文章编号:1671-4229(2024)02-0048-09

基于深度学习的禽流感病毒溢出风险预测研究

刘耀华[a]，范馨月[a]，徐雪健[a]，王　娜[a]，寇　铮[a]，强小利[b*]
(广州大学 a. 计算科技研究院，b. 计算机科学与网络工程学院，广东 广州　510006)

摘　要：禽流感病毒基因组由 8 个长短不一的基因节段组成，全长约为 14 ~ 16 kb。由于病毒本身特殊的分子遗传机制，病毒通过基因点突变和基因组重排快速变异，引发病毒感染宿主范围的改变，持续威胁人类健康，因此，自然界禽流感病毒溢出风险预测具有重要公共卫生意义。文章联合使用卷积神经网络和循环神经网络表征病毒基因组序列，分别在特定类群数据集和全类群数据集上进行训练和测试，并对模型的迁移预测能力进行评估。实验结果显示：①特定类群模型对各自数据集的预测性能良好，AUROC 值和 AUPR 值分别达到 0.966 和 0.876 以上，但泛化能力较差；②除 H9N2 类群外，全局模型性能良好，AUROC 值和 AUPR 值均达到 1.000；③基于消融实验，发现注意力机制和 Embedding 层对模型性能影响较大；④进一步测试模型的泛化能力，迁移预测的 AUROC 值和 AUPR 值分别可达 0.984 和 0.941 以上；⑤可视化注意力权重系数矩阵，为模型提供生物学可解释性。性能良好的深度学习预测模型可用于禽流感病毒跨种感染的早期预警。

关键词：禽流感病毒；深度学习；溢出风险；基因组
中图分类号：TP391　　**文献标志码**：A

Prediction of spillover risk for avian influenza virus based on deep learning

LIU Yao-hua[a], *FAN Xin-yue*[a], *XU Xue-jian*[a], *WANG Na*[a], *KOU Zheng*[a], *QIANG Xiao-li*[b*]
(a. Institute of Computing Science and Technology,
b. School of Computer Science and Cyber Engineering, Guangzhou University, Guangzhou 510006, China)

Abstract: The influenza virus genome consists of eight genetic segments of varying lengths, with a total length of approximately 14 ~ 16 kb. Due to the special molecular genetic mechanism of the virus, it undergoes rapid mutations through gene point mutation and genome rearrangement, which leads to changes in its biological infection characteristics and poses a continuous threat to health. Therefore, accurate prediction of natural avian influenza virus spillovers is crucial for public health. This paper, employs a combination of convolutional neural network (CNN) and recurrent neural network (RNN) to represent viral genome sequences. The model's transferability on both specific group datasets and entire datasets was evaluated. The experimental results demonstrate excellent prediction performance of the specific group model on the respective datasets, with AUROC exceeding 0.966 and AUPR values surpassing 0.876. However, its generalization ability is limited. Conversely, except for the H9N2 group, the global model performs well with AUROC and AUPR values reaching 1.000 across all groups. Based on ablation experiments, it was found that attention mechanism and sequence embed-

收稿日期：2023 - 10 - 30；**修回日期**：2023 - 11 - 10
基金项目：国家自然科学基金资助项目(61972109,62172114)；广东省基础与应用基础研究基金资助项目(2022A1515011468)；广州市科技计划资助项目(202201020237, SL2022A03J01035)
作者简介：刘耀华(1999—)，女，硕士研究生. E-mail：yaohua_liu@e.gzhu.edu.cn
* 通信作者. E-mail：qiangxl@gzhu.edu.cn
引文格式：刘耀华，范馨月，徐雪健，等. 基于深度学习的禽流感病毒溢出风险预测研究[J]. 广州大学学报(自然科学版)，2024，23(2)：48-56.

ding method significantly impact model performance while further testing its generalization ability reveals AUROC values above 0.984 and AUPR values over 0.941 for transfer predictions respectively. Visualizing the attention weight matrix provides biological interpretability for the model. The high-performing deep learning prediction model can be effectively utilized for early warning systems against cross-species infections caused by avian influenza viruses.

Key words: avian influenza virus; deep learning; spillover risk; genome

禽流感病毒(Avian Influenza Virus)属于甲型流感病毒,其基因组由 8 个节段组成[1],总长约为 14~16 kb。病毒基因组编码了病毒所有功能性蛋白质,其中,PB1 蛋白参与 RNA 转录延伸过程中催化核苷酸的加入[2],PA 蛋白主要参与病毒蛋白的磷酸化[3],PA、PB1 和 PB2 一起组成流感病毒转录和复制所必需的 RNA 依赖的 RNA 聚合酶复合物[4]。血凝素蛋白 HA 主要功能为识别宿主细胞膜上的唾液酸受体,促使病毒囊膜和细胞膜进行融合,为病毒入侵宿主细胞提供便利。病毒基因组一旦发生位点突变,节段对应编码的蛋白质可能发生功能性改变,从而导致生物学特性的改变。同时,感染不同病毒亚型的禽类体内也可发生病毒基因组重排,形成新的病毒亚型,具备潜在的跨种感染人类的能力,可引发流感大流行。因此,基于深度学习的禽流感病毒溢出风险预测是一个值得探索的科学问题[5]。

