6	95.1	92.0	93.5	95.2	87.2	
CAMP-ANN	89.1	91.2	96.4	88.8	91.2	78.3	
AMPscanner	79.0	55.5	88.9	68.3	75.5	57.2	
iAMP-2L	81.7	68.6	94.1	79.3	74.2	66.1	
AMPLify	95.3	95.1	92.2	93.6	92.3	89.9	
DLMAMP	96.5	98.7	91.9	95.2	99.2	92.6	

2.2 模型预测结果分析

将小型开放阅读框数据与抗菌肽数据进行比对，去除已知的抗菌肽序列，剔除冗余序列后得到 1 个包含 2 537 905 个序列的待预测数据集。将该数据集输入到模型中，得到了需要实验验证的 258 749 个序列。随后将模型评分在 0.90~0.99 之间的 312 个多肽使用 ToxinPred141 (<http://crdd.sodd.net/raghava/toxinpred>) 和 HemoPred (<http://codes.bio/hemopred>) 在线预测工具来评估模型输出结果的毒性和溶血潜力。最终得到了 9 个既无毒性也无溶血性的多肽，分别是：“Arg-Lsy-Gln-Gly-Phe-Arg-Arg; Met-Thr-Lys-Trp-Trp-Ala; Ser-Phe-Arg-Arg-Phe-Thr-Tyr-Gly-Asn-Leu-Val-Thr-Thr-Phe-Thr-

Ser-Ser-Lys; Val-Pro-Tyr-Leu-Leu-Leu-Pro-His-Ser-Val-Ser-Ala-Leu-Val-Leu-Thr-Phe-Arg-Arg; Arg-Val-Ala-Lys-Leu-Pro-Lys-Pro-Leu; Leu-Cys -Leu-Ile-Gln-Asn-Ile- Lys-Thr- Phe-Asn-Asn-Gly-Ser-Leu-Gly-Ser-Arg-Ile-Asp-Glu-Glu-Arg-Ser-Ile-Gly; Ala-Thr-Asn-Ile-Pro-Phe-Lys-Val-His-Phe-Arg-Cys-Lys-Ala-Ala-Phe-Cys; Asn-Phe-Gln-Gln-Arg-Ile-Ser-Trp-Phe-Ser-His-Arg” 记为 P1~P9，进行后续的分子对接与分子动力学模拟验证。

使用 Kraken2 软件对所有原始数据进行注释，将预测结果与原始数据进行对比。结果显示，筛选出的多肽主要来源于荧光假单胞菌、枯草芽孢杆菌、空间链霉菌、黑曲霉、哈茨木霉、阿维菌素链霉菌、根霉菌、大肠杆菌等土壤微生物。图 2 展示了筛选结果（P1~P9）与一组细菌种群（如图右侧所列）之间的关系映射。通过桑基图可视化分析，可以清晰地观察到 9 种筛选结果与一种或多种细菌之间的丰度差异，流的宽度代表了相应的肽在不同细菌种群中相对丰度的高低，P1 在荧光假单胞菌中有较高的丰度，P2 在哈茨木霉菌中有较高的丰度，P3 在根内丛枝菌中有较高的丰度，P4、P7 和 P9 在大肠杆菌中有较高的丰度，P5 和 P6 在黑曲霉中有较高的丰度，P8 在马弗链霉菌中有较高的丰度。最右侧的百分含量代表了特定的细菌种群 P1~P9 的总丰度。从而提供了不同种群中抗菌活性潜力的初步指标。这种表示法揭示了各肽类在不同细菌种群中的分布格局，有助于进一步筛选和验证具有潜在抗菌效力的肽类分子。

