计算机科学
COMPUTER SCIENCE

基于transformer的门控双塔模型预测H1N1流感抗原性

李川, 李维华, 王迎晖, 陈伟, 文俊颖

引用本文

李川, 李维华, 王迎晖, 陈伟, 文俊颖基于transformer的门控双塔模型预测H1N1流感抗原性[J]. 计算机科学, 2022, 49(11A): 211000209-6.

LI Chuan, LI Wei-hua, WANG Ying-hui, CHEN Wei, WEN Jun-ying. Gated Two-tower Transformer-based Model for Predicting Antigenicity of Influenza H1N1 [J]. Computer Science, 2022, 49(11A): 211000209-6.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.211000209

基于 transformer 的门控双塔模型预测 H1N1 流感抗原性

李 川 李维华 王迎晖 陈 伟 文俊颖
云南大学信息学院 昆明 650503
(lichuanfirst163@163.com)

摘 要 流感病毒血凝素蛋白的快速演变导致新的病毒株不断产生,新的病毒株可能引起季节性流感甚至全球流感大爆发。及时检测出抗原变异体对疫苗的筛选和设计至关重要。鲁棒的抗原性预测模型是应对疫苗挑战的有效方法。各种端到端的特征学习工具为蛋白组学提供了良好的特征表示方法,但是现有的甲型流感预测模型还不能有效地提取并利用血凝素蛋白氨基酸序列中的特征。基于 transformer 设计一个门控双塔模型,通过输入甲型流感病毒血凝素蛋白的氨基酸序列,利用两个并行的编码器分别从血凝素蛋白氨基酸序列的时间维和空间维上捕捉抗原特征,并学习特征与预测结果间的非线性关系。为了减少数据中的噪声,融合时间维与空间维上的特征时,通过门机制自适应地获取衡量它们相对重要性的权重进行选择性融合,最后使用融合特征预测 H1N1 流感抗原变异株。在 H1N1 数据集上的实验结果表明,该模型利用优秀的非线性特征学习能力提高了抗原变异的预测性能,同时具有良好的鲁棒性。

关键词:甲型流感;H1N1;抗原性预测;transformer;双塔模型;门机制
中图法分类号 TP391

Gated Two-tower Transformer-based Model for Predicting Antigenicity of Influenza H1N1

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

Abstract The rapid evolution of influenza virus hemagglutinin protein has led to the continuous production of new virus strains, which may cause seasonal influenza and even global influenza outbreaks. Timely detection of antigen variants is essential for vaccine screening and design. Therefore, a robust predictive model of antigenicity is an effective method to deal with the challenge of vaccines. Various end-to-end feature learning tools provide good feature representation methods for proteomics, but the existing influenza A prediction models cannot effectively extract and utilize features in amino acid sequences. In this paper, a gated two-tower model is designed based on the transformer. By inputting the amino acid sequence of the influenza A virus hemagglutinin protein, two parallel encoders are used to capture the antigenic characteristics from the time and space dimensions of the hemagglutinin protein amino acid sequence, and learn the nonlinear relationship between features and prediction results. In order to reduce the noise in the data, when fusing the features in the time dimension and the space dimension, the weights that measure their relative importance are adaptively obtained through the gate mechanism for selective fusion, and finally the fusion features are used to predict the H1N1 influenza antigen variants. Experimental results on the H1N1 data set show that the use of the model's excellent non-linear feature learning ability improves the predictive performance of antigenic variation, and at the same time has good robustness.

Keywords Influenza A, H1N1, Antigenicity prediction, Transformer, Two-tower mode, Gate mechanism

1 引言

季节性流感是对公共卫生和全球经济的严重威胁,在全球每年造成多达 50 万人死亡和数百万人患病[1]。甲型流感病毒中的 H1N1 亚型是流感病毒传播的主要亚型之一[2]。目前接种疫苗是预防流感病毒最有效的手段[3]。流感病毒表面糖蛋白血凝素(Hemagglutinin,HA)作为宿主免疫的主要靶点,容易发生突变[4]。靶点上积累的突变会导致病毒产生抗原漂移(Antigenic Drift),使病毒可以逃避免疫,给疫苗设计带来了巨大的挑战[5]。因此,快速、可靠地确定病毒株抗原性对于疫苗设计和流感监测至关重要。目前,抗原距离来源于免疫分析,如血凝抑制(Hemagglutination Inhibition,HI)[6]。然而,该方法成本高昂且费时费力,所以基于计算方法的抗原性预测成为理想的替代方法。

目前,基于计算的抗原性预测主要是传统机器学习模型。Smith 等使用 HI 数据创建了抗原图谱,并确定了 1968 至 2003 年间甲型 H3N2 流感病毒的抗原进化[7]。Lees 等利用 5 个抗原位点及周围 131 个氨基酸的变化对病毒株之间的抗原距离进行建模[8]。Liao 等提出一种结合评分和回归方法来预测抗原变异的方法[9]。Zhou 等提出一种蛋白质序列的

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)