生物信息学是一门综合数学、生物学和计算机科学的交叉学科,旨在运用数学和计算机科学的方法挖掘生物学数据内在规律,解决生物学问题。近年来,卷积神经网络在分子模式识别[6-7]、DNA-蛋白质结构和功能预测[6-10]和 RNA/MicroRNA 功能识别[11-12]等领域取得了巨大成就,其原因是卷积运算可以检测不同尺度数据中的关联模式,将原始输入映射到一个自动确定的隐藏空间,适合监督学习模型的构建。残差网络[13]、密集连接网络[14]和双路径网络[15-16]具备良好预测能力[17],用于症状-疾病网络[18]、基因共表达网络[19]、蛋白质-蛋白质相互作用网络[20]和细胞系统层次结构等问题的建模[21]。近年来,核酸合成与测序成本大大降低[22-23],传染病病原体的全基因组测序数据快速积累,提高了基于病毒基因组数据深度学习预测模型的准确率,有利于服务公共卫生事业。

神经网络模型在深度学习中发挥着重要作用,通过使用反向传播算法,可以迭代更改神经网络参数,从而建立监督学习预测模型[24]。本文构建了一个基于深度学习的禽流感病毒溢出风险预测模型(Avian Influenza Virus Spillover Risk Prediction,AIVSRP),结合卷积神经网络和循环神经网络对基因组序列进行特征提取。分别在特定类群数据集、全类群数据集上进行训练和测试,测试该模型的泛化能力。基于消融实验测试预训练向量、序列嵌入方法和注意力机制对模型性能的影响,并可视化注意力权重系数矩阵,为模型提供生物学可解释性。

1　数据集构建

禽流感病毒的基因组主要包含 8 个节段,分别为 PB2、PB1、PA、HA、NP、NA、MP 和 NS。本文所使用的禽流感病毒数据集共有 869 株,其中包含 440 株来自禽类的禽流感病毒株(下称阴性样本)和 429 株来自人类的禽流感病毒株(下称阳性样本)。本文使用多尺度聚类(MDS)将 429 株来自人类的禽流感病毒株分为 3 类:第一类共有 8 株禽流感病毒,主要包括 H9N2、H7N3 及 H7N7(下称 H9N2 类);第二类有 99 株,全部为 H5N1 类型禽流感病毒(下称 H5N1 类);第三类有 322 株,主要为 H7N9 类型(下称 H7N9 类)。将 440 株禽源禽流感病毒株根据病毒类型以及 K 最近邻算法,以阳性样本的 3 个类别的聚类中心为中心位点将 440 株阴性样本按阳性样本比例分入 3 个类别中。由于第一类病毒株样本量较少,将第一类样本引入轻微数量的随机突变,扩增后的第一类样本量为 80 株。最终的数据集类别与数目如表 1 所示。

已有生物学实验证明,禽流感病毒基因组第四节段编码的 HA 蛋白对宿主范围的选择具有关键作用。根据这一特性,将人源禽流感病毒样本的第四节段随机替换为禽源第四节段,作为新的人工阴性样本(禽源禽流感病毒株)加入数据集

中,使得模型能够学习到 HA 节段对溢出风险的影响,以增加模型的鲁棒性。增加阴性样本后的数据集类别个数如表 2 所示。

表 1 禽流感病毒原始数据集

Table 1 The original dataset of avian influenza virus

禽流感病毒类别	H9N2 类	H5N1 类	H7N9 类
阳性样本个数	8 + 72	99	322
阴性样本个数	76	86	278

表 2 扩增后禽流感病毒数据集

Table 2 The amplified avian influenza virus dataset

禽流感病毒类别	H9N2 类	H5N1 类	H7N9 类
阳性样本个数	8 + 72	99	322
阴性样本个数	156	185	600

2 数据预处理

禽流感病毒具有 8 个基因组节段,数据集中最长节段长度为 2 936 base pairs(bp)。本文将所处理的各节段长度固定为 3 000 bp(长度未达到的使用“N”填充)。为保证序列信息的完整性,使用长度为 6,步长为 1 的滑动窗口对 DNA 序列进行扫描,最终形成 2 995 个 6-mer(基因组子序列)。根据这些 6-mer,构建长度为 4^6+1 的序列字典(所有含有“N”的单词,统一按照“NULL”处理)。扫描出的 6-mer 对应序列字典中的词编号,原 DNA 序列最终形成 2 995 × 1 的词索引,并作为网络的实际输入。

3 深度学习模型

传统判断禽流感病毒宿主范围需要进行全基因组测序分析,寻找突变位点,结合实验验证,时间和经济开销较高。本文提出一种深度学习方法,根据一种病毒的 8 个 DNA 节段进行预测,判断其是否感染人,模型如图 1 所示。

