计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.241000145

PPIS-MFH:集成 ViT 的多特征混合网络预测蛋白质相互作用位点

胡昭龙　胡春玲　胡瑞捷　郭龙菊
合肥大学人工智能与大数据学院　合肥 230601
(sebant@163.com)

摘　要　通过深入研究蛋白质-蛋白质相互作用位点(PPIS),能够揭示生命在分子层面运作的深层原理。然而现有方法鉴定PPIS复杂且耗时,需要更精确的模型进行PPIS预测。尽管基于注意力机制和卷积神经网络(CNN)的深度学习方法在PPIS预测方面取得了进展,但在氨基酸特性表征上仍存在局限。为了有效捕捉蛋白质序列中远距离的依赖关系,并准确地表征氨基酸的特性,提出了一种用于预测蛋白质-蛋白质相互作用位点的多特征混合网络(Multi-feature hybrid networks)——PPIS-MFH,通过结合全局序列特征与局部序列特征对PPIS进行预测。对于局部序列特征,PPIS-MFH模型融合了Vision Transformer(ViT)模块,该模块能够捕获蛋白质序列中的远距离依赖性,并提取局部特征。对于全局序列特征,模型PPIS-MFH通过由文本卷积神经网络(TextCNN)并引入注意力机制的文本循环神经网络(TextRNN-Attention)构成的特征交叉网络,利用双向门控循环单元网络来识别蛋白质序列中氨基酸间的内在联系。在4个数据集上对PPIS-MFH模型进行了评估,将其与8种同类方法进行了比较。实验结果显示在大多数指标上,所提方法优于其他的同类方法。

关键词:蛋白质-蛋白质相互作用位点;注意力机制;文本卷积神经网络;双向门控循环单元网络;特征交叉网络
中图分类号　TP391

PPIS-MFH:Predicting Protein-Protein Interaction Sites Based on Multi-feature Hybrid Network Integrating ViT

HU Zhaolong,HU Chunling,HU Ruijie and GUO Longju
School of Artificial Intelligence and Big Data,Hefei University,Hefei 230601,China

Abstract　The deeper principles of molecular life can be revealed through an in-depth study of protein-protein interaction sites (PPIS). However,existing methods for identifying PPIS are complex and time-consuming,and more accurate models are needed for PPIS prediction. Although deep learning techniques based on attention mechanisms and convolutional neural networks(CNNs) have made progress in PPIS prediction,they still face limitations in capturing amino acid features. To effectively capture long-range dependencies in protein sequences and accurately characterize amino acid properties,this paper proposes a multi-feature hybrid network(MFH),PPIS-MFH,for predicting protein-protein interaction sites. Protein-protein interaction sites are predicted by combining both global and local sequence features. For local sequence features,the PPIS-MFH model incorporates a Vision Transformer(ViT) module,which captures long-range dependencies and extracts local features from protein sequences. For global sequence features,the model employs a bidirectional gated recurrent neural network to discern intrinsic connections between amino acids in protein sequences. This is achieved through a feature crossover network that combines a text convolutional neural network(TextCNN) with an attention mechanism,specifically a text recurrent neural network(TextRNN-Attention). In this study,the PPIS-MFH model was evaluated on four datasets and compared with eight similar methods. The experimental results show that,on most metrics,the proposed method outperforms other similar methods.

Keywords　Protein-protein interaction site,Attention mechanism,Text convolutional neural network,Bidirectional gated recurrent neural network,Feature crosses network

1　引言

在生物信息学与计算生物学领域,蛋白质-蛋白质相互作用(PPI)已经成为一个重要的研究方向。PPI在许多细胞功能中扮演着至关重要的角色,包括信号传导、分子运输和新陈代谢等过程。深入探索蛋白质间的相互作用对于理解生命活动的结构和功能至关重要[1-2],在这些相互作用中,PPI位点作为特定氨基酸残基的集合,通常只占据蛋白质表面的小部分区域,却显著地影响着蛋白质间相互作用的特异性与结合强度[3-4]。准确预测PPI位点不仅能够帮助人们理解PPI的结合模式与作用原理,还有助力于设计针对性的小分子抑制剂或促进剂。因此,对PPI位点的探究对于阐明生命的分子基础、识别新的生物标志物以及药物靶标具有极其重要的意义。