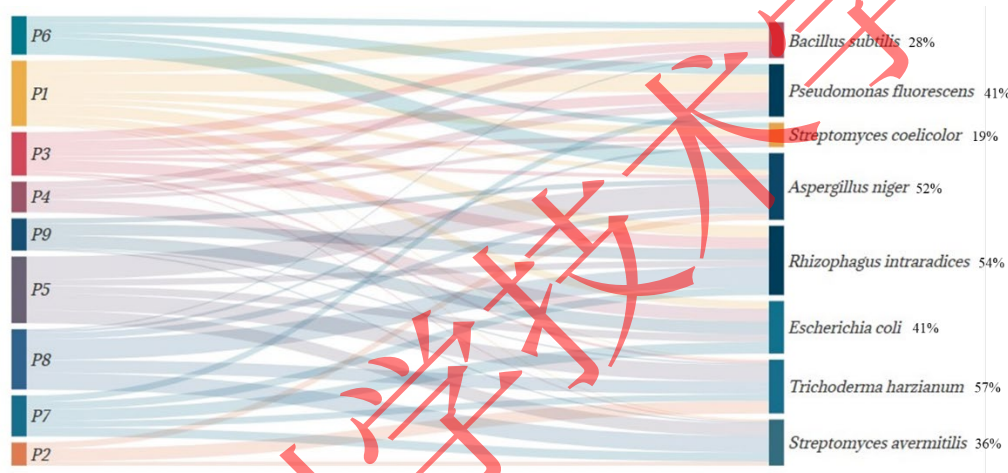


图 2 9 个筛选结果的相对丰度

Fig.2 The relative abundance of the 9 screened results.

2.3 分子对接结果分析

生物膜是一种由细菌分泌的黏性物质构成的保护层，能够让细菌群体在各种环境条件下存活，并增强其抗药性^[21-22]。理解抗菌肽如何破坏这一防御机制是开发新型抗感染策略的关键。目前已有部分抗菌肽被发现能够干扰生物膜的形成，它们可通过影响细菌的群体感应系统，这是一种细菌用于感知和响应细胞密度的信号系统，来减少细菌的生物膜形成^[23]。一些抗菌肽还可以直接与生物膜成分相互作用，破坏其结构或干扰生物膜内的水通道和营养物质传输。此外，抗菌肽也可能通过抑制生物膜相关的酶活性或细菌粘附分子的表达来阻止生物膜的形成和维护^[24]。研究表明，调节 C2E 水平或阻断其与 MrkH 等关键蛋白的相互作用，能够显著影响生物膜的稳定性和功能^[25]。

本研究对肽与 MrkH 蛋白分子的相互作用进行了评估。使用 Autodock Vina 将模型中得分最高的肽对接到 MrkH 蛋白(PDB ID:5KEC)的活性口袋中，结果见表 2。将所有对接结果与 MrkH 蛋白的原配体进行比较。由表 2 可知，多肽 P1、P4、P5、P9 的结合能为-9.082、-10.882、-10.624 和-9.213kcal/mol，与 C2E 配体（-8.800kcal/mol）相比具有更好的结合能力^[26-27]。故后续验证中选取评分最高的 P4 多肽进行验证。图 3 展示了肽-蛋白相互作用与原有配体-蛋白相互作用的可视化对比，其中突出了目标蛋白质-配体复合体的分子结合位点和电荷表面电位。图 3（a）~（b）二维展示了原有配体和 P4 多肽与蛋白质活性位点关键氨基酸的相互作用。（c）~（d）则展示了配体-蛋白质复合体的三维分子模型。由图 3（a）可知，MrkH 蛋白与 C2E 配体之间的相互作用主要通过氢键介导。位于配体一侧的残基 LYS 63 和 ARG

65 通过氢键与氨基和磷酸基团相互作用；另一侧，残基 PHE 208、PRO 112 和 ASP 143 也通过氢键与氨基和碳原子相互作用。可以推断出，肽与蛋白质的结合主要是由氢键、静电相互作用和疏水相互作用驱动的，并且氢键起主导作用。总的来说 P4 与其他肽相比具有更强的结合能。由图 3（b）可知，多肽 P4 与 MrkH 蛋白结合能可能受到自身与蛋白结合袋之间氢键数量的影响^[28-29]。考虑到化合物的结构和性质，多肽和 C2E 之间的结合位点存在差异是可以预料的^[30]。

表 2 安全且具有抗生物膜特性的高分多肽
Tab. 2 Safe high-scoring polypeptide with antibiofilm properties

