计算机科学
COMPUTER SCIENCE

基于注意力机制与集成学习的甲型H5N1流感病毒抗原相似性预测

王迎晖, 李维华, 李川, 陈伟, 文俊颖

引用本文

王迎晖, 李维华, 李川, 陈伟, 文俊颖基于注意力机制与集成学习的甲型H5N1流感病毒抗原相似性预测[J]. 计算机科学, 2022, 49(11A): 210900032-6.

WANG Ying-hui, LI Wei-hua, LI Chuan, CHEN Wei, WEN Jun-ying. Prediction of Antigenic Similarity of Influenza A/H5N1 Virus Based on Attention Mechanism and Ensemble Learning [J]. Computer Science, 2022, 49(11A): 210900032-6.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210900032

基于注意力机制与集成学习的甲型 H5N1 流感病毒抗原相似性预测

王迎晖　李维华　李　川　陈　伟　文俊颖
云南大学信息学院　昆明 650503
(742854584@qq.com)

摘　要　甲型流感病毒可能导致季节性流感病毒疫情甚至全球大爆发。流感病毒血凝素蛋白的持续和累积变化会产生新的抗原株,致使疫苗效力降低甚至失效。抗原相似性预测对流感疫情监测和疫苗选择是至关重要的。甲型 H5N1 病毒源于禽类,可引起人类肺炎和多器官衰竭。针对流感病毒及其抗原特点,设计一个预测病毒抗原相似性的神经网络模型,该模型分别基于 K-mer 嵌入与位置特异性矩阵表示序列信息并进行融合;在此基础上,设计融合注意力机制的集成深度学习模型用于抗原相似性预测。实验结果表明,相比基准模型,该模型显著提高了模型预测的准确率、精确率、F1 值和 MCC 值。从实验中可以看出该模型具有良好的鲁棒性和扩展性,在抗原相似性预测领域有很好的应用潜力。

关键词:抗原性相似性;甲型流感;H5N1;集成学习;注意力机制
中图法分类号　TP391

Prediction of Antigenic Similarity of Influenza A/H5N1 Virus Based on Attention Mechanism and Ensemble Learning

WANG Ying-hui,LI Wei-hua,LI Chuan,CHEN Wei and WEN Jun-ying
School of Information Science and Engineering,Yunnan University,Kunming 650503,China

Abstract　Influenza A virus can lead to seasonal influenza virus outbreaks or even global outbreaks. Continued and cumulative changes in the hemagglutinin protein of influenza viruses can lead to the antigenic variants that reduce vaccine effectiveness or even cause vaccine failure. Therefore,antigenic similarity prediction is critical for influenza outbreak surveillance and vaccine selection. Although A/H5N1 virus originates in poultry,they can cause pneumonia and multiple organ failure in humans. In view of influenza virus and the antigenic characteristics,this paper designs a neural network model to predict the antigenic similarity between viruses. Specifically,the model represents amino acid sequences based on the K-mer embedding and position specific scoring matrices(PSSM),then integrates the features. Furthermore,integrated deep learning model fused with attention mechanism for antigen similarity prediction. Experimental results show that the model significantly improves the accuracy,precision,F1 and MCC compares with the baseline models. Experimental results show that the model has good robustness and extensibility,and has good application potential in the field of antigenic similarity prediction.

Keywords　Antigenic similarity,Influenza A,H5N1,Ensemble learning,Attention mechanism

1　引言

甲型流感病毒是一种呼吸道病毒。根据表面蛋白血凝素(Hemagglutinin,HA)和神经氨酸酶(Neuraminidase,NA)的差异,流感病毒可分为不同的亚型,如 H5N1,H1N1。流感表面蛋白血凝素会频繁进行突变,累积的氨基酸突变导致病毒产生抗原漂移(Antigenic Drift)并产生抗原变异体[1],导致季节性流感疫情,甚至造成全球范围的流感大爆发。每个流感季,在全球范围内,流感会导致极高的发病率与死亡率[2]。例如,甲型 H1N1 在 1978－1919 年导致 2 000 多万人死亡[3]。1968 年爆发的 H3N2 流感病毒,至今每年仍导致 50 万人死亡[4]。1997 年人类从禽类感染甲型 H5N1 流感病毒,其主要特征是导致肺炎和多器官衰竭,具有高致病性[5],引起急性肺损伤和急性呼吸窘迫综合征,死亡率高达 52.79%[6-7]。目前,疫苗是预防流感和阻止流感疫情最有效的措施。然而,疫苗的免疫效果取决于疫苗株和流行株之间的抗原相似性(也称抗原距离或抗原差异)。因此,抗原相似性预测对流感疫情监测和疫苗选择是至关重要的。

