张浅阅, 李博文, 李敏,等. 基于支持向量机的抗冠状病毒多肽预测模型[J]. 智能计算机与应用,2026,16(5):126-130.
DOI:10.20169/j.issn.2095-2163.24082601

基于支持向量机的抗冠状病毒多肽预测模型

张浅阅[1], 李博文[2], 李 敏[2], 卢俊坤[1], 王明恩[1], 吴 臻[1], 简勋勋[1], 贺碧芳[2], 龙金金[3]
(1 贵州装备制造职业学院, 贵阳 551400; 2 贵州大学医学院, 贵阳 550025;3 贵阳市口腔医院, 贵阳 550005)

摘 要: 抗冠状病毒多肽是一类特异性针对冠状病毒的多肽,对冠状病毒具有优秀的抑制能力。然而鉴别和筛选潜在的抗冠状病毒多肽面临着数据量大、复杂度高的问题,急需一种精准高效的预测工具来辅助科研人员。本研究基于最新的抗冠状病毒多肽数据库中的信息,构建了一个抗冠状病毒多肽预测模型,在五折交叉检验中取得 86.18%的准确率,显著高于随机猜测的结果,能够为提高疫苗研发、药物设计和治疗方案的效率和准确性提供有力支持。

关键词: 冠状病毒; 抗冠状病毒多肽; 预测模型; 生物信息学
中图分类号: TP311.1;R373 **文献标志码**: A **文章编号**: 2095-2163(2026)05-0126-05

Support vector machine-based prediction model for anti-coronavirus peptides

ZHANG Qianyue[1], LI Bowen[2], LI Min[2], LU Junkun[1], WANG Ming'en[1], WU Zhen[1], JIAN Xunxun[1], HE Bifang[2], LONG Jinjin[3]
(1 Guizhou Equipment Manufacturing Polytechnic, Guiyang 551400, China; 2 Medical School of Guizhou University, Guiyang 550025, China; 3 Guiyang Hospital of Stomatology, Guiyang 550025, China)

Abstract: Anti-coronavirus peptides are a class of peptides with excellent inhibitory capabilities against coronaviruses. However, identifying and screening potential anti-coronavirus peptides faces challenges due to large data volumes and high complexity, necessitating a precise and efficient predictive tool to assist researchers. This study constructed an anti-coronavirus peptide prediction model based on the latest information from an anti-coronavirus peptide database. The model achieved an accuracy of 86.18% in five-fold cross-validation, significantly outperforming random guessing. This research is expected to provide robust support for improving the efficiency and accuracy of vaccine development, drug design, and therapeutic strategies.

Key words: coronavirus; anti-coronavirus peptides; prediction model; bioinformatics

0 引 言

冠状病毒是一类在自然界中广泛存在的单链 RNA 病毒。2019 年至今,新型冠状病毒已对全球社会经济造成大规模的公共卫生危机[1]。抗菌肽(Antimicrobial Peptides, AMPs)是一类由多种生物体产生的小分子肽,具有广谱抗菌活性,能够有效对抗细菌、真菌、病毒和寄生虫等病原体。是宿主防御机制的重要组成部分,广泛存在于动植物及微生物中[2]。由于其具有广谱抗菌活性,病原体难以对其产生抗药性,因此受到了研究人员的广泛关注[3-4]。许多抗菌肽不仅具有抗细菌、抗真菌的作用,还表现出一定的抗病毒活性,这样的多肽被称为抗病毒多肽。抗冠状病毒多肽属于抗病毒多肽的一个子集,因其在预防和治疗冠状病毒感染中的潜力受到广泛关注[5-6]。如,Zheng 等[7]发现的抗冠状病毒多肽 P6、P8 和 P10 对 SARS-CoV 病毒有着优秀的抑制效果。Xia 等[8]发现的抗冠状病毒多肽 EK1C4,通过抑制冠状病毒复制,对多种冠状病毒产生抑制作用。Li 等[9]发现的抗冠状病毒多肽 Muproporin-M1