编号	序列	氨基酸数量/个	溶血性	毒性	模型评分	结合能/kcal • mol ⁻¹
1	Arg-Lsy-Gln-Gly-Phe-Arg-Arg	7	无	无	0.91	-9.082
2	Met-Thr-Lys-Trp-Trp-Ala	7	无	无	0.83	-6.796
3	Ser-Phe-Arg-Arg-Phe-Thr-Tyr-Gly-Asn-Leu-Val- Thr-Thr-Phe-Thr-Ser-Ser-Lys	18	无	无	0.95	-8.386
4	Gly-Thr-Ala-Trp-Arg-Trp-His-Tyr-Arg-Ala-Arg-Ser	12	无	无	0.91	-10.882
5	Val-Pro-Tyr-Leu-Leu-Leu-Pro-His-Ser-Val-Ser- Ala-Leu-Val-Leu-Thr-Phe-Arg-Arg	19	无	无	0.92	-10.624
6	Arg-Val-Ala-Lys-Leu-Pro-Lys-Pro-Leu Leu-Cys-Leu-Ile-Gln-Asn-Ile-Lys-Thr-Phe-Asn-	9	无	无	0.93	-6.639
7	Asn-Gly-Ser-Leu-Gly-Ser-Arg-Ile-Asp-Glu-Glu-Arg- Ser-Ile-Gly	26	无	无	0.88	-5.837
8	Ala-Thr-Asn-Ile-Pro-Phe-Lys-Val-His-Phe-Arg- Cys-Lys-Ala-Ala-Phe-Cys	17	无	无	0.85	-6.055
9	Asn-Phe-Gln-Gln-Arg-Ile-Ser-Trp-Phe-Ser-His-Arg	12	无	无	0.82	-9.213

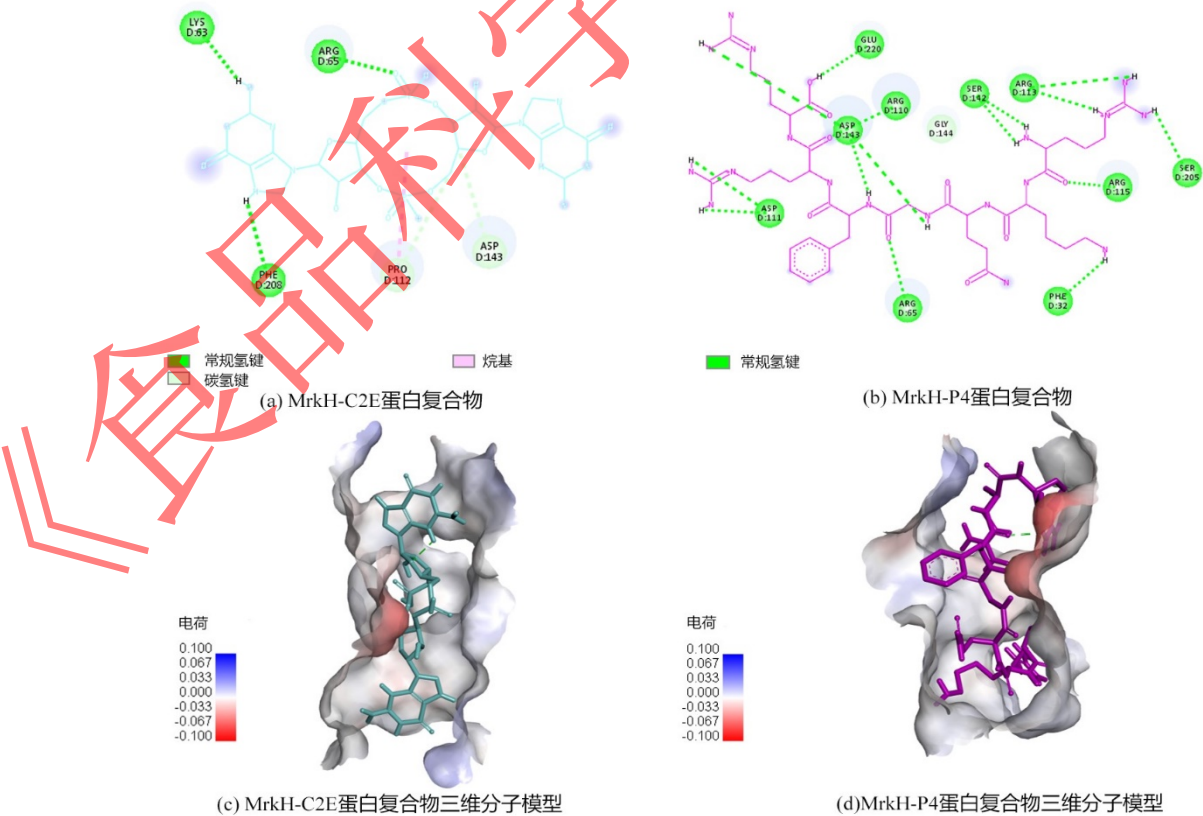


图 3 多肽 P4 和 C2E 与 MrkH 蛋白相互作用示意图

Fig.3 Schematic illustration of the interaction between peptides P4 and C2E with MrkH protein.