抗原距离来源于免疫分析,如血凝抑制(Hemagglutinin-Inhibition,HI)试验。然而,HI 试验花费昂贵且耗时,无法获得所有抗原与抗原血清的相似性[8]。因此,基于计算方法的流感抗原相似性预测在流感早期检测、监测和疫苗筛选方面具有重要的作用。

基于计算方法的流感抗原相似性预测主要包括基于矩阵补全的方法和基于序列的方法。基于矩阵补全的方法通常假设抗原和抗血清关系可以映射到低维的空间中,并且可以通过部分 HI 滴度数据还原全部 HI 滴度数据。Smith 等基于 1968－2003 年监测到 H3N2 的 HI 值,将 HI 值转换为欧几里

基金项目:国家自然科学基金(32060151)
This work was supported by the National Natural Science Foundation of China(32060151).
通信作者:李维华(lywey@163.com)

得距离并绘出抗原图[9]，达到对未知 HI 值预测的目的。Cai 等人通过矩阵补全算法获得完整的 HI 滴度数据矩阵，并将其抗原关系投影到二维空间中[10]。Wang 等是在低秩矩阵补全的基础上，基于 HA 蛋白序列的病毒相似性和机器学习推断抗原与抗血清之间的抗原距离[11]。这些方法为流感抗原相似的预测性奠定了基础，但往往局限于 HI 滴度数据，忽略了病毒突变产生的变化。

基于序列的抗原相似性预测方法，通常利用机器学习对 HA 序列和毒株之间的相似关系建模。HA 蛋白包含 HA1 和 HA2 两条链，其中 HA1 的突变比 HA2 频繁[12]。因此，基于序列的方法主要以 HA1 为分析对象。例如，Peng 等人基于 HA1 序列和贝叶斯分类器预测流感病毒抗原相似性，证明流感病毒的 HA 亚型都经历过 HA1 蛋白的突变[13]。Liao 等人针对 HA1 序列提出一种定量评估氨基酸差异的评分方法，利用回归模型进行预测[14]。Lees 等[15]基于 HA1 序列中的关键位点，结合支持向量机模型预测流感病毒抗原相似性。Ren 等人基于甲型 H1N1 流感的 HI 数据，将随机森林和支持向量回归模型应用于流感病毒抗原性预测，且分析 HA1 多肽氨基酸变化与抗原性变异之间的关系[16]。基于序列的方法充分利用机器学习的优势挖掘 HA 蛋白序列特征和抗原相似性之间的关系，提升抗原相似性预测性能。然而，依赖于人工特征的传统机器学习局限了模型的性能。

区别于传统的机器学习，深度学习可以学习特征之间的非线性关系，因此被广泛应用到蛋白质相关的预测中，包括蛋白质识别[17]、蛋白质二级结构的预测[18]等。在病毒的抗原性预测上，Yin 等利用循环神经网络探索了甲型 H3N2 流感抗原相似性预测[19]，实验结果表明神经网络模型有助于改进流感病毒抗原性预测。然而，单一的神经网络模型仅关注输入的一类特征。

现有的研究表明，有效的蛋白序列特征表示有助于提高模型的预测性能。例如，基于 ProtVec 表示的蛋白质特征[24]显著提高了蛋白质分类的预测效果；Patrick[25]的工作表明氨基酸序列经过 K-mer 划分并结合 Word2Vec 编码方式能有效捕捉氨基酸上下文的关键信息。Qiu 等人的工作表明位置特异性矩阵(Position Specific Scoring Matrices，PSSM)与其他特征融合有助于改进流感病毒抗原性预测[26]。

本文针对 H5N1 流感病毒及其抗原特性，设计一种 HA 氨基酸序列特征表示方法。其次，为了整合不同神经网络在提取特征上的优势，基于卷积神经网络(Convolutional Neural Network，CNN)和长短期记忆网络(Long Short-Term Memory，LSTM)，构建一个三层一维卷积结合长短期记忆网络(Three One-dimensional Convolutions Long Short-term Memory，TCLSTM 模块；基于残差神经网络[27](Residual Networks，ResNets)与注意力机制设计一个注意力残差网络(Residual Networks with Attention Mechanism，ResAM)模块；进一步通过集成策略融合 TCLSTM 与 ResAM 得到 TREL(TCLSTM and ResAM Ensemble Learning)，用于 H5N1 抗原相似性预测。本文的主要工作和创新在于：

(1)基于 word2vec 获取 K-mer 嵌入，并与 PSSM 中的氨基酸进化信息进行融合表示 HA 序列特征，该方法具有集成更多抗原特征的灵活性，也为预测模型的可扩展性提供了支撑；

