计算机科学
COMPUTER SCIENCE

基于图卷积网络的甲型流感H3N2抗原性预测

何明龙, 赵锟, 李维华, 李川

引用本文

何明龙, 赵锟, 李维华, 李川. 基于图卷积网络的甲型流感H3N2抗原性预测[J]. 计算机科学, 2023, 50(11A): 230100113-6.

HE Minglong, ZHAO Kun, LI Weihua, LI Chuan. Antigenicity Prediction of Influenza A/H3N2 Based on Graph Convolutional Networks [J]. Computer Science, 2023, 50(11A): 230100113-6.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230100113

基于图卷积网络的甲型流感 H3N2 抗原性预测

何明龙　赵　锟　李维华　李　川
云南大学信息学院　昆明 650503
(hml13017483632@163.com)

摘　要　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,能够逃避人类免疫并引起季节性流感或流感大爆发。及时识别新的抗原变异体,对疫苗筛选和流感预防是至关重要的。图嵌入模型在部分数据缺失的情况下仍然可以实现有效的相互关系建模。针对甲型流感病毒 H3N2,提出一种基于图卷积神经网络的抗原性预测方法,获取流感毒株低维稠密嵌入向量,同时对序列信息进行编码并作为补充特征,利用深度神经网络模型对特征进行融合并学习关键的抗原特征,完成抗原性预测。在两个数据集上的实验结果表明,该方法相比其他同类方法,显著提升了抗原相似性预测性能,具有良好的鲁棒性和可扩展性。此外,从实验中还可以看出,图卷积神经网络可以有效地获取抗原相似关系的抗原特征。

关键词:甲型流感;H3N2;抗原相似性;图卷积网络;神经网络
中图法分类号　TP391

Antigenicity Prediction of Influenza A/H3N2 Based on Graph Convolutional Networks

HE Minglong,ZHAO Kun,LI Weihua and LI Chuan
School of Information Science and Engineering,Yunnan University,Kunming 650503,China

Abstract　Continual and accumulated mutations in the hemagglutinin(HA) protein of influenza A virus generates novel antigenic strains that can evade human immunity and cause seasonal influenza or influenza pandemics. Timely identification of new antigenic variants is crucial for the selection of vaccines and influenza prevention. Graph embedding models can effectively model interactions even when some data is missing. For influenza A virus H3N2,this paper proposes an antigenicity prediction method based on graph convolutional networks to obtain the low-dimensional dense embedding vector of influenza strain. Then,it encodes the sequence information as supplementary features. Furthermore,deep neural networks is adopted to fuse these features and learn the dominative features for antigenicity prediction. Experimental results on two datasets show that,compared with those of existing methods,the proposed method significantly improves the performance of antigenic similarity prediction,and has good robustness and scalability. In addition,it can be seen from experiments that graph convolutional networks can effectively obtain the antigenic features of the antigenic similarity relationship.

Keywords　Influenza A,H3N2,Antigenic similarity,Graph convolutional network,Neural network

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,甲型流感病毒可分为不同的亚型,如 H1N1 和 H3N2 等。为了逃避先前感染或接种疫苗后获得的群体免疫,负责与细胞受体结合并导致病毒入侵的 HA 的累积氨基酸突变导致病毒产生抗原漂移(Antigenic Drift),引起季节性的甲型流感疫情,甚至造成全球范围的甲型流感大爆发[1]。因此,甲型流感病毒在全世界范围内都严重地威胁着人类的健康。

目前,疫苗是预防甲型流感和阻止甲型流感疫情最有效的手段。疫苗的免疫效果主要取决于疫苗株和流行株之间的抗原距离(抗原差异或抗原相似性)[2-3]。然而,表征抗原相似性的血凝抑制(Hemagglutinin-Inhibition,HI)试验[4-5]成本高,且费时费力。已有的研究表明,抗原性预测为流感疫苗的筛选和更新提供了直接的科学依据[2]。因此,快速、准确地预测甲型流感病毒的抗原相似性,对疫苗的监控、筛选和生产是非常重要的[6]。

目前,针对抗原相似性预测的研究工作主要利用机器学习方法对 HA 序列突变与抗原相似性之间的关系进行建模,围绕关键位点[7]进行抗原特征的选择和表示,并利用模型完成抗原性预测。Smith 等[2]基于 1968 年到 2003 年间监测到的 HI 值,将已知 HI 值转换为欧氏距离,并将它们映射到 2D 空间,建立隐含着抗原与抗血清之间距离的抗原图,对抗原差