PPI位点预测方法可分为两大类:生物学实验方法和计

基金项目:国家自然科学基金面上项目:面向动态知识图谱的局部图表示学习研究(62306100)
This work was supported by the National Natural Science Foundation of China:Research on Local Graph Representation Learning for Dynamic Knowledge Graph(62306100).
通信作者:胡春玲(huchunling@hfuu.edu.cn)

算预测方法。其中生物学实验方法包括质谱分析(MS)、X 射线晶体学以及核磁共振(NMR)等[5-8],这些方法能够精确地识别出 PPI 位点,然而其往往伴随着较高的经济成本与时间成本。因此,开发一种既准确又高效的计算预测方法对于提高 PPI 位点的鉴定效率十分重要。目前 PPI 位点的计算预测方法主要可以分为两大类:基于序列的方法与基于结构的方法[9-11]。其中基于序列的方法仅依赖于蛋白质的一级结构信息,即只通过氨基酸序列进行 PPI 位点的预测,而基于结构的方法除去序列信息外还需借助蛋白质的三维结构信息。

由于序列信息的获取相对容易,且与结构信息相比更为稳定,因此基于序列信息的 PPI 位点预测成为了一个备受关注的研究热点。Pitre 等[12]基于已知的蛋白质相互作用数据,通过分析蛋白质对之间的短多肽序列来识别相互作用位点;Ofran 等[13]从蛋白质序列中提取潜在的相互作用模式,同时根据接触残基的组成来识别相互作用位点。随后提出的模型广泛使用了贝叶斯分类器[14]、学习向量量化[15]、逻辑回归、人工神经网络[16]、浅层神经网络[17]等方法。Koike 等将每个氨基酸残基的邻域序列转换成高维向量,并使用支持向量机(SVM)作为分类器[18],Wang 等则采用了随机森林算法来进行位点预测[19]。Hou 等将蛋白质序列通过滑动窗口的方式划分为固定长度的窗口以提取蛋白质序列的局部特征,并基于随机森林(RF)对每个窗口进行分类[9]。

近年来,深度学习技术也为 PPI 位点预测带来了新的突破。Zeng 等采用了文本卷积神经网络(TextCNN)来提取蛋白质序列的全局特征,并将其与通过滑动窗口捕获的局部特征结合起来,以此进行 PPI 位点的预测,表现出稳健的性能[20]。TextCNN 通过运用多个不同尺寸的卷积核,有效地提取到了蛋白质序列中的局部信息,同时捕捉到了不同尺度下氨基酸的相互作用。此外,TextCNN 通过最大池化层操作,筛选出最关键的特征,从而在减少特征维度和降低计算复杂度的同时保持了信息的完整性。Zhang 等则提出了一种基于简化的长短期记忆网络(LSTM)方法来进行 PPI 位点的预测,这种简化版的 LSTM 是针对传统循环神经网络(RNN)的一种改进,通过减少参数量以及降低计算复杂度来提高模型的效率与泛化能力[21]。Lu 等设计了一种融合注意力机制的卷积神经网络(CNN)方法,用于识别 PPI 位点,该方法通过对每个特征图执行注意力机制,根据残基对 PPI 位点的不同影响程度赋予相应的权重,以此强调关键特征并抑制噪声干扰,进一步提升了预测的准确性[22]。Cong 等提出了一种基于卷积神经网络的方法,结合了混合特征、自注意力机制和模型集成,用于预测 PPI 位点[23]。虽然这些方法在 PPIS 预测方面取得了进展,但在氨基酸特性表征上仍存在局限性。同时,基于 CNN 的方法因其固定的感受野大小,每个卷积核只能感知固定大小的局部区域,随着层数的加深,感受野逐渐扩展,但这种方法对于捕捉远距离依赖关系往往存在局限。而 ViT 通过自注意力机制,允许每个氨基酸残基直接与序列中的其他所有残基交互,计算其之间的依赖关系。

此外,蛋白质序列的局部特征在 PPI 位点预测中的重要性已经得到广泛认可。Wang 等使用滑动窗口提取邻近残基的序列特征和残基进化率,进而通过集成学习方法来提取局部特征[24]。同时,蛋白质序列的全局特征也已被证明了在 PPI 位点预测方面的有效性,Zeng 等提出了一种端到端的深度学习框架,通过结合局部和全局的蛋白质序列特征来预测 PPI 位点,并通过在两个独立的验证数据集上对比其他方法的性能,证明了全局特征对于提高 PPI 位点预测的准确率是有效的[20]。