(2)基于集成策略构建 TREL 预测模型，整合不同神经网络在提取特征上的优势，捕捉氨基酸序列中有区分度的抗原特征。

2 抗原特征编码

甲型 H5N1 流感病毒的 HA 是由相同亚基组成的三聚体，每个亚基由 HA1 链和 HA2 链组成。因为 HA1 的突变比 HA2 频繁[12]，所以本文仅对 HA1 链中 320 位的氨基酸进行编码。

2.1 HA 序列的特征

本文设计的 HA 系列特征主要包括 K-mer 嵌入和 PSSM。

K-mer 嵌入在 HA 序列分割为若干长度为 K 的 K-mer 基础上，基于 word2vec 方法[28]获得。假设 L 为 HA 序列的长度，K 为其子序列的长度，S 为步长，子序列的数量为 C，那么 $C=\lfloor(L-K)\rfloor+1$。例如，$K=3$，$S=1$，那么{“DTLCIGYHANNS”}的编码子序列为{“DTL”}{“TLC”}{“LCI”}{“CIG”}{“IGY”}{“GYH”}。本文选择 $K=3$，并采用 CBOW(Continuous Bag-Of-Words)模型训练获取每个氨基酸嵌入向量。因为 H5N1 的 HA1 链包含 320 个氨基酸，所以每条序列最终得到 320×20 的矩阵，即 $\boldsymbol{M}_1\in\mathcal{R}^{320\times 20}$。

PSSM 利用 NCBI 的 PSI-blast[29]在非冗余蛋白质数据库中对比数据集中的 HA 序列。本文使用的参数 *E-value* 为 0.001，迭代次数是 3，因此一条长 $L=320$ 的 HA 序列得到的 PSSM 向量矩阵 $\boldsymbol{M}_2$ 为：

$$\boldsymbol{PSSM}=\boldsymbol{M}_2\begin{bmatrix}P_{1,1} & P_{1,2} & \cdots & P_{1,19} & P_{1,20}\\ P_{2,1} & P_{2,2} & \cdots & P_{2,19} & P_{2,20}\\ \vdots & \vdots & \ddots & \vdots & \vdots\\ P_{L,1} & P_{L,2} & \cdots & P_{L,19} & P_{L,20}\end{bmatrix} \tag{1}$$

其中，$\boldsymbol{M}_2\in\mathcal{R}^{320\times 20}$。其次，PSSM 矩阵中不存在对‘－’的向量，所以在 PSSM 矩阵对应序列中‘－’出现的位置插入 0 向量。

2.2 特征融合

为了充分利用氨基酸的嵌入特征和进化信息，本文基于矩阵加将 K-mer 嵌入和 PSSM 分别得到的特征矩阵 $\boldsymbol{M}_1$ 与 $\boldsymbol{M}_2$ 进行融合：

$$\boldsymbol{M}=\boldsymbol{M}_1+\boldsymbol{M}_2 \tag{2}$$

由于 $\boldsymbol{M}_1$ 与 $\boldsymbol{M}_2$ 都是 320×20 的矩阵，所以融合得到的 $\boldsymbol{M}\in\mathcal{R}^{320\times 20}$。

由于模型预测两个毒株 V_i 和 V_j 之间的抗原相似性，所以模型输入的特征使用两个菌株的特征差来表示：

$$\boldsymbol{M}_*=\boldsymbol{M}^i-\boldsymbol{M}^j \tag{3}$$

其中，$\boldsymbol{M}^i$ 和 $\boldsymbol{M}^j$ 分别是毒株 V_i 和 V_j 的融合特征，且 $\boldsymbol{M}_*\in\mathcal{R}^{320\times 20}$。因此，两条 H5N1 的 HA 序列之间的抗原特征映射到 320×20 的向量空间中。

3 抗原相似性预测集成模型

TREL 抗原相似性预测模型的主要框架如图 1 所示，主要包括 TCLSTM 与 ResAM 两个独立模块。TCLSTM 模块学习特征的局部与远程的相互作用，ResAM 模块学习特征的深层次关系。最后通过结果加权平均的集成策略实现抗原相似性预测。