2.4 分子动力学模拟结果分析

用 GROMACS 2020.3 软件对对接结果中亲和力最高的 P4 进行动力学模拟, 并与原有配体 C2E 的模拟结果进行对比, 结果见图 4。研究采用均方根偏差(RMSD)来量化结构偏差, 当蛋白质构象偏离其初始结构时, 这种变化的程度可以很容易地在分子动力学模拟过程中通过 RMSD 推断出来。较小的 RMSD 表明配体-受体结合稳定, 而较大的 RMSD 可能表明显著的结构变化或不稳定^[31]。由图 4 (a) 可知, C2E 与 MrkH 蛋白的复合物 (MrkH-C2E) 和 P4 与 MrkH 蛋白的复合物 (MrkH-P4) 二者的曲线都表现出相对的稳定性, 在整个模拟过程中只有很小的波动, 这通常表明主要的分子结构没有发生变化^[32-34]。复合物 MrkH-C2E 的平均 RMSD 为 0.44nm, 复合物 MrkH-P4 的平均 RMSD 为 0.52nm, P4 与 MrkH 结合引起的结构调整比原配体稍大。但在整个 100 ns 的 RMSD 实验结果显示, MrkH-P4 复合物的轨迹处于稳定状态, 与 MrkH 结合位点有很强的联系。且在 100ns 内 RMSD 的数值没有迅速上升, 这表明模拟的分子并没有迅速偏离其初始结构^[35-36]。

均方根波动(RMSF)通过测量轨道内原子位置的标准偏差[图 4(b)], 来确定蛋白质中单个残基的平均柔韧性。P4 和 C2E 的 c 域(残基 192~198)有较大的波动。观察该区域的结构柔韧性发现, P4 和 C2E 的 RMSF 非常相似, ASP192 残基的 RMSF 值分别为 0.44、0.38nm; 残基 62 处, P4 和 C2E 的 RMSF 值分别为 0.31、0.37nm。

为了研究分子动力学模拟过程中蛋白质表面与溶剂水的有效接触面积, 计算溶剂可及表面积(solvent accessible surface area, SASA)参数[图 4(c)]。通过比较 SASA 发现, 在 100ns 分子动力学模拟过程中, P4 与 MrkH 蛋白形成了高度稳定的复合物, 类似于 MrkH-C2E 蛋白复合物的稳定性^[37]。

旋转半径(R_g)可进一步验证复杂结构的稳定性、紧凑性、可折叠性和展开能力。整个 100ns 模拟过程中沿着分子对接轨迹对两种复合物进行了分析, 见图 4 (d)。C2E 的 R_g 平均值为 2.083, P4 的平均值为 2.080, 这一接近的数值表明两种复合物在空间结构的紧凑性或整体构型上可能非常相似。半径体积反映的是复合物在溶液中尺寸的指标, 较小的 R_g 值通常意味着结构更为紧凑。因此, 几乎相同的 R_g 值表示 P4 和 C2E 在与 MrkH 蛋白相互作用时, 可能采取了类似的三维形状, 这点表明它们可能具有相似的结合特性和潜在的生物效应。