因此,本文构建了一个集成 TextCNN,TextRNN-Attention,GRU 以及 ViT 等模块的深度学习模型 PPIS-MFH 来提高 PPI 位点预测的准确性与鲁棒性。考虑到模型的通用性,本文采用了最常用的特征作为模型的输入,包括原始蛋白质序列、PSSM 与蛋白质二级结构。然后,为了有效提取蛋白质序列的局部特征以鉴别 PPI 位点,使用 ViT 来处理蛋白质序列中的远距离依赖关系。对于蛋白质序列的全局特征,则使用 TextRNN-Attention 与 TextCNN 构成的特征交叉网络捕捉不同氨基酸之间的潜在关系。本文的主要工作如下:

(1)提出将 ViT 网络应用于提取蛋白质序列的局部特征以进行 PPI 位点预测。

(2)提出通过 TextRNN-Attention 与 TextCNN 构成的特征交叉网络来提取 PPI 位点的全局特征,将 TextRNN-attention 应用于 PPI 位点预测中。

(3)在 4 个数据集上对 MFH-PPIS 进行测试,实验结果表明,本文提出的方法优于其他同类方法。

2 数据集与材料

从计算的角度来分析,蛋白质相互作用位点预测致力于解决如下挑战:针对一个特定的蛋白质,其由氨基酸序列 X 构成,本研究的任务是确定一个最优的映射函数 $F(X)$,该函数能够将序列 X 映射到一个由 0 和 1 构成的标签序列 Y。在这个序列 Y 中,1 用来表示结合残基的位置,而 0 则代表非结合残基的位置。通过对映射函数 F 进行优化,并利用深度学习方法对其进行训练,可以提升 F 的性能并使其具备对训练集以外的蛋白质序列预测相应 Y 序列的能力。

2.1 数据集

与以往的研究类似,本文使用了 4 个数据集(Dset_186,Dset_72,PDBset_164,Dset_331)来评估 PPIS-MFH 方法的性能。其中 Dset_186,Dset_72,PDBset_164 均获取自 PDB 数据库,其中选取的蛋白质序列同源性≤25%,以确保序列间的多样性,同时所有蛋白质的分辨率均≤3.0Å,这意味着序列数据具有一定的准确性与清晰度。考虑到公开数据集 Dset_186,Dset_72 和 PDBset_164 中的蛋白序列数据互不相同且源自不同的研究团队,本文采用文献[20]的方式将这 3 个数据集合并为一个融合数据集,以保持训练集与测试集的分布一致性。此外,将公开数据集 Dest_448 通过 DSSP 工具[25-26]与 PSI-BLAST 搜索工具[27]生成蛋白质原始序列对应的 PSSM 和 DSSP,去除未定义序列后得到数据集 Dset_331,共包含 331 个有效蛋白质数据。依据文献[20]所采用的标准,当氨基酸在蛋白质结合过程中的绝对溶剂可及性变化小于 1Å2 时,该氨基酸被认为是相互作用位点。反之,若其绝对溶剂可及性变化大于或等于 1Å2,则被认为是非相互作用位点。统计结果如表 1 所列。

表 1 各数据集相互作用位点数量统计

Table 1 Statistics on the number of binding sites in each dataset

数据集	相互作用位点	非相互作用位点
Dset_186	1923	16217
Dset_72	5517	30702
PDBset_164	6096	27585
Dset_331	11255	72420

本文对融合数据集 Dset_186_72_PDB164 以及 Dest_331 中的序列长度进行了统计，并在表 2 中展示了其中蛋白质序列长度的分布情况。在本研究中，将数据集按照 1∶6 的比例分割成测试集和训练集进行实验。

表 2　各数据集蛋白质序列长度统计

Table 2　Protein sequence length statistics for each dataset

数据集	蛋白质序列长度							
	1～100	100～200	200～300	300～400	400～500	500～600	600～700	700＋
Dset_186_72_PDBset_164	85	176	68	56	23	7	4	3
Dset_331	25	99	100	68	28	6	5	0

2.2　输入特征

本文选用了 3 组基于蛋白质序列的特征进行训练，包含原始蛋白质序列、二级结构以及位置特异性得分矩阵(PSSM)。其中原始蛋白质序列和 PSSM 的特征维度都为 20 维，蛋白质的二级结构特征维度为 9 维。综合这 3 组特征，本文构建了一个 49 维的特征向量作为模型的输入。下文详细说明这 3 个特征组的具体细节。