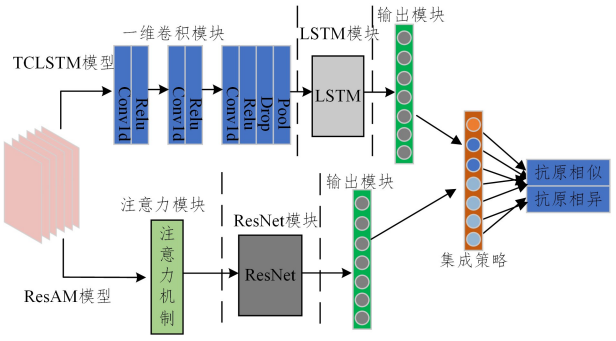

图 1　TREL 模型结构

Fig. 1　Architecture of TREL

3.1　TCLSTM 模块

TCLSTM 模块的主要框架如图 1 所示，主要包括卷积模块、LSTM 模块和输出模块。卷积模块通过一维卷积学习局部相互作用，LSTM 模块利用 LSTM 捕捉远程相互作用，输出模块利用学习到的特征得到抗原相似性预测的标签结果。

3.1.1　卷积模块

卷积模块包括 3 个卷积层和 1 个池化层。卷积层完成映射：

$$f_{\mathrm{conv1d}_k}=\delta_1(\boldsymbol{W}_k\cdot\boldsymbol{M}_{*\,i:i+j-1}+\boldsymbol{b}_k) \tag{4}$$

$$h_{\mathrm{conv1d}}=f_{\mathrm{conv1d}_1}(f_{\mathrm{conv1d}_2}(f_{\mathrm{conv1d}_3}(\boldsymbol{M}_*))) \tag{5}$$

其中，$\boldsymbol{W}_k$ 表示第 k 层一维卷积的卷积核，δ_1 是激活函数 ReLU[30]。

卷积结果送入到池化层，完成映射：

$$\boldsymbol{h}_c=f_{\mathrm{pool}}(\boldsymbol{h}_{\mathrm{conv1d}}) \tag{6}$$

3.1.2　LSTM 模块

LSTM[31]通过门机制来控制神经单元状态的更新，可以捕获序列的长期依赖。将卷积模块最终获得的特征表示为 $\boldsymbol{h}_c=[x_1,x_2,\cdots,x_T]$，其中 $x_i\in\mathcal{R}^k$。LSTM 在 t 时刻的状态 h_t 计算如下：

$$h_t=f_{\mathrm{LSTM}}(x_t,h_{t-1}) \tag{7}$$

LSTM 模块最终获得抗原特征 $\boldsymbol{h}_{\mathrm{TCLSTM}}=[h_1,h_2,\cdots,h_T]$。

3.2　ResAM 模块

ResAM 模块包括注意力单元、ResNet 单元和输出单元。ResNet 单元增加特征的多样性，注意力单元利用自适应权重调整特征。输出模块得到抗原相似性预测的标签结果。

3.2.1　注意力机制模块

利用通道注意力机制[32]，为每个通道分配不同的权重，让模型关注重要的特征。具体公式如下：

$$f_{\mathrm{conv2d}_k}=\delta_1(\boldsymbol{V}_k\cdot\boldsymbol{Input})+\boldsymbol{b}_k \tag{8}$$

$$\begin{aligned}\boldsymbol{h}_{\mathrm{attention}}=&+f_{\mathrm{conv2d}_1}(\delta_1(f_{\mathrm{conv2d}_1}(AvgPool(\boldsymbol{M}_*))))+\\&f_{\mathrm{conv2d}_2}(\delta_1(f_{\mathrm{conv2d}_2}(MaxPool(\boldsymbol{M}_*))))\end{aligned} \tag{9}$$

其中，$\boldsymbol{V}_k$ 表示第 k 层二维卷积的卷积核，$AvgPool$ 是平均池化，$MaxPool$ 是最大池化。

3.2.2　ResNet 模块

ResNet 通过引入残差单元实现直接映射，连接网络的不同层，解决网络退化问题。将注意力机制模块获得的特征表示为 $\boldsymbol{h}_{\mathrm{attention}}=[z_1,z_2,\cdots,z_T]$，其中 $z_i\in\mathcal{R}^k$。ResNet 最终获得的抗原特征 $\boldsymbol{h}_{\mathrm{ResAM}}$ 计算如下：

$$\boldsymbol{h}_{\mathrm{ResAM}}=f_{\mathrm{ResNet}}(\boldsymbol{h}_{\mathrm{attention}}) \tag{10}$$

其中，$\boldsymbol{h}_{\mathrm{ResAM}}=[c_1,c_2,\cdots,c_T]$ 且 $c_i\in\mathcal{R}^k$。

3.3　输出模块

该模块将学习到的特征 $\boldsymbol{h}_{\mathrm{ResAM}}$ 与 $\boldsymbol{h}_{\mathrm{TCLSTM}}$ 经过全连接层，并用于确定每对 H5N1 毒株的标签序列：

$$\boldsymbol{y}_{\mathrm{TCLSTM}}=\delta_2(\boldsymbol{W}\cdot\boldsymbol{h}_{\mathrm{TCLSTM}}+\boldsymbol{b}) \tag{11}$$

$$\boldsymbol{y}_{\mathrm{ResAM}}=\delta_2(\boldsymbol{W}\cdot\boldsymbol{h}_{\mathrm{ResAM}}+\boldsymbol{b}) \tag{12}$$