编码方案，用于预测不同甲型流感病毒的抗原性。该方案将病毒的 HA 序列数据编码成一个数值矩阵，然后利用机器学习模型预测病毒抗原性[10]。Peng 等使用保守的抗原结构建立甲型流感病毒抗原变异预测的通用模型[11]。Yin 等综合多种传统机器学习模型，提出了一种预测 H1N1 流感病毒抗原变异的集成模型[12]。传统机器学习模型仅需要少量数据，但是模型却依赖于特征工程；其次，传统机器学习模型往往建立在有限的氨基酸序列关键位点之上。然而，流感病毒突变率非常高，如果下一代毒株突变的位点超出预测模型的关键位点，那么已建立模型的鲁棒性就会降低。

深度学习可以尽可能避免特征工程依赖并对特征之间的非线性关系进行建模，成为生物医疗信息处理的新兴技术[13]。经典的神经网络模型，包括卷积神经网络（Convolutional Neural Networks，CNN）、循环神经网络（Recurrent Neural Network，RNN）以及含门控机制的长短期记忆网络（Long Short-Term Memory，LSTM）和门控循环单元（Gated Recurrent Unit，GRU）等，被广泛应用到基因组学和蛋白组学的预测分析[14-15]。Yin 等使用二维卷积结合残差网络，基于 HA 序列特征设计一个流感病毒抗原性预测的模型[16]，提高了预测精度。然而，CNN 并不能有效地捕捉 HA 序列上氨基酸之间的远程关联。RNN 以及变体虽然可以捕捉长距离的依赖，但是因为每一个状态的更新依赖于上一个状态，因此无法实现有效的并行计算。transformer 是一种 encoder-decoder 结构，利用 self-attention 和 Position Embedding 不仅可以捕捉长序列依赖，也可以实现并行训练[17]。基于深度学习的各种端到端特征学习工具为蛋白组学提供各种嵌入方式，不仅显著提高了蛋白组学的预测性能，也为蛋白组学的研究提供了独特的视角。现有的研究表明序列特征表示中的空间维特征和时间维特征同样影响模型预测的性能，但是现有预测模型还不能有效地利用序列中两个维度上的特征。

本文针对甲型 H1N1 抗原性进行预测，基于 transformer 设计了一个门控双塔模型（Gated Two-tower Transformer-based Model，GTDM）。该模型并行地从氨基酸序列的时空维上学习抗原特征，通过门机制选择性地融合，然后用于 H1N1 抗原性预测。本文的贡献主要有以下几点：

（1）基于 transformer 设计门控双塔模型，利用并行的 encoder 捕捉序列中时空维上蕴含的特征及其非线性关系，同时通过门机制自适应地衡量它们的相对重要性，并生成对应的权重，通过权重进行选择性融合；

（2）在 H1N1 数据集上的实验结果表明，与现有的模型相比，GTDM 有效地提升了模型非线性特征学习能力和抗原变异的预测性能。

2 GTDM 模型

本文设计的 GTDM 模型如图 1 所示。该模型基于输入病毒V_i和V_j对应的 HA 序列，预测V_i和V_j之间的抗原相似关系。GTDM 模型包含数据处理、输入模块、特征提取模块、特征信息融合模块和输出模块。