1)原始蛋白质序列

原始蛋白质序列能够精确地表示各个氨基酸在蛋白质序列中的具体位置。鉴于大多数蛋白质由 20 种不同的氨基酸构成，本文采用了 20 维的 one-hot 编码方式来表示蛋白质序列中的氨基酸类型。

2)二级结构

蛋白质的二级结构定义了多肽链主链的空间排列，表征着局部肽段在三维空间中的折叠模式。本文应用 DSSP 工具提取蛋白质序列的二级结构信息。在 DSSP 的分类标准中，共有 8 种主要的二级结构类型可被识别，分别是 310-螺旋(G)、α-螺旋(H)、π-螺旋(I)、β-折叠(E)、β-桥(B)、β-转角(T)、弯曲(S)以及不规则或无定形环(L)。因此，本文使用一个 8 维的 one-hot 编码向量来表示这 8 种状态。此外，考虑到 DSSP 的结果中有些氨基酸未被明确指定二级结构，本文增加一个维度，即采用 9 维的 one-hot 向量来编码包括未指定状态在内的二级结构信息。

3)位置特异性得分矩阵

本文使用 PSI-BLAST 搜索工具，在 UniProtKB 数据库中进行 3 次迭代搜索，并定义截断阈值 $e=0.001$，以此来构建位置特异性得分矩阵。该矩阵为每个氨基酸的位置映射出 20 种氨基酸的出现概率，形成一个 20 维的概率向量，以此表示序列中的进化信息。

3　PPIS-MFH 模型

图 1 展示了 PPIS-MFH 模型结构，该模型接受两种类型的输入特征：局部序列特征和全局序列特征。模型分别对这两种特征进行独立的特征提取，并将最终得到的局部和全局特征拼接后，输入到分类模块进行分类。

在处理局部序列特征时，本文采用了基于滑动窗口的方法，将原始蛋白质序列切割为多个局部子序列。然后，利用 ViT 网络中的多头自注意力机制及其对多尺度特征的处理能力捕获蛋白质序列中的远距离依赖关系，得到最终的局部序列特征。其中，滑动窗口通过选择合适的窗口大小和起始位置，可以精确地从蛋白质序列中提取出特定长度的子序列，以进行进一步的特征提取。