其中，$\boldsymbol{W}$ 和 $\boldsymbol{b}$ 是预测模块的参数，δ_2 是激活函数 sigmoid。输出结果序列 $\boldsymbol{y}=[y_1,y_2,\cdots,y_i]$，其中 $y_i\in[0,1]$。如果抗原相异，则 $y_i=1$；反之抗原相似，$y_i=0$。$\boldsymbol{y}_{\mathrm{TCLSTM}}$ 是 TCLSTM 模块输出的标签序列，$\boldsymbol{y}_{\mathrm{ResAM}}$ 为 ResAM 模块输出的标签序列。

3.4　集成策略

本文的集成策略是对结果加权平均，加入权重参数 α 用于控制独立模块对融合模型的影响程度。公式如下：

$$\boldsymbol{y}_{\mathrm{TREL}}=\alpha\boldsymbol{y}_{\mathrm{TCLSTM}}+(1-\alpha)\boldsymbol{y}_{\mathrm{ResAM}} \tag{13}$$

其中，$\boldsymbol{y}_{\mathrm{TREL}}$ 是融合模型得到的预测标签序列。

3.5　模型训练

通过训练数据集训练模型的参数，本文的训练数据包含 H5N1 毒株对的 HA1 序列以及对应的标签。每对 H5N1 毒株的标签根据抗原距离[33]确定：

$$D_{ij}=\sqrt{\frac{H_{ii}\times H_{jj}}{H_{ij}\times H_{ji}}} \tag{14}$$

其中，H_{ij} 是抗血清 V_i 抑制菌株 V_j 凝集红细胞的最大稀释度。如果 $D_{ij}\geqslant 4$，则菌株 V_i 和菌株 V_j 抗原相异，反之抗原相似[14]。

本文使用二元交叉熵损失函数计算预测误差，具体定义如下：

$$J(\theta)=-\frac{1}{n}\sum_{i=1}^{n}(y_i\log(\tilde{y}_i))+(1-y_i)\log(1-f(\tilde{y}_i))) \tag{15}$$

其中，θ 是模型参数，y_i 是 H5N1 毒株对的真实标签，$\tilde{y}_i$ 是预测标签，n 是训练数据的数量。

本文使用随机梯度下降算法 Adam[34]（Adaptive Moment Estimation）更新网络模型参数，同时采用 Dropout 策略来避免模型出现过拟合，提高模型的鲁棒性。

4　实验与结果分析

4.1　实验环境与参数设置

本文实验环境的主要参数为：处理器：Intel i7-10700K 3.80 GHz，图形加速卡：NVIDIA GeForce RTX 3070 8 GB，内存 32GB，操作系统：Windows10；采用深度学习框架 pytorch1.8.0 以及内置的神经网络进行模型的训练和测试。实验采用的参数如表 1 所列。

表 1　模型参数配置

Table 1　Parameters of TREL

参数	取值
迭代次数	100
批大小	32
优化器	Adam
学习率	0.001
Dropout	0.5
α	0.6

4.2　实验数据集

本文使用 Yin 提供的甲型 H5N1 流感病毒数据集[35]分析 TREL 模型的性能。该数据集包括 666 对 H5N1 流感病毒抗原关系数据集和 260 条 H5N1 流感病毒 HA 序列数据集。

4.3 性能评估指标

本文采用准确率(Acc)、精确率(Pre)、召回率(Rec)、F1分数(*F*1)、马修斯相关系数(Matthews Correlation Coefficient,MCC)作为评价的指标:

$$Acc=\frac{TP+TN}{TP+TN+FP+FN} \tag{16}$$

$$Pre=\frac{TP}{TP+FP} \tag{17}$$

$$Rec=\frac{TP}{TP+FN} \tag{18}$$

$$F1=1-\frac{FN+FP}{2TP+FN+FP} \tag{19}$$

$$MCC=\frac{TP\times TN-FP\times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{20}$$

其中,*TP* 和 *TN* 分别代表分类正确的抗原相似和抗原相异的样本数;*FP* 和 *FN* 分别代表分类错误抗原相似和抗原相异的样本数;*MCC* 是实际分类与预测分类之间的相关系数,取值范围是[−1,1],值越趋于 1 表示预测越准确,反之则预测效果越差。