基金项目: 贵州省教育厅 2024 年度自然科学研究项目(青年科技人才成长项目)(黔教技[2024]298 号);贵州装备制造职业学院院级科研重点项目(ZBKY2024-19)。

作者简介: 李博文(1998—),男,博士。主要研究方向:电子科学与技术; 李 敏(1999—),女,学士。主要研究方向:电子科学与技术; 卢俊坤(1995—),男,硕士,讲师。主要研究方向:深度学习; 王明恩(1996—),男,硕士。主要研究方向:智能农机导航与定位; 吴 臻(1989—),男,硕士,助教。主要研究方向:机械电子; 简勋勋(1999—),女,硕士,助教。主要研究方向:医疗器械。
通信作者: 张浅阅(1997—),男,硕士。主要研究方向:电子科学与技术。Email:972896854@qq.com; 贺碧芳(1988—),女,博士,副教授。主要研究方向:生物信息学。Email:bfhe@gzu.edu.cn; 龙金金(1995—),女,硕士。主要研究方向:医学信息工程。Email:1277578386@qq.com。

收稿日期: 2024-08-26

可以降低冠状病毒的传染性，从而产生对 SARS-CoV 病毒的抑制作用。

近年来，随着研究的不断深入，科研人员面临着越来越复杂的预测任务和不断提高的研究需求，当前的挑战在于如何从海量的数据集中精准预测具有抗冠状病毒潜力的多肽，然而，现有的抗冠状病毒多肽预测工具在预测准确性方面存在不足，科研人员亟须构建更为先进的模型来满足复杂的预测要求。

2021 年，Pang 等[10]提出了首个专门用于识别抗冠状病毒多肽的预测器 PreAntiCoV。该研究收集了 4 种具有不同功能的多肽数据，开发了 4 个具有相同正样本但不同负样本的预测工具。在独立测试中，PreAntiCoV 的灵敏度（Sensitivity，Sn）、特异性（Specificity，Sp）和马氏相关系数（Matthews Correlation Coefficient，MCC）分别为 0.738、0.855 和 0.223。同年，Timmons 等[11]提出了一种新的方法，命名为 ENNAVIA，使用神经网络预测抗病毒多肽和抗冠状病毒多肽，该方法从相关数据库中收集数据，并整合为 4 个数据集：ENNAVIA-A、ENNAVIA-B、ENNAVIA-C 和 ENNAVIA-D。其中，ENNAVIA-A 和 ENNAVIA-B 用于预测抗病毒多肽，而 ENNAVIA-C 和 ENNAVIA-D 用于预测抗冠状病毒多肽。同年，Kaushik 等[12]构建了 CORONAPEP 用于生成抗冠状病毒多肽，该工具通过对输入数据进行分析，从冠状病毒的目标蛋白中进行检索，来预测抗冠状病毒多肽，对寻求靶点更为具体的多肽用户尤为有用。

目前，已有的多肽预测工具主要基于公共数据库中的抗菌肽与抗病毒多肽数据进行开发。这类工具多用于预测抗菌肽和抗病毒多肽，同时，用于构建模型的数据过于陈旧，该数据局限性导致现有预测工具在面对新型抗冠状病毒多肽时，预测准确性和实用性存在明显不足。随着抗冠状病毒多肽研究的深入，科研人员亟需能够精准预测最新抗冠状病毒多肽的预测模型，这一需求不仅加快新型抗冠状病毒多肽的发现和验证，也能提升现有多肽预测工具的准确性和适用性。因此，构建一个基于最新数据的抗冠状病毒多肽预测模型显得尤为重要。本研究基于最新的抗冠状病毒多肽数据库 ACovPepDB[13]开展数据收集和模型构建。ACovPepDB 汇集了最新发现的抗冠状病毒多肽信息，提供了更加全面和准确的数据支持。通过对这些数据进行清洗和预处理，确保数据的唯一性和准确性。在此基础上，采用多种特征提取技术和机器学习算法，构建了一个高效的抗冠状病毒多肽预测模型。