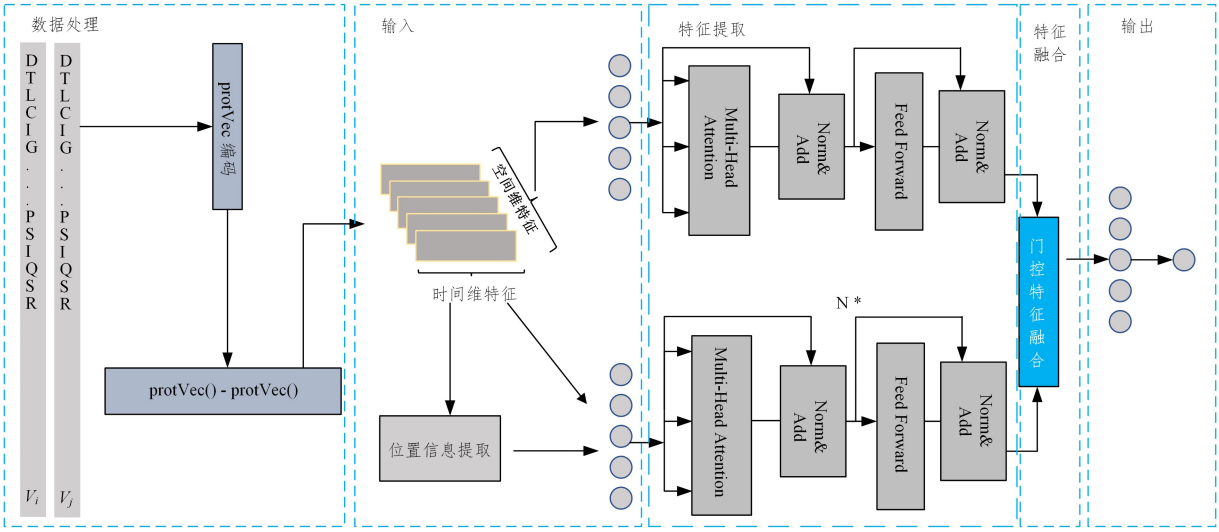

图 1 GTDM 模型的结构图

Fig. 1 Architecture of GTDM model

数据处理模块将输入的 HA 序列转换为对应的特征矩阵，然后做差运算；输入模块接收特征矩阵并与位置向量进行融合；特征提取模块对接收到的数据分别在时间维和空间维上提取特征；特征融合模块使用门（gate）机制学习两个维度上特征的权重并进行融合，得到具有区分力的抗原特征；输出模块利用融合的抗原特征预测是否产生变异。

2.1 数据处理

氨基酸序列需要转化成能够被模型识别计算的数字特征，获得氨基酸序列对应的特征表示是预测和建模的第一步。根据不同的氨基酸编码方式，可以获得 HA 序列特征的不同表达方式。

目前，氨基酸编码方式大致可以分为二进制编码（Binary）、基于理化性质的编码（Physicochemical Properties）、基于进化信息的编码（Evolution-based）、基于结构的编码（Structure-based）和机器学习编码（Machine-learning）5 类。经过实验验证，本文使用机器学习编码中的 ProtVec 编码可以更大程度地发挥模型的性能（详细实验数据见表 2）。ProtVec 编码是一种预训练的氨基酸编码，基于词嵌入技术在蛋白质数据上训练获得氨基酸三元组的嵌入向量[18]。

如图 2 所示，在 ProtVec 中每 3 个氨基酸映射为一个 100 维的特征向量[18]。对输入长度为 n 的 HA 序列，按照 $k=3$ 且步长 $s=1$ 的 k-mer[19]将氨基酸序列划分为 $n-2$ 个氨基酸

三元组。采用 protVec 将毒株映射到$\mathbb{R}^{(n-2)\times d}$的向量空间中。

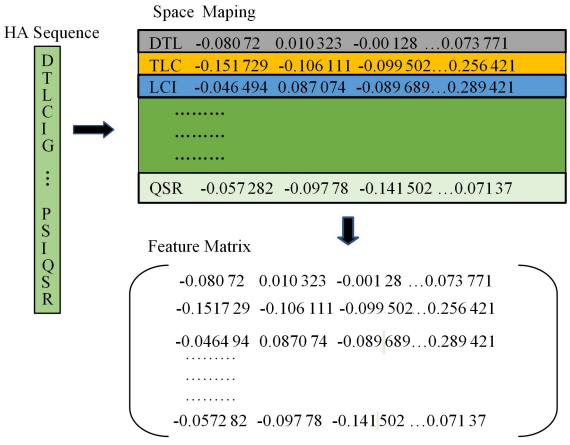

图 2　protVec 编码示意图

Fig. 2　Schematic diagram of protVec encoding

2.2　输入模块

对两条不同的毒株V_i和V_j进行编码，分别得到$\boldsymbol{X}_i$和$\boldsymbol{X}_j$。通过矩阵差运算获得毒株对的特征表示：

$$\widetilde{\boldsymbol{X}}=\boldsymbol{X}_i-\boldsymbol{X}_j \tag{1}$$

$\widetilde{\boldsymbol{X}}\in\mathbb{R}^{(n-2)\times d}$，且每一行元素表示 HA 序列对相应位置的氨基酸变化特征。

由于 transformer 基于注意力机制实现，所以无法利用序列上的顺序信息[17]。transformer 中引入单独计算位置信息的 Position Embedding(PE)：

$$\boldsymbol{PE}_i^t=\begin{cases}\sin\left(\dfrac{t}{10\,000^{\frac{i}{d}}}\right), & i=2k,k<\dfrac{d}{2},k\in\mathrm{N}\\ \cos\left(\dfrac{t}{10\,000^{\frac{i-1}{d}}}\right), & i=2k+1,k<\dfrac{d}{2},k\in\mathrm{N}\end{cases} \tag{2}$$

其中，d 是氨基酸三元组的嵌入维度，t 是一个氨基酸三元组在 HA 序列上的位置且 $0\leqslant t<n-2$，i 表示 t 位置向量上的第 i 维，且 $0\leqslant i<d$。

位置信息$\boldsymbol{PE}^t$与输入的嵌入向量$\boldsymbol{X}^t$具有相同的维数，所以二者相加产生了一个新的特征表示：

$$\boldsymbol{X}^t=\boldsymbol{PE}^t\oplus\widetilde{\boldsymbol{X}}^t \tag{3}$$

其中，$\oplus$是对应元素相加。因为只有时间维度上的 encoder 才能利用氨基酸三元组的顺序信息，所以 $\boldsymbol{PE}$ 仅在时间维上的 encoder 中与毒株对特征进行融合。