基金项目:国家自然科学基金(32060151);云南大学第十三届研究生科研创新项目(2021Y280);云南省中青年学术与技术带头人后备人才培养计划项目(202305AC160014)
This work was supported by the National Natural Science Foundation of China(32060151),13th Postgraduate Scientific Research Innovation Project of Yunnan University(2021Y280) and Reserve Talents for Young and Middle-aged Academic and Technical Leaders in Yunnan Province Training Program(202305AC160014).
通信作者:李维华(lywey@163.com)

异进行可视化。Lee 等[8]基于序列变化的氨基酸数目，建立一个抗原变异体的预测模型。Liao 等[9]通过多元回归分析出 19～23 个关键位点，再使用线性模型预测抗原性。Lees 等[10]更新 5 个结合区域上以及附近的氨基酸，并基于这些结合区域的氨基酸变化建立预测模型。Peng 等[11]将 HA1 序列人工划分为 10 个区域带，利用每个区域带中氨基酸的差异数作为抗原特征。Yao 等[12]为了获得更好的性能，结合反映氨基酸物理化学、生化性质指数的 AAindex[13]数据库，提出一种联合随机森林回归方法，用于优化 H3N2 流感病毒抗原性的预测。现有的方法表明，基于 HA 序列的抗原性预测是可行的，也为疫苗的快速、早期筛选提供了支撑。

然而，基于关键氨基酸位点或者特征进行建模，忽略了氨基酸替换发生的遗传背景；同时，甲型流感病毒突变率非常高，毒株突变的位点可能超出预测模型的关键位点，所以最终导致预测方法的鲁棒性还不足[14]；其次，传统机器学习由输入空间的单个线性变换组成，这可能忽略了蛋白序列中氨基酸及其位置之间的非线性关系。

深度学习作为机器学习一个新的研究方向，由于具有良好的自适应能力、泛化能力以及更强的预测能力，被应用于甲型流感病毒的抗原性预测。例如，Yin 等[15]首次尝试了基于 ProtVec[16]表示氨基酸序列特征，并使用了二维卷积神经网络对流感抗原性预测进行探索，建立端到端的预测模型。实验结果表明，深度学习在流感抗原性预测上具有很好的应用潜力。然而，现有基于深度学习的方法仅捕捉氨基酸序列的特征，而没有充分利用毒株间存在的间接交互关系，因此还无法获得满意的预测结果。

本文使用 A/H3N2 流感病毒数据，构建抗原性网络，利用图卷积神经网络获取网络全局结构信息，学习流感毒株的低维稠密嵌入向量，并融合 HA 序列，搭建一个用于甲型流感病毒抗原相似性预测的深度学习框架；结合 CNN 和 RNN 提取抗原特征，并利用注意力层对提取的特征进行选择性融合，充分发挥深度学习的优势，避免预先设定抗原特征和关键位点，提高了流感抗原性预测的鲁棒性。

2 抗原特征表示

2.1 基于图卷积神经网络的毒株嵌入

图神经网络(Graph Neural Network，GNN)将深度学习应用到图上，挖掘网络中蕴含的非线性关系，得到节点或者图的向量表示。图卷积网络(Graph Convolutional Network，GCN)[18]是最典型的 GNN，本质上每个节点的特征通过邻域节点特征的变换、聚合进行更新。对 H3N2 数据进行图(网络)建模，则使用 GCN 可从中学习图的全局结构信息，充分利用毒株间的间接交互关系，为抗原性预测提供基础。

血凝抑制试验测定毒株 v_i 制备的抗血清抑制毒株 v_j 凝集红细胞的最大稀释度 H_{ij}，但是 HI 滴度并不能很好地度量毒株 v_i 和 v_j 之间的抗原相似性。故 Archetti-Horsfall 距离[19]作为一种毒株抗原相似性的定量度量被引入：

$$D_{ij}=\sqrt{\frac{H_{ii}*H_{jj}}{H_{ij}*H_{ji}}} \tag{1}$$

现有的研究表明，Archetti-Horsfall 距离是目前最鲁棒的抗原相似性的度量方法，可以有效地识别出抗原变异体。Archetti-Horsfall 距离具有对称的特点，而且当 $D_{ij}\geqslant 4$ 时，称毒株 v_i 和 v_j 抗原相异，反之称毒株 v_i 和 v_j 为抗原相似。

本文采用无向图对 n 条流感毒株 $\{v_1,v_2,\cdots,v_i,\cdots,v_n\}$ 的抗原关系进行建模，构建抗原性网络。其中，每条毒株表示为无向图中的一个节点，毒株间的已知抗原关系表示为无向图的边。