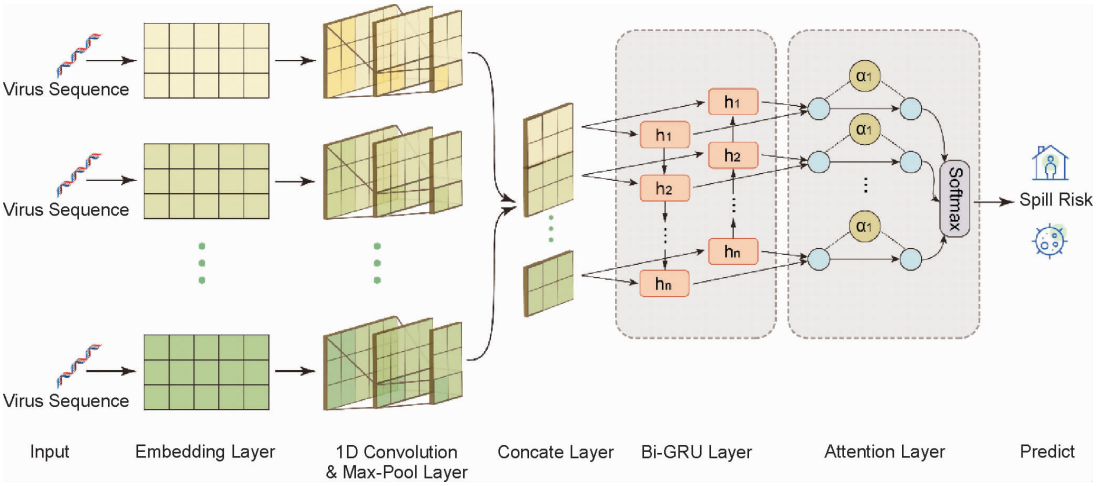

图 1 禽流感病毒溢出风险预测模型结构

Fig. 1 Structure of avian influenza virus spillover risk prediction model

模型使用预训练过的 Embedding 层[25],对输入的序列进行嵌入,将长度为 6 的基因组子串序列(6-mer)映射到 N 维向量空间,使得具有相似含义的子串在向量空间中更加接近,便于捕获 DNA 序列上有效分类信息。

本文结合卷积神经网络和循环神经网络对基因序列进行特征提取。首先使用一维卷积(1D Convolution)提取序列中的局部相关特征。一维卷积层输出通道数为 64,卷积核大小为 40,使用 ReLU 激活。其后为最大池化层(Max-Pooling),池化核尺寸为 20,步长为 20,下采样降低了模型复杂度,提高了计算效率。为了防止过拟合,模型融合层使用了批归一化(Batch Normalization)。

考虑到基因序列信息存在上下文关系,本模型使用双向 GRU(Bi-GRU)来捕获特征。通过正向传播和逆向传播 2 个循环神经网络分别对输入序列进行处理,将 2 个 GRU 得到的输出进行整合,整合方式为将 2 个方向相反的输出拼接。在 Bi-GRU 层后面添加了一个注意力层,提取特征之间的注意力系数矩阵。使用 softmax 函数对加权

聚合后的向量进行激活得到最终预测结果。损失函数使用二元交叉熵,优化器使用 Adma,训练 Epoch 设置为 15。

本文的 3 个病毒类群(H9N2、H5N1、H7N9),阴性数据与阳性数据比例都约为 2∶1,使用 5 折交叉验证计算各项指标。本研究中,由于数据集不平衡,使用的评估指标是 AUROC(ROC 曲线下的面积)及 AUPR(PR 曲线下的面积),AUROC 和 AUPR 的值越接近 1,模型的性能越好。由于病毒基因序列在特定数据集上由亲缘关系划分,存在差异性,本文设置了对照试验用于验证模型的泛化能力,并设置消融实验分析模型每个模块对模型性能的影响。

4　结果分析

4.1　特定类群模型与全局模型

根据 3 个病毒类群数据集(H9N2、H5N1、H7N9),本文训练了 3 个特定类群模型(AIVSRP-specific)模型和一个全局模型(AIVSRP-general)。在 3 个特定类群模型中,每个模型以训练的数据集名称命名,每个模型分别测试 3 个病毒类群测试集的 AUROC 值和 AUPR 值,具体结果如表 3 和表 4 所示。

表 3　AIVSRP-specific 模型 AUROC 值

Table 3　AUROC value of AIVSRP-specific model

AUROC 值	“H9N2”测试集	“H5N1”测试集	“H7N9”测试集
“H9N2”模型	**0.966**	0.774	0.659
“H5N1”模型	0.563	**0.993**	0.503
“H7N9”模型	0.499	0.502	**0.999**

表 4　AIVSRP-specific 模型 AUPR 值

Table 4　AUPR value of AIVSRP-specific model

AUPR 值	“H9N2”测试集	“H5N1”测试集	“H7N9”测试集
“H9N2”模型	**0.876**	0.692	0.546
“H5N1”模型	0.422	**0.968**	0.353
“H7N9”模型	0.339	0.351	**0.997**