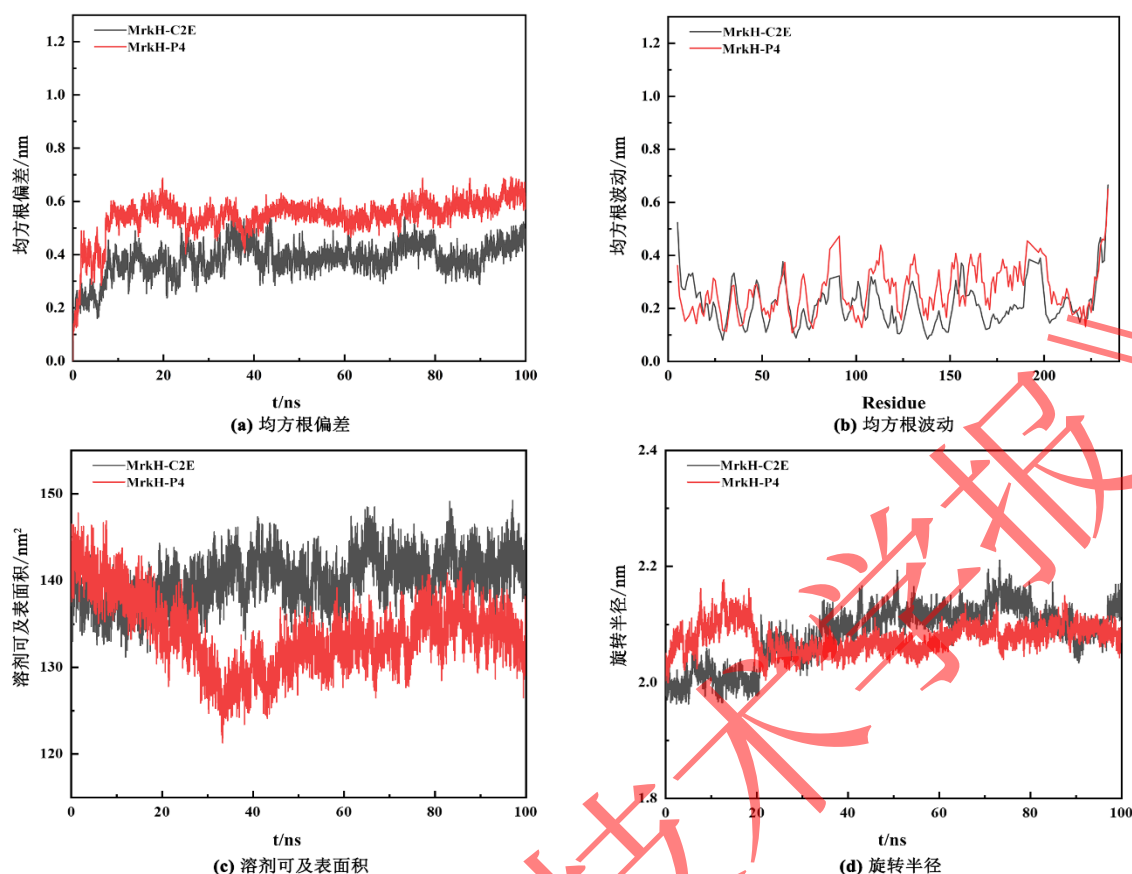


图 4 多肽 P4 和 C2E 与 MrkH 蛋白配合物的分子模拟结果

Fig.4 Molecular simulation analyses of complexes over 100 ns simulation time include

2.5 抗菌肽抑菌活性分析

不同质量浓度的 P4 多肽溶液的抑菌活性见图 5。由图 5 可知，该多肽不仅对肺炎克雷伯菌具有抑制作用，同时还对金黄色葡萄球菌和大肠杆菌具有抑制作用。其中 1 组抑菌圈直径最大，分别为 13.8mm，12.5mm，13.1mm。添加不同质量浓度的 P4 多肽溶液后不同菌液的光密度和菌落总数见图 6。由图 6 可知，随着 P4 质量浓度的降低，受试菌被抑制的效果亦逐渐减弱。P4 对肺炎克雷伯氏菌、大肠杆菌和金黄色葡萄球菌的最小抑菌浓度分别为 31.5、50.5、70.5 $\mu\text{g}/\text{mL}$ 。最小抑菌浓度的抑菌圈为第 5 组，抑菌圈直径为 5.1mm，6.6mm 和 8.1mm。通过观测添加 P4 溶液的菌液的 CFU 和 OD_{600} 的相关性发现，不同的质量浓度下的多肽溶液对细菌生长产生了不同的影响。由图 6 (a1)、(a2) 可以看出，未添加 P4 溶液的肺炎克雷伯氏菌整个生长阶段的 OD_{600} 在 0.4~2.5 内波动，细菌悬浮液的浓度与 OD_{600} 正相关，且在 14h 处到达对数生长期，20h 后到达稳定状态。由图 6 (b1)、(b2) 可知大肠杆菌整个生长阶段的 OD_{600} 在 0.4-0.9 内波动，4h 进入对数生长期，24h 后到达稳定状态。由图 6 (c1)、(c2) 金黄色葡萄球菌整个生长阶段的 OD_{600} 在 0.4-1.1 内波动，10h 进入对数生长期，20h 后达到稳定状态。添加了 P4 溶液后的菌液均展现出不同程度上的生长抑制。综合 OD_{600} 、CFU 计数与 MIC 结果，P4 多肽显示出了一定的抗菌活性。