1　材料与方法

1.1　数据来源

为构建能够有效识别抗冠状病毒多肽的预测模型，本研究所使用的数据集分为两个部分：（1）阳性数据集：阳性数据集由经过实验验证的抗冠状病毒多肽组成，这些多肽在实验中被证明具有抗冠状病毒活性。数据来源为抗冠状病毒多肽数据库 ACovPepDB。（2）阴性数据集：包含不具有抗冠状病毒活性的多肽序列。数据来源于 UniProt 数据库。

为保证数据的唯一性和准确性，本研究采取了以下步骤对阳性数据集进行了预处理：（1）去重筛选：ACovPepDB 收录了大量抗冠状病毒多肽信息，但其记录条目是基于实验过程的，而不是直接以多肽序列的形式呈现。因此，必须对数据进行筛选和清理，并删除重复序列，以获得用于模型构建的多肽序列。（2）删除含有非标准氨基酸的序列：非标准氨基酸（如“X”、“B”、“Z”等）指标准在 20 种氨基酸之外的氨基酸，或表示不确定的氨基酸残基。为了进行特征提取和预测模型的构建，需要确保数据的一致性和可用性，因此必须排除含有非标准氨基酸的序列。（3）删除带修饰的多肽序列：部分人工合成的抗冠状病毒多肽存在多种修饰，这类序列 会影响模型训练，故需删除所有带有多肽修饰的序列。经以上步骤，共收集 165 条抗冠状病毒多肽序列，并通过该组数据构建阳性数据集。

在处理阴性数据集时，本研究采用了与阳性数据集一致的预处理流程（如图 1 所示）。（1）删除重复序列；（2）删除含有非标准氨基酸的序列；（3）删除带有修饰的序列。同时，本研究使用了 CD-HIT 工具对阴性数据集进行去冗余，以避免由于序列过于相似产生过高估计。首先是初步去冗余，将阴性数据集中相似度在 80%以上的多肽序列删除，以确保数据集内部没有过于相似的序列。其次，为避免阴性数据集中包含与阳性数据集过于相似的序列，还进行了二次去冗余，将阴性数据集中与阳性数据集相似度在 40%以上的序列删除，这一步骤提高了模型的区分能力[14]。同时，为确保模型在训练过程中处理的数据具有相似的结构特征，对阴性数据集进行筛选，使其序列长度范围与阳性数据集的序列长度范围保持一致。最后，为避免因数据量差别过大而导致预测偏差，需要平衡阳性和阴性数据集的数量，因此在阴性数据集中随机挑选与阳性数据集

中数量相同的序列,并确保所选序列的长度范围与阳性数据集一致,避免了模型在训练过程中对某一类数据的偏向。经以上步骤,共收集 165 条非抗冠状病毒多肽序列,并通过该组数据构建阴性数据集。