由表 3 和表 4 可知,由特定数据集所训练出来的特定类群模型在自身数据集上的 AUROC 和 AUPR 值分别达到 0.966 和 0.876 以上,在其他类群数据集上的 AUROC 和 AUPR 值最优表现为 0.774 和 0.692。由此可知,特定类群训练出来的模型对自身类群有较好的预测能力,但泛化能力较差。

全局模型将 3 个类群数据集看作一个整体数据集,训练一个全局模型,分别测试该模型在 3 个病毒类群的效果,结果如表 5 所示。

表 5　AIVSRP-general 模型对各个类群的预测结果

Table 5　Prediction results of AIVSRP-general model for each group

AIVSRP-general	“H9N2”测试集	“H5N1”测试集	“H7N9”测试集
AUROC 值	0.948	1.000	1.000
AUPR 值	0.834	1.000	1.000

由表 5 可知,AIVSRP-general 模型针对 3 个数据集的预测性能良好,在 H5N1 和 H7N9 数据集上的 AUROC 值和 AUPR 值均为 1,在包含轻微突变的 H9N2 数据集效果略差,AUROC 值和 AUPR 值分别为 0.948 和 0.834,略低于用 H9N2 训练的 AIVSRP-specific 模型。AIVSRP-general 模型整体不存在遗忘某类数据特征的情况,预测结果良好。

4.2　预训练向量

为了测试 Embedding 层是否使用预训练向量对模型性能的影响,本文设计了 Embedding 层没有使用预训练向量的模型(AIVSRP no-pre 模型)与 AIVSRP 模型进行对比,结果如图 2 所示。图 2(a)为用特定数据集训练的 AIVSRP-specific 模型和 AIVSRP-specific no-pre 模型 AUROC 值对比结果,图 2(b)为用特定数据集训练的 AIVSRP-specific 模型和 AIVSRP-specific no-pre 模型的 AUPR 值对比结果,图 2(c)为用 general 数据集(H9N2 + H5N1 + H7N9)训练的 AIVSRP-general 模型和 AIVSRP-general no-pre 模型 AUROC 值对比结果,图 2(d)为用 general 数据集训练的 AIVSRP-general 模型和 AIVSRP-general no-pre 模型的 AUPR 值对比结果。

由图 2(a)和图 2(b)可知,H9N2 数据集训练的 AIVSRP-specific 模型在 H9N2 测试集上的 AUROC 值和 AUPR 值与 AIVSRP-specific no-pre 模型的相比提升了 0.093 和 0.114,H7N9 数据集训练的 AIVSRP-specific 模型在 H7N9 测试集上的 AUROC 值和 AUPR 值与 AIVSRP-specific no-pre 模型

的相比提升了 0.007 和 0.008。由图 2(c)和图 2(d)可知 AIVSRP-general 模型在 3 类病毒测试集上的 AUROC 值和 AUPR 值与 AIVSRP-general no-pre 模型的相差不大。

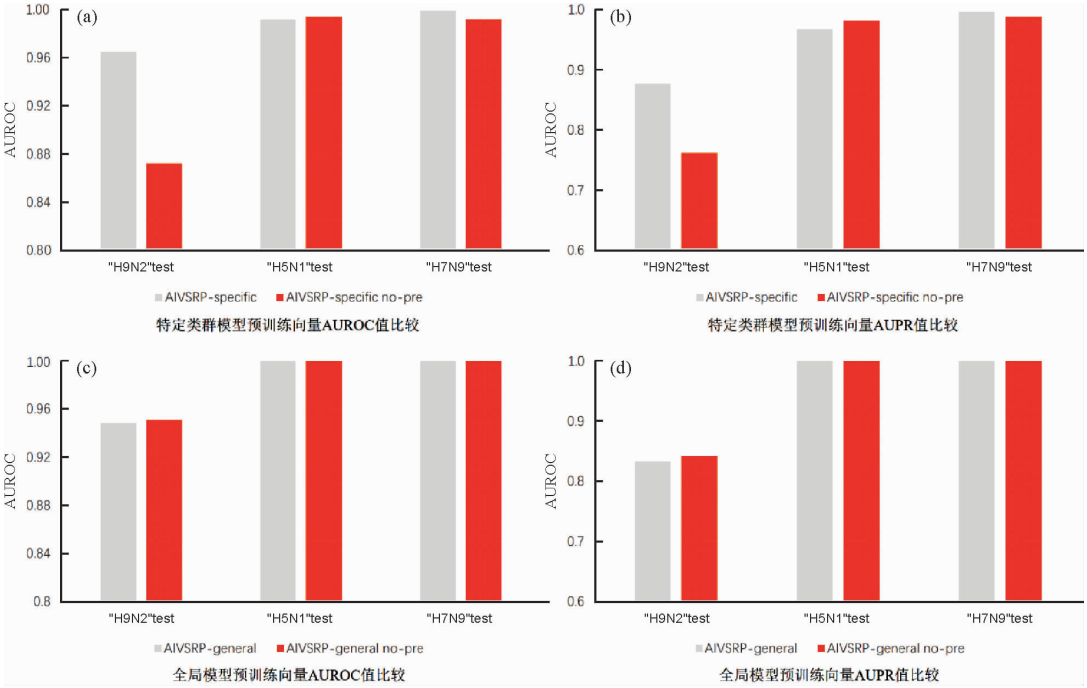

图 2 预训练向量对预测模型的影响

Fig. 2 Impact of pre-trained vectors on prediction models

综上所述,3 个特定类群训练的 AIVSRP-specific no-pre 模型与 AIVSRP-specific 模型性能相差较小,general 数据集训练的 AIVSRP-general no-pre 模型与 AIVSRP-general 模型相比模型性能接近,AUROC 值和 AUPR 值相差在 0.004 和 0.008 以内。由此可知,预训练向量对本文模型效果影响不大。