本文通过两层的 GCN 从抗原性网络中学习毒株的嵌入表示，主要完成如下的映射：

$$M=\hat{A}\,\mathrm{ReLU}(\hat{A}XW^{(0)})W^{(1)} \tag{2}$$

其中，$X\in R^{n\times n}$ 是节点的特征矩阵，$W^{(0)}$ 和 $W^{(1)}$ 是模型的权重矩阵，M 是毒株的嵌入表示；$\hat{A}=\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}$。

为了学习毒株的嵌入表示，使用 Kipf 等[18]定义的交叉熵损失函数：

$$J=-\sum_{l\in\gamma_L}\sum_{f=1}^{F}Y_{lf}\ln Z_{lf} \tag{3}$$

其中，γ_L 是具有标签的节点索引的集合；F 是输出特征的维度；Y 是标签指标矩阵。

使用 GCN 学习抗原性网络的全局结构信息，捕捉网络中蕴含的相似关系，将毒株映射为低维嵌入向量，不仅支持 H3N2 的抗原性聚类和可视化，也为融合序列数据提供支持。

2.2 HA 序列的特征表示

H3N2 流感病毒的 HA 负责通过细胞膜上的唾液酸使病毒与宿主细胞结合，是由相同亚基组成的三聚体。每个亚基由 HA1 链和 HA2 链组成，分别含有 329 和 175 个残基。其中，HA1 的突变频率高于 HA2，而且 HA1 位于流感病毒膜外，是病毒的主要抗体结合部位[10]。虽然 NA 序列中的氨基酸也可以获得替换，但是 NA 与抗体结合较弱，一般认为它不是主要的抗原决定因素[19]。因此，在本文实验中，使用 HA1 链上的氨基酸序列作为抗原特征的补充。

因为构成病毒的常见氨基酸有 20 种，但氨基酸发生变异时会出现缺省和未知变异 X，因此本文采用 21 维的向量对氨基酸进行正交编码。例如，丙氨酸(A)表示为 100000000000000000000。其次，以 3 个氨基酸为一组，使用 3 个氨基酸正交编码的或运算结果进行特征表示。例如，QKL(谷氨酰胺(Q)、赖氨酸(K)和亮氨酸(L))为 HA1 链上 3 个连续的氨基酸，如果它们的编码依次为 000000000000001000000，000000000000000010000和 001000000000000000000，则 QKL 的序列特征为 001000000000001010000。基于氨基酸三元组的特征编码，对输入长度为 m 的一条 HA1 序列，按照 $k=3$ 且步长 $s=1$ 的 k-mer 将氨基酸序列划分为 $m-2$ 个氨基酸三元组，并表示为 $\overline{M}=R^{(m-2)\times 21}$ 的向量。

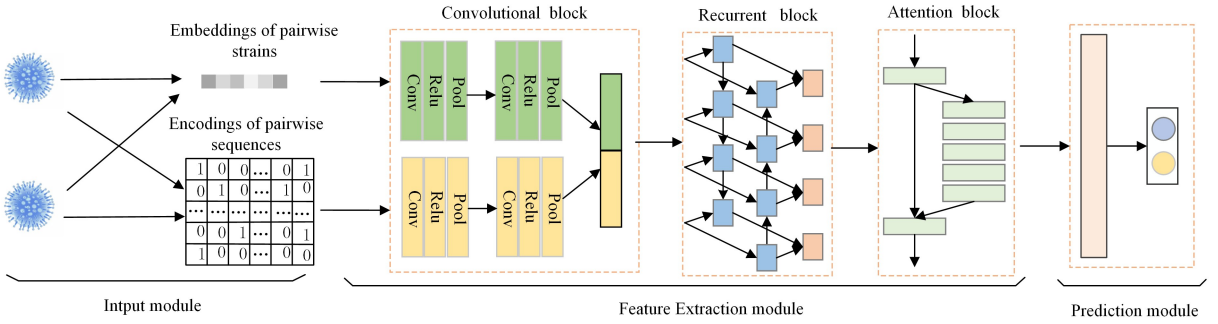
为了有效地计算毒株对之间的抗原相似性，本文将两条毒株 v_i 和 v_j 的节点嵌入向量 M_i 和 M_j 进行拼接得到 $\tilde{M}$，同时对两个毒株的 HA1 编码 $\overline{M}_i$ 和 $\overline{M}_j$ 按位进行或运算，得到毒株对的序列特征表示 $\bar{\bar{M}}$。