4.4 实验设计与结果分析

为了验证 TREL 模型的性能,本文设计 6 个实验,采用 5 折交叉验证并取平均值作为最后结果。

实验 1 本文使用 Liao 的方法、Lees 的方法、Peng 的方法和 IAV-CNN 作为基线,验证 TREL 在 H5N1 抗原相似性预测中优势。

实验 2 分析不同的 H5N1 序列特征表示对模型性能的影响。

实验 3 分析集成学习对独立模块性能的影响。

实验 4 分析注意力机制对模型性能的影响。

实验 5 分析本文设计的特征在不同的分类模型上的性能表现。

IAV-CNN 是目前 H5N1 预测性能最优的模型。由于对比模型 IAV-CNN 的实验是先将 80%的数据集作为训练集,20%的数据作为测试集,其中训练集做 5 折交叉验证进行模型的选择,然后单独测试集测试,因此为了对比,实验 1 采用相同的数据训练划分方法。表 2 列出了本文的 TREL 模型和基线模型的性能。从整体上看,TREL 模型优于所有基线模型。例如,与 IAV-CNN 模型相比,TREL 模型的准确率提高了4.4%,精确率超过 1%,召回率超过 8.2%,F1 分数超过 5%,*MCC* 超过 11.3%。结果表明,基于集成学习的 TREL 模型改进了 H5N1 病毒抗原相似性预测的性能。

表 2 TREL 模型与基线模型的性能比较

Table 2 Comparison between TREL and baselines

Model	*Acc*	*Pre*	*Rec*	*F*1	*MCC*
Liao 等	0.818	0.825	0.809	0.802	0.616
Lees 等	0.759	0.751	0.735	0.737	0.483
Peng 等	0.765	0.754	0.745	0.747	0.499
IAV-CNN	0.889	0.910	0.894	0.897	0.746
TREL	0.933	0.920	0.976	0.947	0.859

表 3 列出了 TREL 模型在不同特征上的表现。从实验结果可以看到,在单一编码中 1-mer 的特征表示可以获得最佳的性能;采用 K-mer 与 PSSM 相加方式相融合可以提升模型的性能,且 3-mer 与 PSSM 进行特征融合后的编码在 TREL 模型上的准确率达到 0.889,优于其他特征表示方式。分析原因可能是 3-mer 含有的氨基酸上下文元素信息与 PSSM 特征中含有的氨基酸进化信息互补,提高了预测准确度。

表 3 不同特征表示方式的影响

Table 3 Impact of different feature representations

Method	*Acc*	*Pre*	*Rec*	*F*1	*MCC*
1-mer	0.884	0.915	0.896	0.904	0.762
2-mer	0.882	0.913	0.896	0.903	0.758
3-mer	0.876	0.918	0.879	0.897	0.748
PSSM	0.869	0.891	0.900	0.894	0.728
1-mer&PSSM	0.878	0.906	0.896	0.900	0.747
2-mer&PSSM	0.878	0.912	0.888	0.899	0.748
3-mer&PSSM	0.889	0.915	0.903	0.908	0.769

表 4 列出了 TREL 模型和其独立模块的性能。TREL 模型对比 TCLSTM 与 ResAM 两个独立模块可以看出,独立模块之间的模块预测效果差距小,经过模块融合后,预测的准确度提高了 1.1%,且精确率、*Rec*、*F*1、*MCC* 均有不同程度的提升,表明预测模型融合了 TCLSTM 对局部和远程依赖特征提取。的能力与 ResAM 模块对深层次特征提取的能力。综上所述,融合后的 TREL 模型相比独立模型具有更好的鲁棒性和预测效果。

表 4 独立模块与集成学习模型性能的比较

Table 4 Performance comparison of independent model and TREL

Model	*Acc*	*Pre*	*Rec*	*F*1	*MCC*
TCLSTM	0.875	0.900	0.898	0.898	0.741
ResAM	0.875	0.910	0.885	0.896	0.742
TREL	0.889	0.915	0.903	0.908	0.769

表 5 列出了注意力机制加入 ResNet 模块和集成模型 TREL 后的性能。可以看出,注意力机制的加入,提升了预测模型的效果,表明注意力机制可以挖掘到更深层次的特征信息,并且根据权重让预测模型关注关键特征信息。

表 5 注意力机制对模型预测性能的影响

Table 5 Effect of attention mechanism on model prediction performance

Model	*Acc*	*Pre*	*Rec*	*F*1	*MCC*
ResNet	0.864	0.905	0.872	0.887	0.721
ResAM	0.875	0.910	0.885	0.896	0.742
TREL 无注意力	0.875	0.901	0.896	0.897	0.742
TREL	0.889	0.915	0.903	0.908	0.769