在接下来的特征提取模块中，并行 encoder 采用相同的结构学习特征，所以将它们的输入都表示为$\widetilde{\boldsymbol{X}}$。

2.3　特征提取模块

GTDM 模型包括两个并行的 encoder 块，分别从输入特征的时间维和空间维上学习抗原特征。encoder 块中的每个 encoder 由堆叠的多头注意力层(Multi-Head Attention)和前馈网络层(Feed Forward)组成，这两层分别连接着残差正则化层(Add&Norm)[17]。

Multi-Head Attention 由多个自注意力(self-attention)子模块组成[17]。self-attention 更新方式如下：

$$self\_Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=softmax\left(\frac{\boldsymbol{QK}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{4}$$

$$head_i=self\_Attention(\boldsymbol{QW}_i^Q,\boldsymbol{KW}_i^k,\boldsymbol{VW}_i^v) \tag{5}$$

$$\boldsymbol{X}_{\mathrm{att}}=MultiHead(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=concat(head_1,\cdots,head_h)\boldsymbol{W}^O \tag{6}$$

其中，$\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}$ 是通$\widetilde{\boldsymbol{X}}$进行不同的线性变换之后得到的，$\sqrt{d_k}$是缩放因子，$\boldsymbol{W}_Q,\boldsymbol{W}_k,\boldsymbol{W}_v$是模型学习到的参数。self-attention 可以在不同位置共同关注来自不同表示子空间的信息，最后将不同的 self-attention 结果拼接起来就得到多头注意力特征[17]。

Add&Norm 表示残差连接和正则化，防止模型过深带来的梯度消失或梯度爆炸问题。它们分别表示如下：

$$\bar{\boldsymbol{X}}=\widetilde{\boldsymbol{X}}+\boldsymbol{X}_{\mathrm{att}} \tag{7}$$

$$\bar{\bar{\boldsymbol{X}}}=\mathrm{norm}(\bar{\boldsymbol{X}}) \tag{8}$$

Feed forward 是一个全连接层：

$$\boldsymbol{X}=\max(0,\bar{\bar{\boldsymbol{X}}}\boldsymbol{W}_1+\boldsymbol{b})\boldsymbol{W}_2+\boldsymbol{b}_2 \tag{9}$$

其中，$\boldsymbol{W}_1,\boldsymbol{W}_2,\boldsymbol{b}_2$是模型参数。

因为 GTDM 模型采用的双塔机制，所以模型同时从时间维和空间维上捕捉，得到特征$\boldsymbol{X}_t$和$\boldsymbol{X}_s$。

2.4　特征融合模块

虽然 GTDM 模型采用双塔机制捕捉到时间维和空间维上的特征$\boldsymbol{X}_t$和$\boldsymbol{X}_s$，但是这些特征在模型预测中并不是同等重要。因此，GTDM 模型基于 gate 机制自适应地为特征赋予权重并进行选择性融合。融合门结构如图 3 所示。

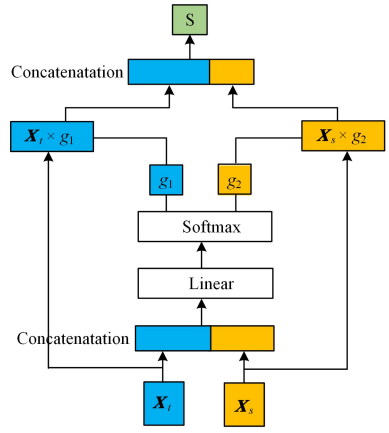

图 3　融合门结构

Fig. 3　Structure of fusion gate

融合门的映射如下：

$$\boldsymbol{Z}=\boldsymbol{W}_g\times\mathrm{Concat}(\boldsymbol{X}_t,\boldsymbol{X}_s)+\boldsymbol{b}_g \tag{10}$$

其中，$\boldsymbol{W}_g$是训练模型的参数，$\boldsymbol{b}_g$是偏差，Concat 表示对特征$\boldsymbol{X}_t$和$\boldsymbol{X}_s$进行融合。映射得到向量 $\boldsymbol{Z}=(z_1,z_2)$。将 $\boldsymbol{Z}$ 输入到 softmax 函数得到权重：

$$g_i=\frac{\mathrm{e}^{z_i}}{\sum_{j=1}^{2}\mathrm{e}^{z_j}} \tag{11}$$

根据权重g_1,g_2，将权重值分配给$\boldsymbol{X}_t$和$\boldsymbol{X}_s$进行融合，最终得到特征数据 $\boldsymbol{S}$：

$$\boldsymbol{S}=\mathrm{Concat}(\boldsymbol{X}_t\times g_1,\boldsymbol{X}_s\times g_2) \tag{12}$$

2.5　输出模块

输出模块将融合特征 $\boldsymbol{S}$ 依次通过全连接层和 Softmax 层，预测毒株对之间的抗原相似性：

$$y=\mathrm{softmax}(\boldsymbol{w}\times\boldsymbol{S}+\boldsymbol{b}) \tag{13}$$