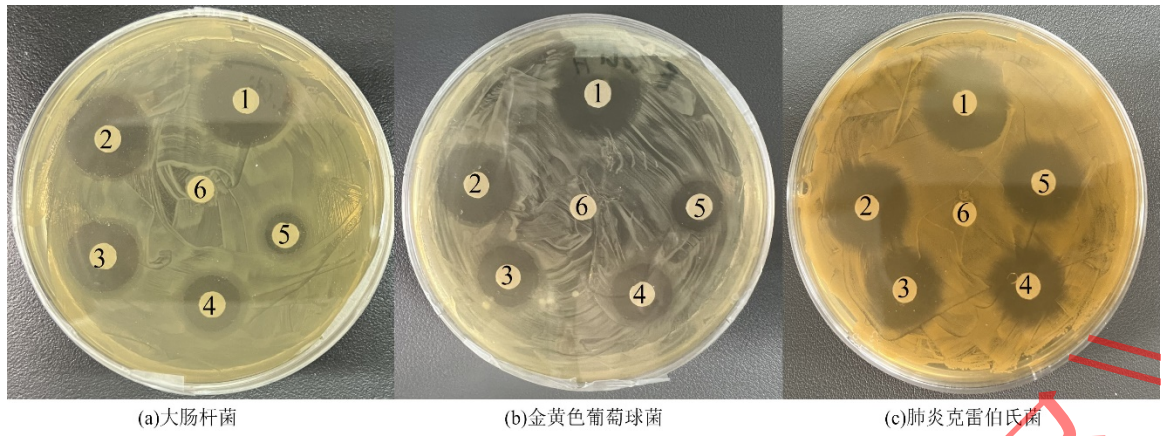


图 5 P4 的抑菌圈实验结果
Fig.5 The antibacterial zone test results for P4

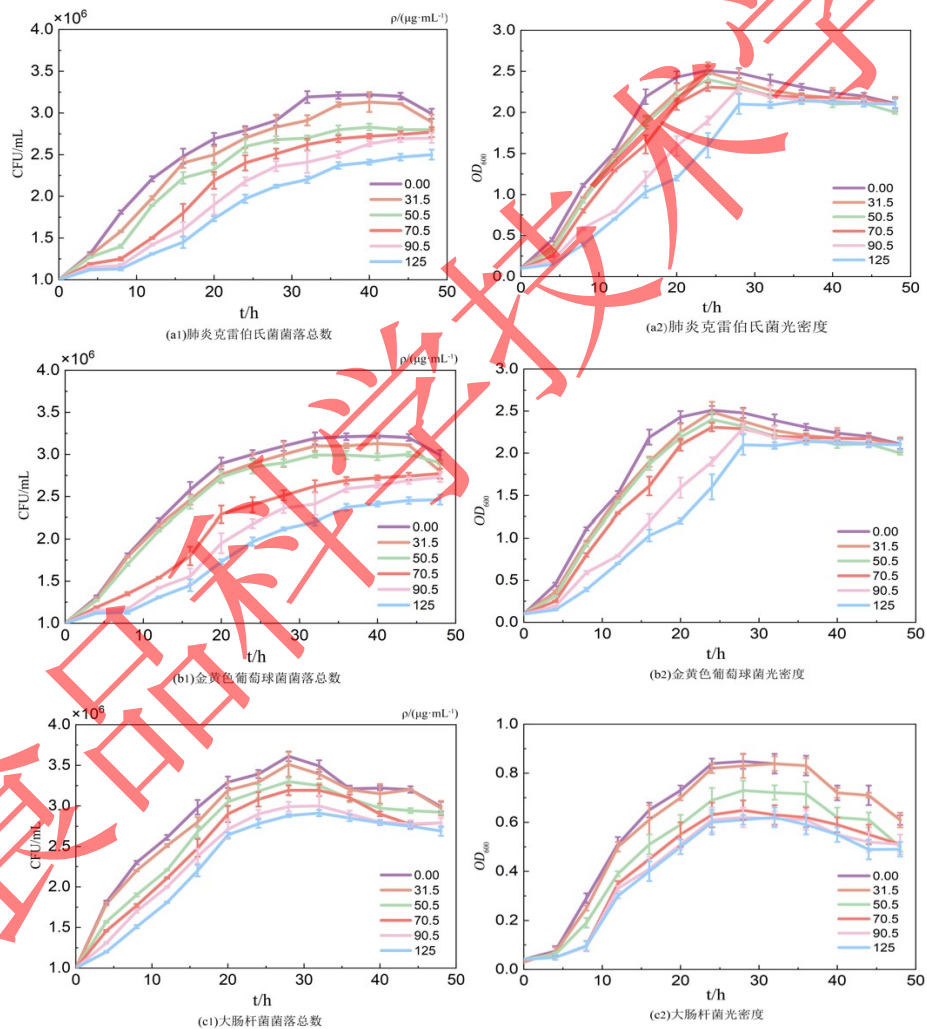


图 6 添加不同质量浓度的 P4 多肽溶液后菌液的光密度和菌落总数
Fig.6 The optical density and the total number of colonies of bacterial solution after adding P4 polypeptide solution with different mass concentrations