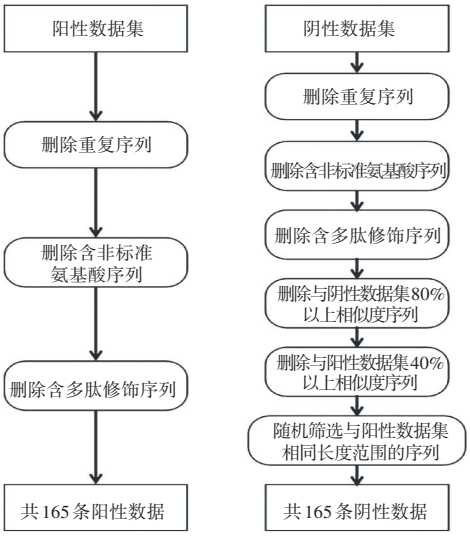


图 1 数据预处理过程

Fig. 1 Data preprocessing

1.2 特征提取与特征选择

为构建高效的预测模型,特征提取和选择是研究中的关键步骤。由于本研究中的阳性/阴性数据集仅包含多肽一级序列信息,因此对序列采用了 20 个天然氨基酸序列(ACDEFGHIKLMNPQRSTVWY)来进行编码。同时,为保证预测模型的准确性,本研究使用了 4 种特征提取方法:(1)氨基酸组成(Amino Acid Composition, AAC)[15];(2)二肽组成(Dipeptide Composition, DPC)[16];(3)K-间距氨基酸对组成(Composition of K-Spaced Amino Acid Group Pairs, CKSAAGP)[17];(4)伪氨基酸组成(Pseudo-Amino Acid Composition, PAAC)[18]。然而,并非所有特征都对模型的预测有帮助,当训练数据相对较少时,高维数据容易导致模型过拟合,或包含许多噪声特征,这些特征不仅没有助益,反而可能干扰模型的学习过程。因此,通过特征筛选可提取最具代表性与信息量的特征,简化模型结构,增强模型的泛化能力和解释性,使其在未知数据集上获得更优的预测表现,同时提升模型的预测性能和降低计算复杂度。

经过以上 4 种特征提取方法,本研究从每条序列中共提取 517 维初始特征向量(表 1),为筛选出对预测任务最有效的特征、构建最优预测模型,本研究进一步采用了 5 种特征提取方法:CHI2、F-score、IG、MIC 和 Pearson 相关系数。

表 1 特征维度数量

Table 1 Feature dimensions

特征提取方法	特征维度数量
氨基酸组成(AAC)	20
二肽组成(DPC)	400
K-间距氨基酸对组成(CKSAAGP)	75
伪氨基酸组成(PAAC)	22
合计	517

1.3 模型构建算法

本研究采用支持向量机(Support Vector Machine, SVM)构建预测模型。SVM 是一种监督学习算法,主要用于分类和回归任务[19]。其核心原理是通过寻找最优超平面,将不同类别的数据点进行分隔。该算法不仅可以处理线性可分的数据,还可通过核函数(如多项式核或径向基函数核)将数据映射到更高维空间,以处理非线性分类问题。通过合理选择核函数与优化参数,SVM 在许多实际问题中表现出色,尤其在文本分类、图像识别和生物信息学等领域[20-22]。因此,本研究采用 SVM 进行模型构建,针对抗冠状病毒多肽生物信息构建高性能预测模型。

1.4 性能评估方法

本研究数据量相对有限,为充分利用有限数据、确保模型评估的可靠性,采用 K 折交叉检验(K-Fold Cross-Validation)方法。该方法将数据集均匀地划分为 K 个大小相近的子集(称为"折")。在每次迭代中,选取其中一个子集作为测试集,其余 $K-1$ 个子集作为训练集。这个过程会重复 K 次,每次选择不同的子集作为测试集。最终,模型的性能指标是这 K 次测试结果的平均值。

实验中,为充分利用有限的数据并确保模型评估的可靠性,K 值被设定为 5,即采用 5 折交叉检验(5-Fold Cross-Validation)。以下是实验的具体步骤:

1)数据准备:将原始数据集随机划分为 5 个大小相等的子集。本研究使用的数据集共包含阴性数据 165 条,阳性数据 165 条。经过 5 折划分,数据被分为训练集和测试集。其中,训练集为 264 条多肽,测试集为 66 条多肽序列,训练集与测试集中的阴性/阳性多肽序列的数量都为 1∶1,并保证序列长度范围一致。

2)交叉检验过程:进行 5 次迭代。在每次迭代中,选择一个不同的子集作为测试集,其余 4 个子集合并作为训练集。在每次迭代中,使用训练集训练

模型，并在测试集上评估模型的性能。

3）结果汇总：将 5 次验证的结果进行汇总，计算模型性能评估指标的平均值。

2　结果与分析

本研究采用 4 种性能指标对模型预测效果进行评估：准确率（Accuracy，Acc）、灵敏度（Sensitivity，Sn）、特异度（Specificity，Sp）和马氏相关系数（Matthews Correlation Coefficient，MCC）。4 种指标的计算公式如下：