其中，矩阵 $\boldsymbol{w}$ 和向量 $\boldsymbol{b}$ 为模型的参数。如果抗原相异，则 $y=1$；反之抗原相似，则 $y=0$。

2.6　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

H1N1 毒株对的 HA 序列以及对应的标签。每对 H1N1 毒株的标签根据抗原距离确定：

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

其中，H_{ij} 表示根据菌株 V_i 的抗血清获得的抑制菌株 V_j 的抗原的抗体的 HI 滴度。如果 $d_{ij}\geqslant 4$，称毒株 V_i 和 V_j 为抗原差异，反之称毒株 V_i 和 V_j 为抗原相似。本文对训练模型定义下面的目标函数：

$$J(\theta)=-\frac{1}{M}\sum_{k=1}^{M}(y\times lb(f(\widetilde{\boldsymbol{X}}))+(1-y)\times lb(1-f(\widetilde{\boldsymbol{X}}))) \tag{15}$$

其中，θ 是模型的参数，$f(\widetilde{\boldsymbol{X}})$ 表示在 θ 参数条件下所预测的标签，M 是输入的毒株对数量。本文使用随机梯度下降 Adam 算法(Adaptive Moment Estimation)[20]，最小化目标函数 $J(\theta)$ 来对模型进行训练，同时采用 dropout 策略[21]来缓解模型过拟合问题。

3 实验设置

3.1 实验环境和数据

实验环境：处理器 Intel(R) Core i7-10700K CPU 3.80GHz，图形加速卡 NVIDIA GeForce GTX 3070 8GB，内存 32GB，操作系统 Windows10；采用 pytorch1.7 深度学习框架。

本文使用 Yin 论文中提供的甲型 H1N1 流感病毒数据集，分析模型的性能。该数据集中包括 1 562 对 H1N1 流感病毒抗原关系数据、294 条 H1N1 流感病毒 HA 序列数据[16]。单独的一条 HA 序列中包含 327 个氨基酸，序列中的部分氨基酸位置存在缺失。

3.2 模型的参数和评价指标

GTDM 模型主要设定输入模块、特征提取模块中的相关参数，具体参数设置如下：根据输入的向量形状为 $\widetilde{\boldsymbol{X}}\in\mathbb{R}^{325\times 100}$，设置时间维数为 325，空间维数为 100。为了保证衔接处各个模块的维度相同，输入模块中的 Embedding 层将输入数据统一映射成稠密度为 64 维的向量。在特征提取模块，双塔中的 encoder 数量都设置为 8。在单独的 encoder 模块中，隐藏层维度设置为 16，多头注意力的头数为 8。初始学习率(learning rate)设为 0.0001；使用 dropout 策略[21]来缓解模型过拟合问题，且丢弃率(dropout rate)设为 0.2；使用随机 Minibatch 算法来优化训练过程，其中的小批量训练参数均设置为 25。对于基线模型的其他参数，我们按照对应论文中推荐的参数进行设置。

本文采用准确率(accuracy)、精确率(precision)、召回率(recall)、F1 分数($F1score$)、马修斯相关系数(Matthews Correlation Coefficient，MCC)作为评价的指标：

$$accuracy=\frac{tp+tn}{tp+tn+fp+fn} \tag{16}$$

$$precision=\frac{tp}{tp+fp} \tag{17}$$

$$recall=\frac{tp}{tp+fn} \tag{18}$$

$$F1=\frac{precision\times recall\times 2}{precision+recall} \tag{19}$$

$$MCC=\frac{tp\times tn-fp\times fn}{\sqrt{(tp+fp)(tp+fn)(tn+fp)(tn+fn)}} \tag{20}$$

其中，tp 表示真实标签和预测标签都是 1 的样本数，tn 表示真实标签和预测标签都是 0 的样本数，fp 表示真实标签是 0 和预测标签是 1 的样本数，fn 表示真实标签是 1 和预测标签是 0 的样本数。按照 0.8∶0.2 将数据划分为训练集和测试集，用五倍交叉验证进行训练，并在测试集上验证泛化能力。

4 实验和结果分析

4.1 与基准模型对比

为了评估本文所构建模型的性能，使用 6 个基线模型与 GTDM 进行对比，实验结果如表 1 所列。

表 1 与基准模型的对比

Table 1 Comparison of our model with baselines

Model	*accuracy*	*precision*	*recall*	*F1*	*MCC*
Lees 等[8]	0.726	0.718	0.718	0.709	0.433
Peng 等[11]	0.715	0.747	0.741	0.713	0.480
Liao 等[9]	0.753	0.760	0.744	0.730	0.487
Zhou 等[10]	0.859	0.896	0.882	0.889	0.756
Tran 等[12]	0.908	0.938	0.902	0.924	0.812
Yin 等[16]	0.920	0.928	0.915	0.924	0.806
GTDM	0.946	0.968	0.942	0.955	0.887