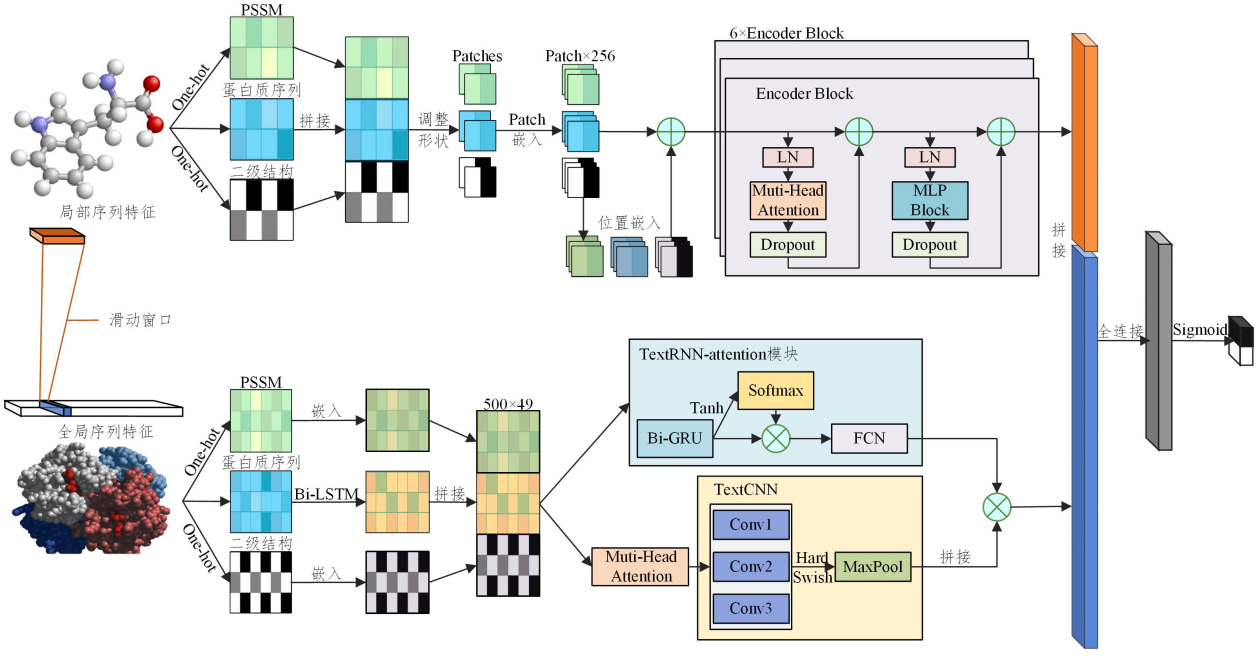


图 1　PPIS-MFH 模型概述

Fig. 1　Overview of the PPIS-MFH model

在处理全局序列特征时，本文通过 Bi-LSTM 模块进一步提取蛋白质序列的进化信息，将提取到的信息经过由 Text-RNN-Attention 和 TextCNN 构成的特征交叉网络输出最终的全局序列特征。

最后，本文将两个模块输出的全局序列特征与局部序列特征通过拼接操作进行融合，并将拼接后的序列输出到分类模块。分类模块由两个全连接层和一个 Sigmoid 函数组成，其输出是一个单元素的二进制值，用于表示序列的指定位置是否为蛋白质结合位点。

3.1 ViT 网络

为了更好地提取蛋白质的局部序列特征，本文引入了 ViT 网络。在 ViT 中，蛋白质序列首先被切分成固定大小的块，每个块作为独立的输入特征，通过线性变换嵌入到高维空间中。然后，通过自注意力机制计算每个块的 Query，Key 和 Value 向量，并通过点积计算 Query 与 Key 之间的相似度，得到注意力权重，表示每个块对其他块的关注程度。基于这些权重，模型对所有 Value 向量进行加权求和，从而更新每个块的表示。为了从不同子空间提取更多信息，ViT 采用多头自注意力机制，将 Query，Key 和 Value 分成多个头部，使得模型能够在多个层次上捕捉氨基酸序列中的不同依赖关系，从而更好地表达序列中的远距离依赖性。

其中，ViT 网络的输入序列是由滑动窗口截取的连续氨基酸片段、位置特异性得分矩阵以及蛋白质的二级结构信息拼接而成。为了将拼接后的蛋白质特征转化为 ViT 模型能够处理的格式，本文将输入数据调整为一个具有 7 个通道、分辨率为 7×7 的类网格表示。具体来说，本文在原始的序列数据基础上增加一个通道维度，以便将其输入到 ViT 模块中。ViT 模块包含 Patch Embedding 层、Positional Embedding 层、Transformer Encoder 模块和 MLP 层 4 个部分。

首先，Patch Embedding 层将输入数据分割成 49 个固定大小的块，每个块通过线性映射后获得了 256 维的通道表示，此时序列的维度是 49×256。此外，拼接一个分类字符 cls_token(class token)，此时序列维度变成 50×256，然后通过 Positional Embedding 层为序列中的每个元素添加对应的位置嵌入信息。加入位置嵌入信息之后，序列的维度保持不变，其定义如下：

$$Y_0=[\boldsymbol{X}_{\text{cls}};\boldsymbol{X}\boldsymbol{W}_{\text{patch}}]+\boldsymbol{X}\boldsymbol{W}_{\text{pos}} \tag{1}$$

其中，$\boldsymbol{X}\in\mathbb{R}^{D_b\times D_{pos}\times D_t}$ 表示输入序列；X_{cls} 表示分类字符 cls_token；$\boldsymbol{W}_{\text{pos}}$ 和 $\boldsymbol{W}_{\text{patch}}$ 分别表示 Positional Embedding 与 Patch Embedding 的线性变换矩阵；$Y_0\in\mathbb{R}^{D_b\times(D_{\text{pos}}+1)\times D_t}$ 是输入数据与位置嵌入信息相加后的输出，其中 D_b 表示批次大小，D_{pos} 表示序列长度，D_t 表示每个块的特征维度。