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

$$Sn = \frac{TP}{TP + FN} \tag{2}$$

$$Sp = \frac{TN}{TN + FP} \tag{3}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{4}$$

其中，TP 为真阳性；TN 为真阴性；FP 为假阳性；FN 为假阴性。

准确率为表示模型预测正确的样本占总样本的比例；灵敏度表示模型在所有阳性数据中正确识别出阳性数据的比例；特异度表示模型在所有阴性数据中正确识别出阴性数据的比例；马氏相关系数是一个平衡指标，该指标考虑了 TP、TN、FP 和 FN 的情况，其值范围为-1～1，其中 1 表示模型分类效果完美；0 表示随机猜测；-1 表示完全不匹配。同时，本研究也使用了曲线下面积（Area Under the Curve，AUC）来对模型进行评估，曲线下面积是模型区分正负类能力的一个指标。ROC 曲线是以假阳性率为横坐标，真阳性率为纵坐标绘制的曲线。AUC 的值范围为 0.5～1，其中 1 表示分类性能完美，0.5 表示模型的分类能力和随机猜测性能一样。AUC 值越大，模型区分正负类的能力越强。

由于本研究使用了 5 种特征选择方法，在评估模型阶段，将 5 种特征选择方法选择出来的特征分别输入支持向量机（SVM）构建 5 个预测模型，以找出预测效果最佳的模型。经由 5 折交叉检验后，各模型预测性能见表 2。

经多次测试比对发现，F - Score 方法选择特征构建的模型在准确率、特异度、马氏相关系数和 AUC 中都表现出了最佳性能。IG 和 MIC 方法选择特征构建的模型性能相同，是灵敏度表现最优的模型，但模型在其他方面表现不佳，说明使用 IG 与 MIC 方法构建的模型可能过拟合阳性样本，导致对阴性样本的区分能力较弱，以致于在测试时，即使样本的特征稍微接近阳性样本，也会被错误预测为阳性，导致假阳性增多。这会导致使用该模型对阴性数据进行预测时，会将大量阴性数据识别为抗冠状病毒多肽，使得科研人员花费过多时间对阴性数据进行抗病毒抑制实验，产生不必要的资源浪费。同时，使用 F - Score 方法选择特征构建的模型预测性能综合预测性能最优，显著优于随机猜测结果，证明了本研究提出模型的有效性。

表 2　模型预测性能对比

Table 2　Comparison of model prediction performance

特征选择方法	准确率/%	灵敏度/%	特异度/%	马氏相关系数	AUC
CHI2	85.89	80.71	90.98	0.730 0	0.908 0
F - Score	86.18	80.70	91.61	0.733 0	0.926 3
IG	84.98	80.72	89.18	0.708 7	0.910 9
MIC	84.98	80.72	89.18	0.708 7	0.910 9
Pearson 相关系数	84.39	79.54	89.27	0.699 0	0.934 1

3　结束语

本研究构建了准确有效的抗冠状病毒多肽预测模型，通过融合多种特征提取技术与支持向量机算法，旨在提升预测抗冠状病毒多肽的准确性和可靠性。本研究基于 UniProt 与抗冠状病毒多肽数据库 ACovPepDB 分别构建阴性、阳性数据集，并进行数据预处理后，基于多种特征选择方法构建预测模型，其中，采用 F - Score 方法筛选特征构建的模型展现出优异性能，模型准确率为 86.18%，灵敏度为 80.70%，特异度为 91.61%，马氏相关系数为 0.733 0，AUC 为 0.926 3。该模型在现有数据集上表现良好，但在阴性样本的识别精度方面仍有提升空间。

本研究未来可从多方向进行扩展和优化，如：进一步探索更多的特征选择和提取方法（如遗传算法

或递归特征消除),以提升模型对复杂数据集的适应性;结合深度学习方法,如卷积神经网络、Transformer 神经网络等方法,进一步提升模型性能;持续更新和扩展模型训练数据集是提高模型泛化能力的最佳方式,有助于增强模型的准确率和稳定性。未来,本模型将聚焦于模型多样化和数据全面性,以提升模型预测效果。

参考文献

[1] ZHOU P, YANG X L, WANG X G, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin[J]. Nature, 2020, 579(7798):270-273.