本文设计的抗原特征表示融合序列特征和 HI 值蕴含的相互关系，不受关键位点的局限，同时具有能够集成更多抗原特征的灵活性，为预测模型的鲁棒性和可扩展性提供支持。

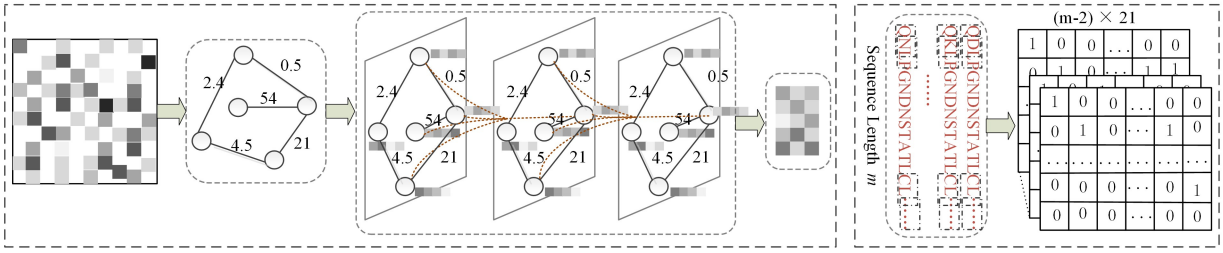
3 预测模型

基于 GCN 的 A/H3N2 流感病毒抗原性预测模型的主要框架如图 1 所示。该模型首先利用 GCN 学习毒株的低维嵌

入向量，并将其作为特征提取模块的输入，同时将毒株 HA1 的氨基酸序列编码作为抗原关系的补充输入数据，利用深度神经网络进行特征融合并捕捉关键的抗原支配特征，最后将学习到的关键特征用于抗原相似性预测。



(a)输入一对菌株并预测其抗原性



(b)基于 GCN 的毒株向量映射　　(c)将 HA1 序列编码

图 1　流感抗原性预测模型

Fig. 1　Influenza antigenicity prediction model

3.1　特征提取

为了从输入中获得更有效的抗原特征，本文集成 CNN、双向 GRU(Bidirectional Gated Recurrent Unit，BiGRU)[20] 和 SENet(Squeeze-and-Excitation Networks)[21] 设计抗原特征提取模块。该模块实现的特征映射可表示为：

$$\widetilde{X}=f_{\mathrm{SENet}}(f_{\mathrm{BGRU}}(f_C(f_C(\widetilde{M}))\,\|\,f_C(f_C(\overline{\overline{M}}))))\tag{4}$$

其中，$f_C(X)=\mathrm{Pool}(\mathrm{ReLu}(\mathrm{Conv}(X)))$，表示依次通过卷积、ReLu 激活函数和最大池化进行特征映射，并且本文通过两层叠加的 $f_C()$ 分别从毒株节点嵌入特征和氨基酸序列编码上提取特征并进行拼接，从而获得局部抗原特征。

$f_{\mathrm{BiGRU}}()$ 表示在获得局部抗原特征的基础上，使用双向 GRU 捕获数据中蕴含的长程依赖，同时有效克服梯度消失和梯度爆炸问题。具体来说，对于当前的输入 x_t，隐状态 h_t 通过从前向后获得的特征 $\overrightarrow{h}_{t-1}$ 和从后向前获得的特征 $\overleftarrow{h}_{t-1}$ 计算得到，具体表示为：

$$h_t=BiGRU(x_t,\overrightarrow{h}_{t-1},\overleftarrow{h}_{t-1})=[GRU(x_t,\overrightarrow{h}_{t-1});GRU(x_t,\overleftarrow{h}_{t-1})]\tag{5}$$

其中，h_{t-1} 表示上一个时刻的隐状态。最后，BiGRU 输出特征序列 $h=[h_1,\cdots,h_T]$。

$f_{\mathrm{SENet}}()$ 表示采用 SENet 自适应地学习通道维度上的权重，并对每个通道特征进行选择性融合。SENet 主要包括 Squeeze 和 Excitation 操作。

SENet 首先将 $f_{\mathrm{BiGRU}}()$ 获得的特征通过卷积操作映射为 $U\in R^{H\times W\times C}$，其中 C 是通道数。接下来，Squeeze 将每个通道上的空间特征编码 $u_c\in R^{H\times W}$ 压缩为一个全局特征。第 c 个通道的特征 $z_c\in R^C$ 通过下面公式进行计算：

$$z_c=\frac{1}{W\times H}\sum_{i=1}^{W}\sum_{j=1}^{H}u_c(i,j)\tag{6}$$

Excitation 操作对 Squeeze 操作获得 $z=[z_1,z_2,\cdots,z_C]$，使用两级的全连接层对每个通道的重要性进行衡量：

$$s=\sigma(W_2\delta(W_1z))\tag{7}$$

其中，$W_1\in R^{\frac{C}{r}\times C}$，$W_2\in R^{C\times\frac{C}{r}}$，$\delta$ 是 ReLu 激活函数[22]，σ 是 sigmoid 函数。