4.3 序列嵌入方法

为了测试序列嵌入方法对模型性能的影响。本文将 one-hot 编码的模型(AIVSRP one-hot 模型)与 AIVSRP 模型作对比。如图 3 所示,图 3(a)为用特定数据集训练的 AIVSRP-specific 模型和 AIVSRP-specific one-hot 模型 AUROC 值对比结果,图 3(b)为用特定数据集训练的 AIVSRP-specific 模型和 AIVSRP-specific one-hot 模型的 AUPR 值对比结果,图 3(c)为用 general 数据集训练的 AIVSRP-general 模型和 AIVSRP-general one-hot 模型 AUROC 值对比结果,图 3(d)为用 general 数据集训练的 AIVSRP-general 模型和 AIVSRP-general one-hot 模型的 AUPR 值对比结果。

由图 3(a)和图 3(b)可知,H9N2 数据集训练的 AIVSRP-specific 模型在 H9N2 测试集上的 AUROC 值和 AUPR 值与 AIVSRP-specific one-hot 模型的相比提升了 0.456 和 0.534,H5N1 数据集训练的 AIVSRP-specific 模型在 H5N1 测试集上 AUROC 值和 AUPR 值与 AIVSRP-specific one-hot 模型的相比提升了 0.492 和 0.619,H7N9 数据集训练的 AIVSRP-specific 模型在 H7N9 测试集上的 AUROC 值和 AUPR 值与 AIVSRP-specific one-hot 模型的相比提升了 0.009 和 0.017。

由图 3(c)和图 3(d)可知,AIVSRP-general 在 H9N2 测试集上的 AUROC 值和 AUPR 值与 AIVSRP-general one-hot 模型的相比提升了 0.400 和 0.470,在 H5N1 测试集上 AUROC 值和 AUPR 值与 AIVSRP-general one-hot 模型的相比提升了 0.419 和 0.620,在 H7N9 测试集上的 AUROC 值和 AUPR 值与 AIVSRP-general one-hot 模型的相比提升了 0.359 和 0.574。

3 种 AIVSRP-specific 模型与 AIVSRP-specific one-hot 模型相比,预测效果良好。AIVSRP-general 模型相较于 AIVSRP-general one-hot 模型性能提升显著。本文模型的序列嵌入方法优于 one-hot 编码。