加入了位置嵌入的输入数据随后被传递到 Transformer Encoder 模块中，该模块由 6 个连续的 Encoder Block 构成，Encoder Block 中的 MLP Block 包含两个全连接层和两个 Dropout 层。在每个块的处理过程中，数据的维度保持与输入时一致。在经过 6 个 Encoder Block 作用后，会将分类字符 cls_token 对应的输出作为 Transformer Encoder 模块的最终输出，并通过 MLP 层与全局序列特征拼接后输入到分类器中，其中 Encoder Block 的定义如下：

$$\boldsymbol{Z}=LN(Y_{L-1}) \tag{2}$$

$$\boldsymbol{Q}_n=\boldsymbol{Z}\boldsymbol{W}_n^Q,\boldsymbol{K}_n=\boldsymbol{Z}\boldsymbol{W}_n^K,\boldsymbol{V}_n=\boldsymbol{Z}\boldsymbol{W}_n^V \tag{3}$$

$$A_n=\text{softmax}\left(\frac{\boldsymbol{Q}_n\boldsymbol{K}_n^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V}_n,n=1,\cdots,k \tag{4}$$

$$\boldsymbol{Y}_L'=\boldsymbol{Y}_{L-1}+Concat(A_1,A_2,\cdots,A_k)\boldsymbol{W}^O \tag{5}$$

$$\boldsymbol{Y}_L=\boldsymbol{Y}_L'+W_2\cdot GELU(W_1\cdot LN(\boldsymbol{Y}_L')+b_1)+b_2 \tag{6}$$

其中，LN 是归一化层；$\boldsymbol{Q}_n,\boldsymbol{K}_n,\boldsymbol{V}_n$ 分别表示查询矩阵(Query)、键矩阵(Key)和值矩阵(Value)；A_n 表示每个注意力头的注意力；$\boldsymbol{W}^O$ 是多头注意力的输出权重矩阵；$\boldsymbol{Y}_L'\in\mathbb{R}^{D_b\times(D_{\text{pos}}+1)\times D_t}$ 与 $\boldsymbol{Y}_L\in\mathbb{R}^{D_b\times(D_{\text{pos}}+1)\times D_t}$ 分别为对应 Encoder Block 中多头自注意力层与前馈神经网络的第 L 层的输出，其中 D_b 表示批次大小，$D_{\text{pos}}+1$ 表示序列长度，D_t 表示每个块的特征维度。

3.2 多特征交叉网络

多特征交互网络是由 TextRNN-Attention 模块与多头注意力文本卷积模块通过点乘操作进行特征融合而成。其中 TextRNN-Attention 通过注意力机制增强重要特征的表示，而 TextCNN 则通过卷积操作进一步提取更为细致的局部特征，这两个网络的协同工作有效提升了全局特征的表达能力。

3.2.1 TextRNN-Attention 模块

在 PPI 位点的预测任务中，全局序列特征与局部上下文特征同等重要。拼接后的 3 组特征，可被视作一段文本序列，TextRNN 可以通过其循环结构传递信息以提取序列中的远距离依赖关系。此外，TextRNN 还考虑到了氨基酸在蛋白质序列中的出现顺序，这对于全局特征的提取十分重要。

本文使用 TextRNN-Attention 模块来提取 3 组特征之间的上下文关系。TextRNN-Attention 在传统文本循环神经网络的基础上引入了注意力机制，该机制能区分特征序列中每个特征向量对于分类任务的贡献程度，在没有引入注意力机制的传统 TextRNN 模型中，通常假设每个特征向量对于分类任务的贡献是相同的。然而，在 PPI 位点预测中，并非所有特征都对结合位点的分类结果具有相同的重要性。通过引入注意力机制，可以根据其重要性给予不同的关注。

值得注意的是，门控循环单元(GRU)在小数据集场景下具有优势，且其结构简单以及参数数量相对较少，这有助于减少模型过拟合的风险。因此，本文采用 GRU 替换原本的 LSTM 网络。在 GRU 层中，当前节点的输入数据以及前一节点输出的信息会被用来决定两个关键的门控状态：复位门(reset gate)和更新门(update gate)。这些门控状态随后被用以筛选和整合历史信息与当前的输入信息。通过这种方式，GRU 单元能够有选择性地更新其内部状态，保留重要的信息同时摒弃无关的细节。其具体定义如下：

$$\boldsymbol{R}_t=\sigma(\boldsymbol{W}_r\cdot[\boldsymbol{H}_{t-1},\boldsymbol{X}_t]) \tag{7}$$

$$\boldsymbol{Z}_t=\sigma(\boldsymbol{W}_z\cdot[\boldsymbol{H}_{t-1},\boldsymbol{X}_t]) \tag{8}$$

$$\widetilde{\boldsymbol{H}}_t=\tanh(\boldsymbol{W}_h\cdot[\boldsymbol{R}_t*\boldsymbol{H}_{t-1},\boldsymbol{X}_t]) \tag{9}$$

$$\boldsymbol{H}_t=(1-\boldsymbol{Z}_t)*\boldsymbol{H}_{t-1}+\boldsymbol{Z}_t*\widetilde{\boldsymbol{H}}_t \tag{10}$$