利用 Excitation 获得的权重对每个通道特征 u_c 进行重标定：

$$\widetilde{x}_c=F_{\mathrm{scale}}(u_c,s_c)=s_c\cdot u_c\tag{8}$$

其中，$F_{\mathrm{scale}}(u_c,s_c)$ 是通道上的乘积。

最后，得到特征 $\widetilde{X}=[\widetilde{x}_1,\cdots,\widetilde{x}_C]$。

3.2　输出模块

输出模块包含全连接层和输出层。在全连接层基础上，对输入的两个流感病毒的抗原得分进行计算：

$$y=\delta(W^{(p)}\widetilde{X}_c+b^{(p)})\tag{9}$$

其中，δ 是激活函数 sigmoid，$W^{(p)}$ 是权值向量，$b^{(p)}$ 是偏置量，y 是抗原标签。

损失函数为二元交叉熵：

$$l(\widetilde{y},y)=-y\ln\widetilde{y}-(1-y)\ln(1-\widetilde{y})\tag{10}$$

其中，y 是 H3N2 毒株对的真实标签，$\widetilde{y}$ 是预测标签。若 $y=1$，则毒株对为抗原相异；若 $y=0$，则毒株对为抗原相似。

3.3　模型训练

对 n 条 A/H3N2 流感毒株 $\{v_1,v_2,\cdots,v_n\}$，采用 N 对已知抗原性关系的毒株用于训练，表示为 $\{((v,v')_1,y_1),\cdots,((v,v')_N,y_N)\}$，其中 $(v,v')_i$ 定义一对毒株对，且 y_i 是训练标签。利用训练数据中的 HI 值构建包含 n 个节点的网络，并基于 GCN 学习每个毒株的节点嵌入表示，同时对每个毒株的 HA1 序列进行编码。最后，将训练数据的节点对嵌入特征和 HA1 序列编码输入模型，用于训练模型的参数。

$\widetilde{y}_i$ 表示模型在参数 θ 下的预测标签，采用最小化正则目标函数训练模型：

$$\mathcal{J}(\theta)=\sum_{i=1}^{N}l(\widetilde{y}_i,y_i)+\lambda\,\|\,\theta\,\|_2\tag{11}$$

其中，θ 为模型所有参数，λ 为正则化参数，$\lambda \| \cdot \|_2$ 是 L2-范数的正则化项，N 是训练集样本数量。

本文采用随机最小批量梯度下降算法[23]优化训练过程，使用 Adam(Adaptive Moment Estimation)[24]作为模型的优化器。同时，训练过程采用 Dropout[25]技术防止训练过程的过拟合。

4 实验设置

4.1 实验环境和数据

本文实验环境的主要参数为：处理器 Intel i7-8750H CPU 2.21 GHz，图形加速卡 NVIDIA GeForce GTX 3070 8GB，内存 32GB，操作系统 Windows10；基于开源深度学习框架 TensorFlow2.2 及其内置的 Keras 构建神经网络并进行模型的训练和测试。

本文采用了两个数据集评估模型的性能，其中，H3N2-I 包含 253 条 H3N2 流感病毒[2]，H3N2-II 包含 697 条 H3N2 流感病毒[15]。数据集的具体统计信息如表 1 所列。

表 1 数据集的统计信息

Table 1 Statistical information of datasets

数据集	毒株数量	抗原关系数量	抗原差异数量	抗原相似数量
H3N2-I	253	31878	27098	4780
H3N2-II	697	1249	606	643

4.2 模型参数设置

本文模型的主要参数为：在 GCN 嵌入中，MAX_DEGREE 为 2，丢弃率(dropout rate)为 0.5，嵌入维度为 128，L2 正则化参数 kernel_regularizer 为 0.00001，训练周期为 1000，早停阈值 early stopping 为 100。第一、二层卷积的输出空间维数 filter 分别为 64 和 128，卷积窗口长 kernel_size 分别为 5 和 3，卷积步长 strides 均为 1，填充模式 padding 均为 valid。每个卷积层之后的池化窗口大小 pool_size 为 2，步长 strides 为 2，填充模式 padding 为 valid。BiGRU 层中输出空间维数为 128。注意力层的维度变化参数为 16。输出模块中包含两个全连接层，第一个全连接层节点数为 128，第二个全连接层节点数为 32，小批量训练参数为 64。