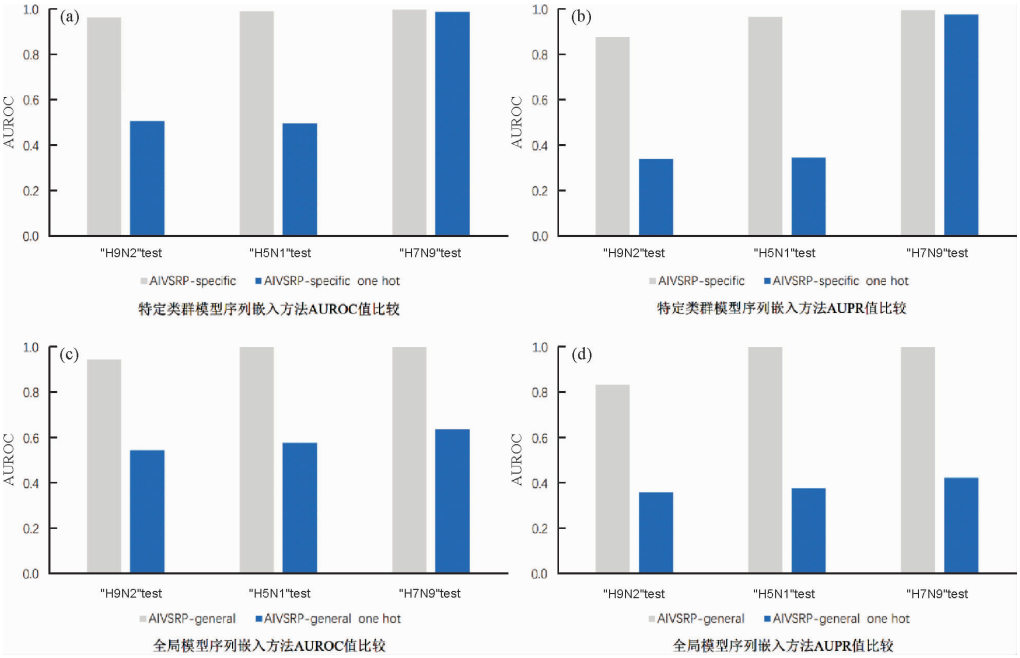

图 3　序列嵌入方法对预测模型的影响

Fig. 3　Impact of sequence embedding methods on prediction models

4.4　注意力机制

为了测试注意力机制在模型中的作用,本文训练没有使用注意力层的模型(AIVSRP no-att)与 AIVSRP 模型进行对比,结果如图 4 所示。图 4(a)为用特定数据集训练的 AIVSRP-specific 模型和 AIVSRP-specific no-att 模型的 AUROC 值对比结果,图 4(b)为用特定数据集训练的 AIVSRP-specific 模型和 AIVSRP-specific no-att 模型的 AUPR 值对比结果,图 4(c)为用 general 数据集训练的 AIVSRP-general 模型和 AIVSRP-general no-att 模型的 AUROC 值对比结果,图 4(d)为用 general 数据集训练的 AIVSRP-general 模型和 AIVSRP-general no-att 模型的 AUPR 值对比结果。

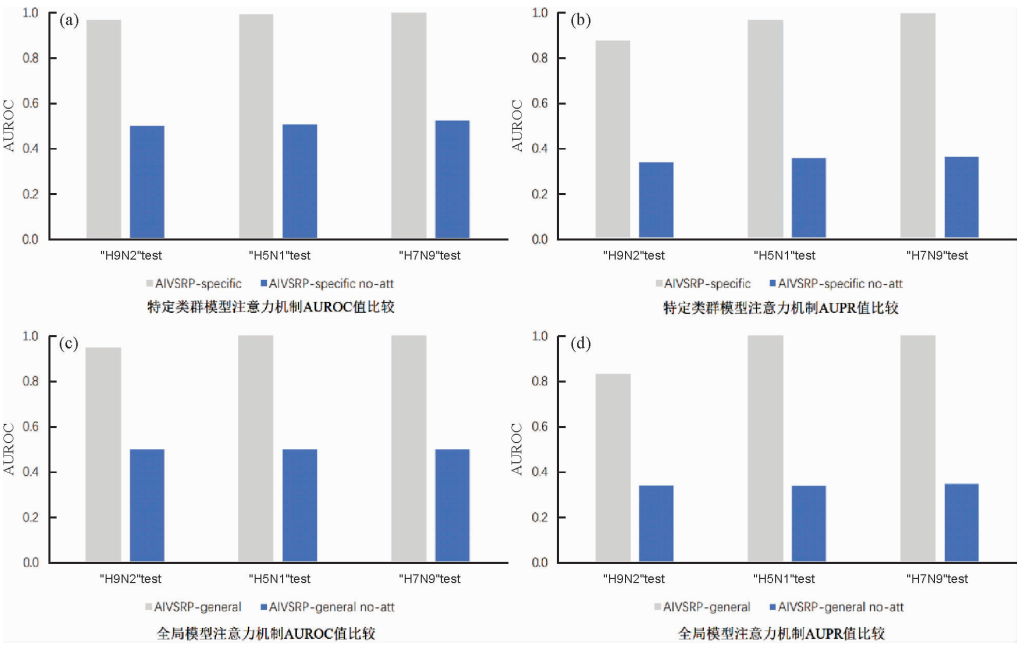

图 4　注意力机制对预测模型的影响

Fig. 4　Impact of attention mechanism on prediction models

由图4(a)和图4(b)可知,H9N2数据集训练的AIVSRP-specific模型在H9N2测试集上的AUROC值和AUPR值与AIVSRP-specific no-att模型的相比提升了0.464和0.537,H5N1数据集训练的AIVSRP-specific模型在H5N1测试集上AUROC值和AUPR值与AIVSRP-specific no-att模型的相比提升了0.485和0.609,H7N9数据集训练的AIVSRP-specific模型在H7N9测试集上的AUROC值和AUPR值与AIVSRP-specific no-att模型的相比提升了0.475和0.633。

由图4(c)和图4(d)可知,AIVSRP-general模型在H9N2数据集上的AUROC值和AUPR值与AIVSRP-general no-att模型的相比提升了0.448和0.493,在H5N1数据集上AUROC值和AUPR值与AIVSRP-general no-att模型的相比提升了0.500和0.661,在H7N9数据集上的AUROC值和AUPR值与AIVSRP-general no-att模型的相比提升了0.500和0.653。

综上,3种AIVSRP-specific模型与AIVSRP no-att模型相比,预测效果良好。AIVSRP-general模型相较于AIVSRP-general no-att模型性能提升显著。注意力机制在本文模型中发挥积极作用。

4.5 模型泛化能力

特定类群模型在经过训练过的数据集类群上表现较好,在非训练数据集类群上表现较差,考虑到病毒持续变异产生新病毒,需要评估模型的泛化能力,即在新类群病毒数据上的预测表现进行探讨。针对上述问题本文尝试了一种泛化模型,该模型是用两种已有的数据集进行训练,测试另一个数据集。其中,H9N2 + H5N1 model表示选取数据集H9N2和H5N1作为训练集,选取H7N9作为测试集,测试模型性能,泛化模型的测试结果见表6。