[2] MUSTAFA S, BALKHY H, GABERE M N. Current treatment options and the role of peptides as potential therapeutic components for Middle East Respiratory Syndrome (MERS): A review[J]. Journal of Infection and Public Health, 2018, 11(1):9-17.

[3] ZASLOFF M. Antimicrobial peptides of multicellular organisms [J]. Nature, 2002, 415(6870):389-395.

[4] BULET P, STöCKLIN R, MENIN L. Anti-microbial peptides: from invertebrates to vertebrates[J]. Immunological Reviews, 2004, 198:169-184.

[5] HEYDARI H, GOLMOHAMMADI R, MIRNEJAD R, et al. Antiviral peptides against Coronaviridae family: A review[J]. Peptides, 2021, 139:170526.

[6] WALLS A C, PARK Y J, TORTORICI M A, et al. Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein [J]. Cell, 2020, 181(2): 281-292.

[7] ZHENG B J, GUAN Y, HEZ M L, et al. Synthetic peptides outside the spike protein heptad repeat regions as potent inhibitors of SARS-associated coronavirus[J]. Antiviral Therapy, 2005, 10(3):393-403.

[8] XIA S, LIU M, WANG C, et al. Inhibition of SARS-CoV-2 (previously 2019-nCoV) infection by a highly potent pan-coronavirus fusion inhibitor targeting its spike protein that harbors a high capacity to mediate membrane fusion[J]. Cell Research, 2020, 30(4):343-355.

[9] 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.

[10] PANG Y, WANG Z, JHONG J H, et al. Identifying anti-coronavirus peptides by incorporating different negative datasets and imbalanced learning strategies [J]. Briefings in Bioinformatics, 2021, 22(2):1085-1095.

[11] 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.

[12] KAUSHIK A C, MEHMOOD A, SELVARAJ G, et al. CoronaPep: An Anti-coronavirus peptide generation tool[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics , 2021, 18(4):1299-1304.

[13] ZHANG Q, CHEN X, LI B, et al. A database of anti-coronavirus peptides[J]. Scientific Data, 2022, 9(1):294.

[14] WANG J, DAI W, LI J, et al. PaCRISPR: A server for predicting and visualizing anti-CRISPR proteins [J]. Nucleic Acids Research, 2020, 48(W1):W348-W357.

[15] BHASIN M, RAGHAVA G P. Classification of nuclear receptors based on amino acid composition and dipeptide composition[J]. Journal Biolgical Chemistry, 2004, 279(22):23262-23266.

[16] SARAVANAN V, GAUTHAM N. Harnessing computational biology for exact linear B-cell epitope prediction: A novel amino acid composition-based feature descriptor[J]. OMICS: A Journal of Integrative Biology, 2015, 19(10):648-658.

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

[19] CORTES C, VAPNIK V. Support-vector networks[J]. Machine Learning, 1995, 20:273-297.

[20] 李永飞, 贺桂英, 张金. 基于 SVM 算法的生鲜农产品质量安全预警研究[J]. 智能计算机与应用, 2023, 13(7):150-154.

[21] 袁梦薇, 何宇, 王旭. 基于 SVM-SMOTE 算法的一维卷积神经网络电力系统暂态稳定评估模型[J]. 智能计算机与应用, 2023, 13(7):173-178.

[22] 秦信芳, 魏嘉银, 姚林, 等. 基于 SVM-LSTM-ATTE 组合模型的高校录取分数线预测[J]. 智能计算机与应用, 2024, 14(2):177-182.