4.3 评价指标

使用准确性(Accuracy)、精确性(Precision)、召回率(Recall)、F1 值作为评价指标，来全面地评估模型的性能。

$$Accuracy=\frac{TP+TN}{TP+TN+FP+FN} \tag{12}$$

$$Precision=\frac{TP}{TP+FP} \tag{13}$$

$$Recall=\frac{TP}{TP+FN} \tag{14}$$

$$F1=2*\frac{Precision\times Recall}{Precision+Recall} \tag{15}$$

其中，TP 为真阳性，代表真实标签和预测标签都是 1 的样本数；TN 为真阴性，代表真实标签和预测标签都是 0 的样本数；FP 为假阳性，代表真实标签是 0 和预测标签是 1 的样本数；FN 为假阴性，代表真实标签是 1 和预测标签是 0 的样本数。

5 实验和结果分析

本文通过 3 组实验分析本文模型的性能。第一组实验是与基线方法对比；第二组实验分析网络中节点的嵌入维度对模型性能的影响；第三组实验分析模型不同模块对性能的影响。

5.1 与基线模型对比

本文对比以下 6 个基线模型。

(1)Lee 的方法[8]：使用位于 5 个抗原结合区域的氨基酸变化数目建立抗原变异的预测模型；

(2)Liao 的方法[9]：根据氨基酸极性、电荷和结构对 20 种氨基酸进行分组，使用多元回归分析鉴定出 19～23 个抗原关键位点，再使用线性模型预测抗原变化；

(3)Lees 的方法[10]：更新 5 个抗原结合区域上以及附近的氨基酸，并基于这些结合区域的氨基酸变化建立预测模型；

(4)Peng 的方法[11]：利用每个抗原区域带中氨基酸的差异数作为抗原特征，建立甲型流感病毒抗原变异预测的通用模型；

(5)Yao 的方法[12]：选取 154 个非保守位点，利用 AAindex[7]提供的氨基酸相似矩阵对每条 H3N2 病毒进行特征表示，基于联合随机森林回归方法进行抗原性预测；

(6)IAV-CNN[15]：基于 ProtVec[16]表示的序列特征，利用二维卷积神经网络预测流感抗原变异。

为了评估每个模型的性能，所有模型采用相同的实验环境，实验结果均使用在可信范围内的最佳数值。其次，本文在两个流感数据集上使用 5 折交叉验证评估模型性能。表 2 列出本文模型和 6 个基线模型的性能。

表 2 与基线模型的性能对比

Table 2 Performance comparison between the proposed model and baseline models

Model	H3N2-I				H3N2-II			
	Accuracy	*Precision*	*Recall*	*F1*	*Accuracy*	*Precision*	*Recall*	*F1*
Lee 等	0.6114	1.0000	0.5239	0.6876	0.7926	0.7135	0.6025	0.6533
Liao 等	0.9210	0.8904	0.8586	0.8641	0.7071	0.7275	0.7048	0.6953
Lees 等	0.9575	0.9240	0.9104	0.9160	0.8200	0.7863	0.7712	0.7698
Peng 等	0.8677	0.7908	0.8945	0.8083	0.5099	0.4565	0.5233	0.3853
Yao 等	0.9815	0.9758	0.9529	0.9649	0.6766	0.6872	0.6788	0.6742
IAV-CNN	0.9658	0.9768	0.9813	0.9790	0.8575	0.8695	0.8483	0.8600
APGCN	**0.9998**	0.9998	**0.9997**	**0.9999**	**0.9060**	**0.8846**	**0.9045**	**0.8993**

从表 2 可以看到，本文设计的模型在两个数据集上的性能指标基本都优于基线方法，故推断利用 GCN，整合毒株之间的抗原相似关系和 HA1 序列，可以表示更丰富的抗原特征；其次，克服关键位点的局限，挖掘抗原特征之间的非线性交互关系，对流感抗原性预测具有支配作用；最后，相较于传统机器学习方法，深度学习在 H3N2 流感数据的特征学习中提供了有效的技术支持，让模型有效克服了传统机器学习中人工特征、线性建模等局限性。

5.2 节点嵌入维度对性能的影响

为分析由 GCN 获取的网络节点嵌入特征维度对模型

性能的影响，我们在确保其他参数都保持一致的情况下，设计了在 H3N2-I 数据集[2] 上嵌入维度分别为 32，64，128，256，512 的对比实验，实验结果如图 2 所示。