表 6 列出了不同模型下 H5N1 抗原相似性预测的性能。对比模型包括传统机器学习模型 K 近邻(K-Nearest Neighbor,KNN)、支持向量机(Support Vector Machine,SVM)和 Adaboost(Adaptive Boosting)算法,以及典型的深度学习模型 LSTM 和门控循环(Gated Recurrent Unit,GRU)网络。从表 6 可以看到,TREL 明显优于传统机器学习的方法,这表明神经网络模型能更有效地捕捉到 H5N1 病毒 HA 序列的特征。从表 6 还可以看到,TREL 的预测模型超出了经典神经网络模型,这表明集成深度学习模型在 H5N1 抗原相似性预测中具有较好的鲁棒性。

表 6 TREL 模型和经典模型的性能比较

Table 6 Performance comparison of TREL and classical model

Model	*Acc*	*Pre*	*Rec*	*F*1	*MCC*
KNN	0.829	0.829	0.832	0.825	0.652
SVM	0.828	0.831	0.820	0.817	0.639
Adaboost	0.844	0.837	0.837	0.835	0.672
LSTM	0.781	0.796	0.897	0.837	0.483
GRU	0.833	0.873	0.856	0.863	0.652
TREL	0.889	0.915	0.903	0.908	0.769

结束语 甲型 H5N1 病毒抗原相似性对疫苗监控、筛选和生产具有重要意义。针对 H5N1 表面蛋白血凝素频繁突变的特点，利用 word2vec 获取 K-mer 嵌入，并与 PSSM 特征进行融合表示氨基酸序列信息，设计融合 TCLSTM 和 ResAM 模块的 H5N1 抗原相似性预测模型 TREL。实验表明，特征融合的方式结合神经网络集成模型能有效提高预测性能。此外，模型也具有较好的扩展性。

为了充分利用蛋白质序列蕴含的信息，使用其他嵌入模型对蛋白质进行特征表示，丰富蛋白质特征是一个改进的研究方向。另外，融合一些蛋白质的其他结构信息来表示蛋白质序列，以及优化的深度模型，也可以为模型性能的提升带来机会。

参 考 文 献

[1] MEDINA R A，GARCIA-SASTRE A. Influenza A viruses：new research developments [J]. Nat RevMicrobiol，2011，9(8)：590-603.

[2] THOMPSON W W，SHAY D K，WEINTRAUB E，et al. Mortality Associated With Influenza and Respiratory Syncytial Virus in the United States [J]. JAMA，2003，289(2)：179-186.

[3] CHAN P K S. Outbreak of Avian Influenza A(H5N1) Virus Infection in Hong Kong in 1997 [J]. Clinical Infectious Diseases，2002，34(Supplement_2)：S58-S64.

[4] RUSSELL C，JONES T，BARR I，et al. The global circulation of seasonal influenza A (H3N2) viruses [J]. Science，2008，320(5874)：340-346.

[5] DE JONG M D，SIMMONS C P，THANH T T，et al. Fatal outcome of human influenza A(H5N1) is associated with high viral load and hypercytokinemia [J]. Nature Medicine，2006，12(10)：1203-1207.

[6] BAUER TT，EWIG S，RODLOFF A C，et al. Acute Respiratory Distress Syndrome and Pneumonia：A Comprehensive Review of Clinical Data [J]. Clinical Infectious Diseases，2006，43(6)：748-756.

[7] PEIRIS J S M，CHEUNG C Y，LEUNG C Y H，et al. Innate immune responses to influenza A H5N1：friend or foe? [J]. Trends in Immunology，2009，30(12)：574-584.

[8] SUN H，YANG J，ZHANG T，et al. Using Sequence DataTo Infer the Antigenicity of Influenza Virus [J]. mBio，2013，4(4)：e00230-00213.

[9] SMITH D J，LAPEDES A S，DE JONG J C，et al. Mapping the Antigenic and Genetic Evolution of Influenza Virus [J]. Science，2004，305(5682)：371-376.

[10] CAI Z，ZHANG T，WAN X F. A computational framework for influenza antigenic cartography [J]. PLoS Comput Biol，2010，6(10)：e1000949.

[11] WANG P，ZHU W，LIAO B，et al. Predicting Influenza Antigenicity by Matrix CompletionWith Antigen and Antiserum Similarity [J]. Frontiers in Microbiology，2018，9：2500.

[12] PLOTKIN J B，DUSHOFF J. Codon bias and frequency-dependent selection on the hemagglutinin epitopes of influenza A virus [J]. Proc. Natl. Acad. Sci. USA，2003，100(12)：7152-7157.

[13] PENG Y，WANG D，WANG J，et al. A universal computational model for predicting antigenic variants of influenza A virus based on conserved antigenic structures [J]. Sci Rep，2017，7：42051.