3 结 论

在快速发展的抗菌研究领域,本文介绍了一种预测抗菌肽的新模型,该模型结合了 LSTM, CNN, Attention 三个独立的深度学习分类器,在同一测试数据集下与其它预测模型相比,本模型精准度达到 98.7%、准确率 96.5%、召回率 91.9%、F1-score 95.2%、特异性 99.2%。应用该模型成功地筛选出了多个具有抗菌潜力的多肽。本研究通过分子对接与分子动力学模拟深入分析了 P4 “Gly-Thr-Ala-Trp-Arg-Trp-His-Tyr-Arg-Ala-Arg-Ser” 该多肽的结合特性。结果显示,该多肽展示出了比 MrkH 蛋白原有配体更强的结合力,而且通过分子动力学的模拟也证明了这一结果。通过抗菌活性检测,该多肽不仅对肺炎克雷伯氏菌表现出抗菌活性,最小抑制浓度为 31.5 $\mu\text{g}/\text{mL}$ 。同时也对大肠杆菌和金黄色葡萄球菌展现出不同的抑制作用。此外,抑菌圈实验的测试结果进一步证实了 P4 对细菌生长产生的影响。因此本研究提供的高通量预测筛选方法,为在不同生境条件下微生物菌群中筛选抗菌肽的研究提供参考,具有一定的理论和实践意义。

参考文献:

- [1] BAHAR A A, REN D C. Antimicrobial peptides[J]. *Pharmaceuticals (Basel, Switzerland)*, 2013, 6(12): 1543-1575.
- [2] REARDON S. Antibiotic resistance sweeping developing world[J]. *Nature*, 2014, 509: 141-142.
- [3] ZHANG Q Y, YAN Z B, MENG Y M, et al. Antimicrobial peptides: mechanism of action, activity and clinical potential[J]. *Military Medical Research*, 2021, 8(1): 48.
- [4] BROGDEN K A. Antimicrobial peptides: pore formers or metabolic inhibitors in bacteria?[J]. *Nature Reviews Microbiology*, 2005, 3: 238-250.
- [5] NUTI R, GOUD N S, SARASWATI A P, et al. Antimicrobial peptides: a promising therapeutic strategy in tackling antimicrobial resistance[J]. *Current Medicinal Chemistry*, 2017, 24(38): 4303-4314.
- [6] KAMYSZ W, OKRÓJ M, ŁUKASIAK J. Novel properties of antimicrobial peptides[J]. *Acta Biochimica Polonica*, 2003, 50(2): 461-469.
- [7] 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.
- [8] RÖMLING U, GOMELSKY M, GALPERIN M Y. C-di-GMP: the dawning of a novel bacterial signalling system[J]. *Molecular Microbiology*, 2005, 57(3): 629-639.
- [9] 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.
- [10] 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.
- [11] LI C K, SUTHERLAND D, HAMMOND S A, et al. AMPLify: attentive deep learning model for discovery of novel antimicrobial peptides effective against WHO priority pathogens[J]. *BMC Genomics*, 2022, 23(1): 77.
- [12] 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.
- [13] THOMAS S, KARNIK S, BARAI R S, et al. CAMP: a useful resource for research on antimicrobial peptides[J]. *Nucleic Acids Research*, 2010, 38(Database issue): D774-D780.
- [14] GOGOLADZE G, GRIGOLAVA M, VISHNEPOLSKY B, et al. DBAASP: database of antimicrobial activity and structure of peptides[J]. *FEMS Microbiology Letters*, 2014, 357(1): 63-68.
- [15] KAVOUSHI K, BAGHERI M, BEHROUZI S, et al. IAMPE: NMR-assisted computational prediction of antimicrobial peptides[J]. *Journal of Chemical Information and Modeling*, 2020, 60(10): 4691-4701.
- [16] 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.