表6 泛化模型的测试结果

Table 6 Test results of the generalized model

模型	指标	
	AUROC	AUPR
“H9N2” + “H5N1”	1.000	1.000
“H9N2” + “H7N9”	1.000	1.000
“H5N1” + “H7N9”	0.984	0.941

由表6可知,在测试的3个泛化模型中,AUROC和AUPR值均在0.984和0.941以上,模型的泛化能力效果显著,在非训练类群上的测试情况表现良好。在已经训练2个类群数据集的情况下,对于新出现的类群数据集具有识别能力,证实本文模型具备一定的泛化能力。

4.6 人工阴性数据对预测的影响

上述模型所使用的数据均为已添加人工阴性样本的数据集。为验证人工阴性数据对模型的影响,本文用剔除人工阴性数据的样本集重新训练特定类群模型与全局模型,预测人工阴性样本(表7),并与上文构建的模型进行比较。

表7 人工阴性数据对预测的影响

Table 7 Impact of artificial negative data on prediction

ACC	“H9N2”测试集	“H5N1”测试集	“H7N9”测试集
AIVSRP-specific(无人工阴性数据)	0.000	0.000	0.000
AIVSRP-general(无人工阴性数据)	0.125	0.000	0.000
AIVSRP-specific(有人工阴性数据)	0.950	0.989	0.990
AIVSRP-general(有人工阴性数据)	0.936	1.000	1.000

由表7可见,剔除人工阴性数据的数据集训练的3种AIVSRP-specific模型和AIVSRP-general模型准确率都较差,包含人工阴性数据的数据集训练的AIVSRP-specific模型和AIVSRP-general模型的准确率则显著提高。可见,如果不添加人工阴性数据,模型对于潜在的阴性数据预测效果很差,而添加了人工阴性数据,能显著提高模型预测的鲁棒性,与生物学实验结果相一致。

4.7 注意力机制权重可视化

由上述实验可知,AIVSRP模型在预测禽流感病毒的溢出风险中取得了良好的性能表现,并且模型中的注意力机制层对模型起到关键作用。为了进一步解释模型的预测结果,本文提取模型中注意力层权重系数,并映射到病毒基因组序列的8个节段,组成了一个8×13的系数矩阵,并将该矩阵进行可视化,如图5所示。图5中纵轴表示禽

流感病毒株基因组的 8 个节段,横轴表示每个片段按比例获得的 13 个权重系数,其权重越大,颜色越深。

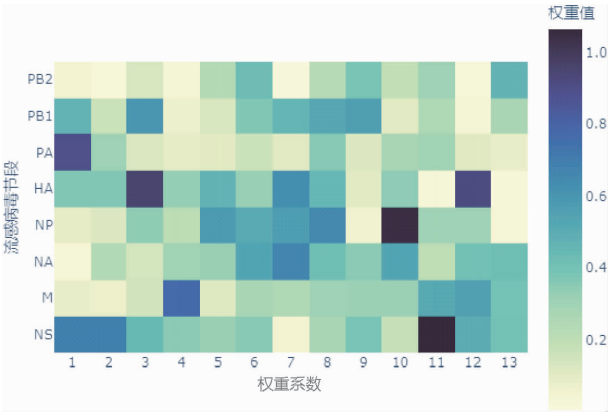

图 5　全局模型注意力权重可视化

Fig. 5　Global model attention weights visualization

如图 5 所示,高权重主要集中在 HA、NP、NS 节段上,尤其表现为 NS 节段和 NP 节段的后半基因序列区域,以及 HA 节段的受体结合区域。禽流感病毒的 NS 节段主要编码 NS1 和 NS2 蛋白。NS1 不仅对于病毒复制至关重要,还参与了抑制宿主细胞天然免疫抗病毒反应的过程,NS2 则主要参与了病毒核糖核蛋白(vRNP)的转运过程。禽流感病毒 NP 蛋白编码于 NP 基因段中,NP 蛋白对于流感病毒的转录和复制具有关键作用,是构成 vRNP 的主要骨架,NP 蛋白与流感病毒宿主范围息息相关。模型捕获的生物学特征,与已有的病毒分子生物学进展一致。

5　结　论

本文结合卷积神经网络和循环神经网络对基因组序列进行特征提取,结合注意力机制,搭建了一个基于深度学习的禽流感病毒溢出风险预测模型。实验结果显示:①特定类群模型对各自数据集的预测性能良好,但泛化能力较差;②除 H9N2 类群外,全局模型的 AUROC 值和 AUPR 值均为 1.000;③基于消融实验,发现 AIVSRP no-pre 模型的 AUROC 值和 AUPR 值与 AIVSRP 模型的基本持平,AIVSRP one-hot 模型 AUROC 值和 AUPR 值和 AIVSRP 模型的相差较大,AIVSRP no-atten 模型的 AUROC 值和 AUPR 值与 AIVSRP 模型的相差较大;④人工阴性样本的添加对模型的识别提供了帮助,提升了模型的鲁棒性;⑤可视化注意力权重矩阵为模型提供了生物学可解释性。

参考文献:

[1] Eisfeld A J, Neumann G, Kawaoka Y. At the centre: Influenza A virus ribonucleoproteins[J]. Nature Reviews Microbiology, 2015, 13(1): 28-41.