[14] LIAO Y C，LEE M S，KO C Y，et al. Bioinformatics models for predicting antigenic variants of influenza A/H3N2 virus [J]. Bioinformatics，2008，24(4)：505-512.

[15] LEES W D，MOSS D S，SHEPHERD A J. A computational analysis of the antigenic properties of haemagglutinin in influenza A H3N2 [J]. Bioinformatics，2010，26(11)：1403-1408.

[16] REN X，LI Y，LIU X，et al. Computational Identification of Antigenicity-Associated Sites in the Hemagglutinin Protein of A/H1N1 Seasonal Influenza Virus [J]. PLoS One，2015，10(5)：e0126742.

[17] ZENG M，LI M，FEI Z，et al. A Deep Learning Framework for Identifying Essential Proteins by Integrating Multiple Types of Biological Information [J]. IEEE/ACM Trans. Comput. Biol Bioinform，2021，18(1)：296-305.

[18] SPENCER M，EICKHOLT J，JIANLIN C. A Deep Learning Network Approach to ab initio Protein Secondary Structure Prediction [J]. IEEE/ACM Trans Comput Biol Bioinform，2015，12(1)：103-112.

[19] YIN R，ZHANG Y，ZHOU X，et al. Time series computational prediction of vaccines for influenza A H3N2 with recurrent neural networks [J]. Journal of Bioinformatics and Computational Biology，2020，18(1)：2040002.

[20] YI H C，YOU Z H，WANG M N，et al. RPI-SE：a stacking ensemble learning framework for ncRNA-protein interactions prediction using sequence information [J]. BMC Bioinformatics，2020，21(1)：60.

[21] ZHANG B，LI J，LÜ Q. Prediction of 8-state protein secondary structures by a novel deep learning architecture [J]. BMC Bioinformatics，2018，19(1)：293.

[22] XU H J，YANG Y，LI G L. Material Recognition Method Based on Attention Mechanism and Deep Convolutional Neural Network[J]. Computer Science，2021，48(10)：220-225.

[23] LI F，ZHU F，LING X，et al. Protein Interaction Network Reconstruction Through Ensemble Deep LearningWith Attention Mechanism [J]. Frontiers in Bioengineering and Biotechnology，2020，8(390).

[24] ASGARIE，MOFRAD M R. Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics [J]. PLoS One，2015，10(11)：e0141287.

[25] NG P. dna2vec：Consistent vector representations of variable-length k-mers [J]. ：arXiv：1701. 06279，2017.

[26] QIU J，QIU T，YANG Y，et al. Incorporating structure context of HA protein to improve antigenicity calculation for influenza virus A/H3N2 [J]. Sci Rep，2016，6：31156.

[27] HE K，ZHANG X，REN S，et al. Deep Residual Learning for Image Recognition [C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). 2016：770-778.

[28] LING W，DYER C，BLACK A W，et al. Two/Too Simple Adaptations of Word2Vec for Syntax Problems[C]// Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics：Human Language Technologies. Association for Computational Linguistics，2015：1299-1304.

[29] ALTSCHUL S F, MADDEN T L, SCHÉÄFFER A A, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs [J]. Nucleic Acids Research, 1997, 25(17): 3389-3402.

[30] HE J C, LI L, XU J C, et al. Relu Deep Neural Networks and Linear Finite Elements [J]. Journal of Computational Mathematics, 2020, 38(3): 502-527.

[31] HOCHREITER S, SCHMIDHUBER J. Long Short-Term Memory [J]. Neural Computation, 1997, 9(8): 1735-1780.

[32] WOO S, PARK J, LEE J Y, et al. CBAM: Convolutional Block Attention Module [C]// Proceedings of the ECCV. 2018: 3-19.

[33] NDIFON W, DUSHOFF J, LEVIN S A. On the use of hemagglutination-inhibition for influenza surveillance: surveillance data are predictive of influenza vaccine effectiveness [J]. Vaccine, 2009, 27(18): 2447-2452.

[34] KINGMA D P, BA J. Adam: A Method for Stochastic Optimization [J]. arXiv: 1412. 6980, 2014.

[35] YIN R, THWIN N N, ZHUANG P, et al. IAV-CNN: a 2D convolutional neural network model to predict antigenic variants of influenza A virus [J]. IEEE/ACM Transactions on Computational Beology and Bioinformatics, 2021, 9: 1-1.

WANG Ying-hui, born in 1997, postgraduate. His main research interests include deep learning and bioinformatics.

LI Wei-hua, born in 1977, Ph. D, associate professor. Her main research interests include data mining and bioinformatics.