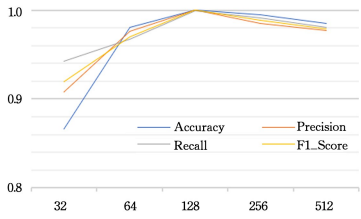


图 2　节点的不同嵌入维度的影响

Fig. 2　Effect of different embedded dimensions of nodes

从图 2 可以看出，模型预测性能首先随着嵌入维度增大上升，当到达峰值之后开始波动，也即是嵌入维度为 128 时模型性能最优，由此推断嵌入层嵌入维度大小在 128 附近时，模型可以获得最佳的性能。因此，本文在其他对实验中，毒株节点的嵌入特征设置为 128 维。

5.3　消融实验

为预测毒株间的抗原关系，在毒株嵌入向量学习和 HA1 氨基酸序列编码的基础上，集成 CNN，BiGRU 和 SENet 提取抗原支配特征，实现抗原相似性预测。

为了验证本文模型中的 GCN 嵌入单元、BiGRU 单元和 SENet 单元的有效性及其必要性，在完整模型中通过移除或者替换组件来设计 5 个变体模型。

Model-1：在完整模型上移除 GCN 嵌入单元；

Model-2：在完整模型上移除了 SENet 单元；

Model-3：在完整模型上移除了 BiGRU 单元；

Model-4：在完整模型上移除了 GCN 嵌入单元、SENet 单元和 BiGRU 单元；

Model-5：在完整模型上移除了 HA1 序列的特征编码，即该模型仅使用节点嵌入特征。

在 H3N2-I[2]，H3N2-II[15] 上进行实验对比，实验结果如表 3 所列。

表 3　不同数据集上的消融实验性能对比

Table 3　Comparison of ablation experiment performance on different data sets

Model	H3N2-I				H3N2-II			
	Accuracy	*Precision*	*Recall*	*F1*	*Accuracy*	*Precision*	*Recall*	*F1*
APGCN	**0.9998**	**0.9998**	**0.9997**	**0.9999**	**0.9060**	**0.8846**	**0.9045**	**0.8993**
Model-1	0.9264	0.9356	0.9432	0.9394	0.8376	0.7897	0.8464	0.8170
Model-2	0.9698	0.9698	0.9700	0.9699	0.8520	0.8236	0.8382	0.8308
Model-3	0.9789	0.9787	0.9855	0.9794	0.8638	0.8493	0.8282	0.8357
Model-4	0.8984	0.8982	0.9000	0.8991	0.7900	0.7300	0.7882	0.7626
Model-5	0.9684	0.9784	0.9807	0.9745	0.8839	0.8692	0.8273	0.8477

通过观察表 3 可知，对比本文设计的完整模型，移除 GCN 嵌入单元的 Model-1 模型在两个数据集上，*Accuracy* 分别降低了 7.34%，6.84%；仅使用 HA1 序列的特征编码和卷积单元的 Model-4 在两个不同数据集上，*Accuracy* 分别降低了 10.14%，11.6%；仅使用节点嵌入特征 Model-5 在两个不同数据集上 *Accuracy*，分别降低了 3.14%，2.21%。

这些结果表明，利用 GCN 获得毒株节点嵌入是一种抗原相似关系预测的有效途径；其次，相比较氨基酸序列特征，节点嵌入特征在本文设计的模型中对模型的性能具有支配作用；此外，序列特征也为模型性能的提升提供了补充作用。

第二，对比本文设计的完整模型，去除 SENet 单元的 Model-2 模型在两个数据集上，*Accuracy* 分别降低了 3%，5.4%；去除 BiGRU 单元的 Model-3 模型在两个不同数据集上，*Accuracy* 分别降低了 2.09%，4.22%。

这些结果表明 BiGRU 单元、SENet 单元对 H3N2 的抗原相似性预测是有效的；其次，SENet 单元通过通道上的注意力机制捕捉输入特征中有区分度的抗原特征，为改进模型性能提供了良好的支持。

结束语　本文基于血凝抑制实验获得的抗原数据，利用 GCN 获取 A/H3N2 毒株的嵌入向量，融合 HA1 的氨基酸序列编码，构建一个端到端的 A/H3N2 抗原相似性预测模型。在两个数据集上的实验结果表明，该方法相比其他同类方法，显著提升了抗原相似性预测性能，具有良好的鲁棒性。其次，从实验中可以看出，基于网络嵌入技术可以有效地获取抗原特征，为模型提供良好的可解释性和可扩展性。