[2] Thompson W W, Shay D K, Weintraub E, et al. Influenza-associated hospitalizations in the United States[J]. Jama, 2004, 292(11): 1333-1340.

[3] Sanz-Ezquerro J J, De La Luna S, Ortín J, et al. Individual expression of influenza virus PA protein induces degradation of coexpressed proteins[J]. Journal of Virology, 1995, 69(4): 2420-2426.

[4] Deng Q, Wang D, Xiang X, et al. Nuclear localization of influenza B polymerase proteins and their binary complexes[J]. Virus Research, 2011, 156(1/2): 49-53.

[5] Qiang X L, Xu P, Fang G, et al. Using the spike protein feature to predict infection risk and monitor the evolutionary dynamic of coronavirus[J]. Infectious Diseases of Poverty, 2020, 9(1): 1-8.

[6] Li Y, Wang S, Umarov R, et al. DEEPre: Sequence-based enzyme EC number prediction by deep learning[J]. Bioinformatics, 2018, 34(5): 760-769.

[7] Dai H, Umarov R, Kuwahara H, et al. Sequence2Vec: A novel embedding approach for modeling transcription factor binding affinity landscape[J]. Bioinformatics, 2017, 33(22): 3575-3583.

[8] Wang M, Tai C, E W N, et al. DeFine: Deep convolutional neural networks accurately quantify intensities of transcription factor-DNA binding and facilitate evaluation of functional non-coding variants[J]. Nucleic Acids Research, 2018, 46(11): e69.

[9] Zhou J, Troyanskaya O G. Predicting effects of noncoding variants with deep learning-based sequence model[J]. Nature

Methods, 2015, 12(10): 931-934.

[10] Kulmanov M, Khan M A, Hoehndorf R. DeepGO: Predicting protein functions from sequence and interactions using a deep ontology-aware classifier[J]. Bioinformatics, 2018, 34(4): 660-668.

[11] Pan X, Shen H B. Predicting RNA-protein binding sites and motifs through combining local and global deep convolutional neural networks[J]. Bioinformatics, 2018, 34(20): 3427-3436.

[12] Yang C, Yang L, Zhou M, et al. LncADeep: An ab initio lncRNA identification and functional annotation tool based on deep learning[J]. Bioinformatics, 2018, 34(22): 3825-3834.

[13] He K, Zhang X, Ren S, et al. Computer Vision-ECCV 2016: 14th European Conference, October 11-14, 2016[C]. Berlin: Springer, 2016.

[14] Huang G, Liu Z, Van Der Maaten L, et al. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017[C]. Piscataway: IEEE, 2017.

[15] Chung J, Gulcehre C, Cho K, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling[J]. arXiv preprint arXiv:1412.3555, 2014.

[16] Godinez W J, Hossain I, Lazic S E, et al. A multi-scale convolutional neural network for phenotyping high-content cellular images[J]. Bioinformatics, 2017, 33(13): 2010-2019.

[17] Li Y, Kuwahara H, Yang P, et al. PGCN: Disease gene prioritization by disease and gene embedding through graph convolutional neural networks[J]. Biorxiv, 2019: 532226.

[18] Zhou X, Menche J, Barabási A L, et al. Human symptoms-disease network[J]. Nature Communications, 2014, 5(1): 4212.

[19] Yang Y, Han L, Yuan Y, et al. Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types[J]. Nature Communications, 2014, 5(1): 3231.

[20] Rual J F, Venkatesan K, Hao T, et al. Towards a proteome-scale map of the human protein-protein interaction network[J]. Nature, 2005, 437(7062): 1173-1178.

[21] Ma J, Yu M K, Fong S, et al. Using deep learning to model the hierarchical structure and function of a cell[J]. Nature methods, 2018, 15(4): 290-298.

[22] 吴琦琨, 赖浪文, 徐怀胜, 等. 新一代数据存储介质——DNA[J]. 广州大学学报(自然科学版), 2020, 19(6): 35-40.

[23] 张新建, 寇铮. 基于DNA计算的优先编码器逻辑分子实现[J]. 广州大学学报(自然科学版), 2020, 19(5): 12-17.

[24] Shrikumar A, Greenside P, Kundaje A. International Conference on Machine Learning: 34th International Conference on Machine Learning, August 06-11, 2017[C]. New York: PMLR, 2017.

[25] Hong Z, Zeng X, Wei L, et al. Identifying enhancer-promoter interactions with neural network based on pre-trained DNA vectors and attention mechanism[J]. Bioinformatics, 2020, 36(4): 1037-1043.

【责任编辑: 陈 钢】