表 1 中前 4 个模型是传统的机器学习模型，后 2 个是深度学习模型。按照论文中所给的实验参数和模型结构进行模型搭建。从表 1 中实验结果可以看到，基于深度学习的方法性能优于基于传统机器学习的方法。传统机器学习建模局限于有限的突变位点和突变区域，难以从全局上去获取 HA 序列上的氨基酸突变信息。此外，传统的机器学习模型对非线性的特征学习能力相对薄弱。深度学习不依赖于 HA 序列上特定的突变位点和突变区域，同时具有优秀的非线性特征学习能力，可以有效提升模型的预测精度和鲁棒性。从表 1 中可以看出，相比目前 Yin 等人提出的最优 IAV_CNN[16]模型，GTDM 的 *accuracy* 上升了 2.6%，*precision* 上升了 4%，*recall* 上升了 2.7%，*F1* 上升了 3.1%，*MCC* 上升了 8.1%。实验结果表明，基于 transformer 的 GTDM 可以有效捕捉序列上局部依赖和远程依赖关系，有效克服 Yin 等人基于二维卷积提出的 IAV_CNN[16]在序列特征学习中的局限性。

4.2 氨基酸特征表示

为了能够充分发挥出模型的性能，我们对比了 5 类共 12 种氨基酸编码方式在 GTDM 模型和 H1N1 数据集上的实验效果。其中，Hydrophobicity_matrix(HM)[22]，Acthely_factors(AF)[23]和 Meiler_parameter(MP1)[24]映射向量维数分别为 20，7，5，这 3 种编码方式是基于氨基酸的物理化学性质的理化性质编码；PAM250[25]，Blosum62[26]，PSSM[27] 3 种编码映射维数皆为 20，在转换 HA 序列特征信息中加入了进化信息的进化编码；miyazawa_energie(ME)[28]，Micheletti potentials(MP2)[29]映射维数都为 20，两种编码方式中加入了对应蛋白质序列的结构信息；Onehot20 采用 20 维的独热码来不重复地表示单个氨基酸；AESNN3[30]，ANN4D[23]，protVec[18]映射维数分别为 3，4，100，是采用词嵌入技术结合蛋白质数据库进行模型训练得到的机器学习编码。

最终结果如表 2 所列，可以看出 protVec 编码的准确率、准确率、召回率、F1 分数和 MCC 的性能达到了 94.6%，96.8%，94.2%，95.5%，88.7%，总体上比其他的氨基酸编码方式的预测结果更加理想，其余编码方式的预测精度最高为

92.3%。因为其余氨基酸编码方式将氨基酸映射到最多 20 维的向量空间中，数据的维数少，GTDM 模型难以发挥自身的优势学习更复杂的维度关系。protVec 编码的映射维数为 100，特征维数大大增加，特征表达更加准确、全面。因此，GTDM 模型可以学习到更复杂的维度关系，充分发挥了深度学习模型的学习优势。

表 2　氨基酸编码方式的对比

Table 2　Comparison of amino acid coding methods

Encoding	*accuracy*	*precision*	*recall*	*F*1	*MCC*
HM	0.888	0.948	0.863	0.904	0.776
MP1	0.917	0.956	0.905	0.930	0.830
AF	0.920	0.946	0.921	0.933	0.834
PAM250	0.901	0.921	0.916	0.918	0.793
Blosum62	0.907	0.940	0.905	0.922	0.809
PSSM	0.920	0.961	0.905	0.932	0.837
ME	0.923	0.966	0.905	0.935	0.845
MP2	0.904	0.930	0.911	0.920	0.801
Onehot20	0.920	0.961	0.905	0.932	0.837
AESNN3	0.904	0.940	0.900	0.919	0.803
ANN4D	0.898	0.925	0.905	0.915	0.787
protVec	**0.946**	**0.968**	**0.942**	**0.955**	**0.887**

4.3　深度学习模型对比

为了验证 GTDM 模型在深度学习模型中的预测效果，我们将其与循环神经网络中的门控循环网络（GRU）、双向循环神经网络（BRNN）以及结合门控循环神经网络和注意力机制的 GRU_ATTENTION 进行对比。实验结果如表 3 所列。

表 3　深度学习模型对比

Table 3　Deep learning model comparison

Model	*accuracy*	*precision*	*recall*	*F*1	*MCC*
GRU	0.850	0.860	0.883	0.871	0.692
BRNN	0.856	0.722	0.591	0.650	0.565
GRU_ATTENTION	0.907	0.947	0.889	0.911	0.815
GTDM	**0.946**	**0.968**	**0.942**	**0.955**	**0.887**

3 个循环神经网络的预测准确度均未低于 85%，高于 Lees[8]，Peng[11]，Liao[9] 等提出的传统机器学习模型，其中 Zhou[10] 预测准确度略高于 GRU 和 BRNN 但依然低于 GRU_ATTENTION。实验结果验证了传统机器学习模型存在局限性，而深度学习模型却能在一定程度上克服这些局限性，提升预测效果。此外，GTDM 在深度学习模型中依然表现出优秀的预测性能以及鲁棒性。