其中，$\sigma(\cdot)$表示 sigmoid 函数，用于将预测值映射到 0 与 1 之间；$\boldsymbol{W}$ 表示权重矩阵；$\boldsymbol{R}_t$ 是在时间步 t 的重置门向量；$\boldsymbol{Z}_t$ 是在时间步 t 的更新门向量；$\boldsymbol{H}_{t-1}$ 是时间步 $t-1$ 的隐藏状态；$\boldsymbol{X}_t$ 为时间步 t 的输入向量；$\widetilde{\boldsymbol{H}}_t$ 是时间步 t 的候选隐藏状态；$\boldsymbol{H}_t$ 表示时间步 t 的最终隐藏状态。

3.2.2 多头注意力文本卷积模块

为了有效处理 3 组特征的差异性，本文使用多头注意力(Multi-Head Attention)机制来处理这些拼接的特征，通过注意力加权的方式挖掘与融合不同特征之间的关联性。多头注

意力中每个头可以关注输入特征的不同部分，从而提供更丰富的信息表示。

考虑到文本卷积网络在挖掘文本序列中相邻元素间联系方面的潜力，本文将融合后的 3 组特征视作连续文本序列。通过应用以 Hard Swish 作为激活函数的文本卷积网络，模型能够有效提取不同长度子序列的特征，并捕捉到不同尺寸感受野中氨基酸之间的关系。

然后，通过最大池化层筛选出每个通道内关键的特征，并对输出特征进行降维处理，以实现更有效的信息压缩与噪声削减。

随后，将最大池化层输出的特征向量进行拼接，形成一个整合 3 个不同尺寸卷积核输出的复合向量，其集成了序列中各部分的关键信息。

$$Y_G^i = MP(\sigma(Conv_i(\boldsymbol{X}_g))), i=1,2,3 \tag{11}$$

$$\boldsymbol{Y}_G = Concat(Y_G^1, Y_G^2, Y_G^3) \tag{12}$$

其中，MP 表示最大池化层；$Conv_i$ 表示第 i 个卷积层；$\boldsymbol{Y}_G \in \mathbb{R}^{D_b \times D_g}$ 是输入数据经过卷积层、最大池化层后的输出，其中 D_b 是批次大小，D_g 是经过池化操作后的特征维度；$\sigma(\cdot)$ 表示 Hard Swish 激活函数；$\boldsymbol{X}_g \in \mathbb{R}^{D_b \times 1 \times D_{\text{len}} \times D_{\text{pos}}}$ 为拼接后的全局序列特征输入数据，其中 D_b 是批次大小，D_{len} 是序列长度，D_{pos} 则是位置嵌入维度。

3.3 分类部分

分类部分包括两个全连接层和一个 sigmoid 函数。本文通过拼接操作将全局特征与局部特征融合，作为第一个全连接层的输入，拼接后的特征经过两层全连接层映射后，最终通过 sigmoid 函数将输出值压缩到 0 到 1 之间进行二元分类，其结果代表该位置上的蛋白质残基是否为结合位点。这种结构不仅有效地整合了不同尺度的特征信息，还通过非线性激活函数增强了模型的表达能力与分类性能。

4 实验结果分析

4.1 评价指标

在 PPI 位点预测任务中，本文将实际发生作用的位点定义为正样本，将没有相互作用的位点定义为负样本。为了全面比较本文提出的模型与其他现有方法在此领域的性能表现，本文引用了 6 个评估指标：准确率（Accuracy，ACC）、精度（Precision）、召回率（Recall）、F 值（F-measure）精度-召回率曲线下面积（Area Under the Precision-Recall Curve，AUC PR）以及马修斯相关系数（Matthews Correlation Coefficient，MCC）。定义如下：

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{13}$$

$$precision = \frac{TP}{TP+FP} \tag{14}$$

$$recall = \frac{TP}{TP+FN} \tag{15}$$

$$F\text{-}measure = \frac{2 \cdot precision * recall}{precision + recall} \tag{16}$$