协同学习网络的拓扑结构和节点氨基酸序列，获得整合抗原相似关系和氨基酸序列的嵌入特征，是未来改进的研究方向，不仅直接支持相似性之外的抗原性分析，也能进一步提高模型的可解释性和实用性。此外，H3N2 是一种典型的甲型流感病毒，作为和 H3N2 类似的其他亚型流感病毒，如 H5N1 和 H7N9，本文的预测模型也为它们基于深度学习进行抗原性预测提供了基础。

参 考 文 献

[1] BLAIR R H，DAWSON E D，TAYLOR A W，et al. Clinical validation of the FluChip-8G influenza a＋ B assay for influenza type and subtype identification[J]. Journal of Clinical Virology，2019，118：20-27.

[2] 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.

[3] SUN H，YANG J，ZHANG T，et al. Using sequence data to infer the antigenicity of influenza virus[J]. MBio，2013，4(4)：e00230-13.

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

[5] World Health Organization. Serological detection of avian influenza A(H7N9) virus infections by modified horse red blood cells haemagglutination-inhibition assay[Z]. 2013.

[6] AMPOFO W K，AZZIZ-BAUMGARTNER E，BASHIR U，et al. Strengthening the influenza vaccine virus selection and development process：Report of the 3rd WHO Informal Consultation for Improving Influenza Vaccine Virus Selection held at WHO headquarters，Geneva，Switzerland，1-3 April 2014[J].

Vaccine, 2015, 33(36): 4368-4382.
[7] KOEL B F, BURKE D F, BESTEBROER T M, et al. Substitutions near the receptor binding site determine major antigenic change during influenza virus evolution[J]. Science, 2013, 342(6161): 976-979.
[8] LEE M S, CHEN J S E. Predicting antigenic variants of influenza A/H3N2 viruses[J]. Emerging Infectious Diseases, 2004, 10(8): 1385.
[9] 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.
[10] 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.
[11] 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]. Scientific Reports, 2017, 7(1): 42051.
[12] YAO Y, LI X, LIAO B, et al. Predicting influenza antigenicity from Hemagglutintin sequence data based on a joint random forest method[J]. Scientific Reports, 2017, 7(1): 1545.
[13] KAWASHIMA S, POKAROWSKI P, POKAROWSKA M, et al. AAindex: amino acid index database, progress report 2008[J]. Nucleic Acids Research, 2007, 36(suppl_1): D202-D205.
[14] NEHER R A, BEDFORD T, DANIELS R S, et al. Prediction, dynamics, and visualization of antigenic phenotypes of seasonal influenza viruses[J]. Proceedings of the National Academy of Sciences, 2016, 113(12): E1701-E1709.
[15] 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 Biology and Bioinformatics, 2021, 19(6): 3497-3506.
[16] ASGARI E, MOFRAD M R K. Continuous distributed representation of biological sequences for deep proteomics and genomics[J]. PloS One, 2015, 10(11): e0141287.
[17] ARCHETTI I, HORSFALL JR F L. Persistent antigenic variation of influenza A viruses after incomplete neutralization in ovo with heterologous immune serum[J]. The Journal of Experimental Medicine, 1950, 92(5): 441-462.
[18] KIPF T N, WELLING M. Semi-supervised classification with graph convolutional networks[J]. arXiv: 1609.02907, 2016.
[19] MORRIS D H, GOSTIC K M, POMPEI S, et al. Predictive modeling of influenza shows the promise of applied evolutionary biology[J]. Trends in Microbiology, 2018, 26(2): 102-118.
[20] CHUNG J, GULCEHRE C, CHO K H, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling[J]. arXiv: 1412.3555, 2014.
[21] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 7132-7141.
[22] NAIR V, HINTON G E. Rectified linear units improve restricted boltzmann machines[C]// Proceedings of the 27th International Conference on Machine Learning (ICML-10). 2010: 807-814.
[23] LI M, ZHANG T, CHEN Y, et al. Efficient mini-batch training for stochastic optimization[C]// Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2014: 661-670.
[24] KINGMA D P, BA J. Adam: A method for stochastic optimization[J]. arXiv: 1412.6980, 2014.
[25] SRIVASTAVA N, HINTON G, KRIZHEVSKY A, et al. Dropout: a simple way to prevent neural networks from overfitting[J]. Journal of Machine Learning Research, 2014, 15(1): 1929-1958.

HE Minglong, born in 1998, postgraduate. His main research interests include deep learning and bioinformatics.

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