4.4　消融实验

为了验证本文 GTDM 模型中各个模块的有效性和必要性，本文设计下面 3 个变体模型。

$GTDM_{\alpha}$：移除 GTDM 中时间维上的 encoder。

$GTDM_{\beta}$：移除 GTDM 中空间维上的 encoder。

$GTDM_{\gamma}$：在 GTDM 使用相同权重替换 gate 机制融合。

将 GTDM 与 3 个变体模型在 H1N1 数据集上进行实验，实验结果如表 4 所列。从表 4 可知，GTDM 的 *accuracy*，*Precision*，*Recall*，*F*1，*MCC* 分别比 3 个变体模型中的最好性能高出 1.6%，1.1%，1.6%，1.5%，1.9%。其次，$GTDM_{\alpha}$ 模型的性能最差，这表明时间维上的 encoder 提取得到的抗原特征对抗原性预测具有支配作用，空间维上的 encoder 对抗原特征进行了有效的补充。另外，进行相同权重特征融合的 $GTDM_{\gamma}$ 模型的预测性能相比仅计算时间维上 $GTDM_{\beta}$ 模型还低，这表明 gate 机制可以自适应地权衡时间维特征和空间维特征对抗原性预测的影响权重，过滤掉噪音，并选择有用的语义上下文。上述的消融实验表明，同时关注了序列时间维和空间维上的特征对于 H1N1 的抗原性预测是有意义的；其次，基于 transformer 的双塔模型不仅可以有效地提取时间维和空间维上的特征，还通过 gate 机制将其进行了有效的融合。

表 4　消融实验

Table 4　Ablation study results

Model	*accuracy*	*precision*	*recall*	*F*1	*MCC*
$GTDM_{\alpha}$	0.917	0.936	0.926	0.931	0.826
$GTDM_{\beta}$	0.930	0.957	0.911	0.940	0.858
$GTDM_{\gamma}$	0.920	0.951	0.916	0.933	0.835
GTDM	**0.946**	**0.968**	**0.942**	**0.955**	**0.887**

4.5　鲁棒性分析

为了逃避免疫，流感病毒血凝素蛋白会随着时间快速突变，持续和累积变化会产生新的抗原株。因此，评估利用历史数据训练的模型在未来季的预测性能是衡量模型鲁棒性的重要方面。

本文分别使用 2016 年、2015－2016 年、2014－2016 年的数据作为测试数据，再分别使用除开验证集部分的数据作为训练数据，同时与 Yin 等提出的 IAV_CNN[16] 模型进行对比。实验结果如表 5 所列。

表 5　鲁棒性分析

Table 5　Robustness analysis

TestData	Model	*accuracy*	*precision*	*recall*	*F*1	*MCC*
2016	GTDM	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	IAV_CNN	0.960	0.900	1.000	0.947	0.919
2015－2016	GTDM	**0.928**	**0.909**	**0.909**	**0.909**	**0.882**
	IAV_CNN	0.835	0.810	0.586	0.680	0.586
2014－2016	GTDM	**0.921**	**0.923**	**0.800**	**0.857**	**0.807**
	IAV_CNN	0.812	0.762	0.533	0.627	0.521

从表 5 的实验结果可以看到，模型在预测未来一年（2016 年）的性能最好，随着预测未来的时间跨度增大，模型性能下降。这是因为随着时间跨度增长，氨基酸突变累积增大，模型预测的性能下降。然而，随着时间跨度的增大，IAV_CNN 模型的预测效果开始出现大幅的降低，但 GTDM 模型下降趋势平缓，且预测效果依然优于 IAV_CNN。在 2016 年、2015－2016 年、2014－2016 年的测试数据上，GTDM 模型的 *accuracy* 分别高出 IAV_CNN 模型 4%，12.6%，10.9%。这个结果表明，GTDM 具有更好的鲁棒性。

结束语　本文针对甲型流感病毒亚型 H1N1 的抗原性预测，基于 transformer 设计了一个门控双塔模型来联合表示建模蛋白质组数据的复杂空间和时间依赖模式。特别地，该模型利用两个并行的编码器分别从氨基酸序列时空维上捕捉特征及其非线性关系，并设计 gate 融合时空维上的特征之间的相互作用，然后用于 H1N1 抗原性预测。在 H1N1 数据集上的实验结果表明，所提模型可以有效地提升模型非线性特征学习能力，并提高了抗原变异的预测性能和鲁棒性。

整合蛋白质的其他特征到序列中，提高蛋白质特征表示的多维度，是改进流感抗原性预测的一个研究方向。另外，进一步探讨其他自适应选择有用特征的方法，也可以为模型性能的提升带来可能。

参 考 文 献

[1] AGOR J K, OZALTIN O Y. Models for predicting the evolution of influenza to inform vaccine strain selection[J]. Hum Vaccin Immunother, 2018, 14(3): 678-683.