- [17] JOSEPH S, KARNIK S, NILAWE P, et al. ClassAMP: a prediction tool for classification of antimicrobial peptides[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2012, 9(5): 1535-1538.
- [18] HUANG J J, XU Y C, XUE Y F, et al. Identification of potent antimicrobial peptides via a machine-learning pipeline that mines the entire space of peptide sequences[J]. Nature Biomedical Engineering, 2023, 7: 797-810.
- [19] DAS P, SERCU T, WADHAWAN K, et al. Accelerated antimicrobial discovery via deep generative models and molecular dynamics simulations[J]. Nature Biomedical Engineering, 2021, 5: 613-623.
- [20] HILPERT K, VOLKMER-ENGERT R, WALTER T, et al. High-throughput generation of small antibacterial peptides with improved activity[J]. Nature Biotechnology, 2005, 23: 1008-1012.
- [21] SINTIM H O, SMITH J A I, WANG J X, et al. Paradigm shift in discovering next-generation anti-infective agents: targeting quorum sensing, c-di-GMP signaling and biofilm formation in bacteria with small molecules[J]. Future Medicinal Chemistry, 2010, 2(6): 1005-1035.
- [22] TAN J W H, WILKSCH J J, HOCKING D M, et al. Positive autoregulation of mrkH by the cyclic di-GMP-dependent MrkH protein in the biofilm regulatory circuit of *Klebsiella pneumoniae*[J]. Journal of Bacteriology, 2015, 197(9): 1659-1667.
- [23] LANGSTRAAT J, BOHSE M, CLEGG S. Type 3 fimbrial shaft (MrkA) of *Klebsiella pneumoniae*, but not the fimbrial adhesin (MrkD), facilitates biofilm formation[J]. Infection and Immunity, 2001, 69(9): 5805-5812.
- [24] AUGUSTYNIAK D, KRAMARSKA E, MACKIEWICZ P, et al. Mammalian neuropeptides as modulators of microbial infections: their dual role in defense versus virulence and pathogenesis[J]. International Journal of Molecular Sciences, 2021, 22(7): 3658.
- [25] ALLEN B L, GERLACH G F, CLEGG S. Nucleotide sequence and functions of mrk determinants necessary for expression of type 3 fimbriae in *Klebsiella pneumoniae*[J]. Journal of Bacteriology, 1991, 173(2): 916-920.
- [26] ARES M A, FERNÁNDEZ-VÁZQUEZ J L, ROSALES-REYES R, et al. H-NS nucleoid protein controls virulence features of *Klebsiella pneumoniae* by regulating the expression of type 3 pili and the capsule polysaccharide[J]. Frontiers in Cellular and Infection Microbiology, 2016, 6: 13.
- [27] ZASLOFF M. An ancient neuropeptide defends the brain against infection[J]. Proceedings of the National Academy of Sciences of the United States of America, 2021, 118(5): e2023990118.
- [28] SCHURTZ T A, HORNICK D B, KORHONEN T K, et al. The type 3 fimbrial adhesin gene (mrkD) of *Klebsiella* species is not conserved among all fimbriate strains[J]. Infection and Immunity, 1994, 62(10): 4186-4191.
- [29] ALCÁNTAR-CURIEL M D, BLACKBURN D, SALDAÑA Z, et al. Multi-functional analysis of *Klebsiella pneumoniae* fimbrial types in adherence and biofilm formation[J]. Virulence, 2013, 4(2): 129-138.
- [30] WANG J M, CIEPLAK P, KOLLMAN P A. How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules?[J]. Journal of Computational Chemistry, 2000, 21(12): 1049-1074.
- [31] WANG C K, SHIH L Y, CHANG K Y. Large-scale analysis of antimicrobial activities in relation to amphipathicity and charge reveals novel characterization of antimicrobial peptides[J]. Molecules, 2017, 22(11): 2037.
- [32] KARACA E, BONVIN A M J J. Advances in integrative modeling of biomolecular complexes[J]. Methods, 2013, 59(3): 372-381.
- [33] WU Q H, PATOČKA J, KUČA K. Insect antimicrobial peptides, a mini review[J]. Toxins, 2018, 10(11): 461.

- [34] MOBLEY D L, DILL K A. Binding of small-molecule ligands to proteins: "what you see" is not always "what you get"[J]. *Structure*, 2009, 17(4): 489-498.
- [35] ARES M A, FERNÁNDEZ-VÁZQUEZ J L, PACHECO S, et al. Additional regulatory activities of MrkH for the transcriptional expression of the *Klebsiella pneumoniae* mrk genes: antagonist of H-NS and repressor[J]. *PLoS One*, 2017, 12(3): e0173285.
- [36] HSU S Y, NAKABAYASHI K, NISHI S, et al. Activation of orphan receptors by the hormone relaxin[J]. *Science*, 2002, 295(5555): 671-674.
- [37] KRIEGER E, VRIEND G. New ways to boost molecular dynamics simulations[J]. *Journal of Computational Chemistry*, 2015, 36(13): 996-1007.

《食品科学技术学报》