[2] YIN R, ZHOU X, IVAN F X, et al. Identification of Potential Critical Virulent Sites Based on Hemagglutinin of Influenza a Virus in Past Pandemic Strains[C]// Proceedings of the 6th International Conference on Bioinformatics and Biomedical Science. Singapore, Association for Computing Machiner, 2017: 30-36.

[3] NEHER R A, BEDFORD T. Nextflu: real-time tracking of seasonal influenza virus evolution in humans[J]. Bioinformatics, 2015, 31(21): 3546-3548.

[4] SAUTTO G A, KIRCHENBAUM G A, ROSS T M. Towards a universal influenza vaccine: different approaches for one goal [J]. Virology Journal, 2018, 15(1): 17.

[5] YIN R, LUUSUA E, DABROWSKI J, et al. Tempel: time-series mutation prediction of influenza A viruses via attention-based recurrent neural networks[J]. Bioinformatics, 2020, 36(9): 2697-2704.

[6] DE JONG J C, PALACHE A M. Haemagglutination-inhibiting antibody to influenza virus[J]. Developments in Biologicals, 2003, 115: 63-73.

[7] SMITH D J, LAPEDES A S, DE JONG J C, et al. Mapping the antigenic andgenetic evolution of influenza virus[J]. Science 2004, 305(5682): 371-376.

[8] 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.

[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] ZHOU X, YIN R, KWOH C K, et al. A context-free encoding scheme of protein sequences for predicting antigenicity of diverse influenza A viruses[C]// Proceedings of the 29th International Conference on Genome Informatics(GIW 2018): genomics. BMC Genomics, 2018: 936.

[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: 42051

[12] YIN R, TRAN V H, ZHOU X, et al. Predicting antigenic variants of H1N1 influenza virus based on epidemics and pandemics using a stacking model[J/OL]. https://doi. org/10. 1371/journal. pone. 0207777, 2018.

[13] LECUN Y, BENGIO Y, HINTON G. Deep learning[J]. Nature, 2015, 521(7553): 436-444.

[14] HOCHREITER S, SCHMIDHUBER J. Long short-term memory [J]. Neural Computation, 1997, 9(8): 1735-1780.

[15] CHO K, VAN MERRIENBOER B, BAHDANAU D. On the properties of neural machine translation: Encoder-decoder approaches [J]. arXiv 2014: 14091259.

[16] 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 Trans Comput Biol Bioinform, 2021, 9: 1-1.

[17] ASWANI A V, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. arXiv: 1706. 03762, 2017.

[18] ASGARI E, MOFRAD M R K. ProtVec: A Continuous Distributed Representation of Biological Sequences[J]. PloS One, 2015, 10: 11.

[19] MARCAIS G, KINGSFORD C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers[J]. Bioinformatics, 2011, 27(6): 764-770.

[20] KINGMA D P, BA J. Adam: A method for stochastic optimization [J]. arXiv: 1412. 6980, 2015.

[21] SRIVASTAVA N, HINTON G, KRIZHEVSKY A. Dropout: a simple way to prevent neural networks from overfitting[J]. The Journal of Machine Learning Research, 2014, 15(1): 1929-1958.

[22] RADZICKA A, WOLFENDEN R. Comparing the polarities of the amino acids: Side-chain distribution coefficients between the vapor phase, cyclohexane, 1-octanol, and neutral aqueous sol-ution[J]. Biochemistry, 1988, 27(5): 1664-1670.

[23] MEILER J, MLLER M, ZEIDLER A, et al. Generation and evaluation of dimension-reduced amino acid parameter representations by artificial neural networks[J]. Mol. Model. Annu., 2001, 7(9): 360-369.

[24] ATCHLEY W R, ZHAO J, FERNANDES A D, et al. Solving the protein sequence metric problem[J]. Proc. Nat. Acad. Sci. United States America, 2005, 102(18): 395-6400.

[25] DAYHOFF M O. A model of evolutionary change in proteins [J]. Atlas Protein Sequence Structure, 1978, 5: 89-99.

[26] HENIKOFF S, HENIKOFF J G. Amino acid substitution matrices from protein blocks[J]. Proceedings of the National Academy of Sciences, 1992, 89(22): 10915-10919.

[27] ALTSCHUL S F, KOONIN E V. Iterated profile searches with PSI-BLAST—a tool for discovery in protein databases[J]. Trends in biochemical sciences, 1998, 23(11): 444-447.

[28] MIYAZAWA S, JERNIGAN R L. Self-consistent estimation of inter-residue protein contact energies based on an equilibrium mixture approximation of residues[J]. Proteins: Structure Function Bioinf, 1999, 34(1): 49-68.

[29] MICHELETTI C, SENO F, BANAVAR J R, et al. Learning effective amino acid interactions through iterative stochastic techniques[J]. Proteins: Structure Function Bioinf, 2001, 42(3): 422-431.

[30] LIN K, MAY A C W, TAYLOR W R. Amino acidencoding schemes from protein structure alignments: Multi-dimensional vectors to describe residue types[J]. Theoretical Biol, 2002, 216(3): 361-365.

LI Chuan, born in 1998, 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.