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{17}$$

其中，TP 表示正确预测为相互作用位点的数目；FN 表示错误预测为非相互作用位点的数目；FP 表示错误预测为相互作用位点的数目；TN 表示正确预测为非相互作用位点的数目。在处理不平衡数据集时，F 值、马修斯相关系数以及精度-召回率曲线下面积这 3 个评价标准更为全面。F 值综合考虑了精度与召回率的平衡，马修斯相关系数适用于度量二元分类问题中类别分布不均衡的情形，而精度-召回率曲线下面积则全面反映了模型在各召回率水平下的精度表现。这些指标共同构成了一个有效的评价体系，使其能够更准确地衡量模型在识别 PPI 位点上的性能。

4.2 实验环境与参数

本文构建了基于 PyTorch 2.2.1 框架的深度学习模型，并结合了 CUDA 11.8 加速库以及 Python 3.10 进行实现。在模型训练过程中，本文选用 Adam 作为主要的优化算法。本研究的实验硬件设备为 GeForce RTX 4070 Super。

为了有效捕获蛋白质序列中的局部上下文信息与全局序列特征，本文将滑动窗口的长度设定为 7，同时限定蛋白质序列的处理长度为 500。同时，在深度学习结构中，本文设置了批量大小（Batch Size）为 128，以确保足够的样本量进行有效学习，同时维持计算资源的合理性。多头注意力文本卷积模块中，为了提取蛋白质序列的全局特征，采用了多尺度卷积神经网络层，其中包含核大小分别为 13，15 和 17 的卷积核。在实现 ViT 网络时，将 Patch Embedding 层的线性映射维度设定为 256，以便提取更丰富的序列信息。在 TextRNN-Attention 模块中，设置 Dropout 率为 0.5 以预防过拟合现象，从而增强模型的泛化能力。至于分类部分，本文将两个全连接层神经元个数分别设置为 1024 和 256。

4.3 实验结果

4.3.1 与其他同类方法的比较

为了衡量 PPIS-MFH 在蛋白质结合位点预测方面的准确度，本研究进行了性能对比分析，将 PPIS-MFH 模型与其他 8 种同类预测方法进行比较，包括 PSIVER，SPPIDER，SPRINGS，ISIS，RF_PPI，DeepPPISP，Attention-CNN 以及 StackingPPINet。

在这些方法中，PSIVER 运用位置特异性得分矩阵与可及性预测等序列衍生特征，借助朴素贝叶斯分类器来实现 PPI 位点的识别；SPPIDER 利用先进的机器学习技术整合了蛋白质表位以及额外的序列和结构数据，以加强对 PPI 位点的识别能力；SPRINGS 结合了进化数据、亲水性度量与溶剂可及性分析，通过一个浅层神经网络来进行 PPI 位点预测；ISIS 开发了一种结合结构特征和进化信息的浅层神经网络方法，用于预测蛋白质结合位点；RF_PPI 利用随机森林算法和多样的特征集合来识别 PPI 位点；DeepPPISP 通过结合局部特征和全局特征的端到端深度学习框架来预测 PPI 位点；Attention-CNN 通过结合注意力机制和卷积神经网络来预测 PPI 位点；StackingPPINet 结合了混合特征、自注意力机制与集成模型来预测 PPI 位点。

表 3 对比了同类方法中是否应用局部序列特征或全局序列特征，其中“Y”表示应用，“N”表示未应用。在提及的方法中，仅有 PPIS-MFH 和 DeepPPISP 这两种方法融合了局部的细节信息与整个蛋白质序列的全局信息进行 PPI 位点的预测分析。

表 3 同类方法中应用局部特征或全局特征的统计

Table 3 Statistics for applying local features or global features in homogeneous methods

方法	局部特征	全局特征
PSIVER	Y	N
SPPIDER	Y	N
SPRINGS	Y	N
ISIS	Y	N
RF_PPI	Y	N
DeepPPISP	Y	Y
Attention-CNN	Y	N
StackingPPINet	Y	N
PPIS-MFH	Y	Y

表 4 对比了 PPIS-MFH 与其他 8 种同类计算模型在测试集上的表现。综合比较各项指标，可以看出 PPIS-MFH 在多数性能评估参数上均超越了对比模型。尽管在准确率方面，PPIS-MFH 略低于 StackingPPINet，但在其他关键性能指标上，它展现出了明显的优势。具体而言，PPIS-MFH 在精度、F 值、recall 值、马修斯相关系数以及 AUC PR 方面取得了 0.317，0.420，0.625，0.238 和 0.362 的成绩，均优于对比模型。

表 4 在 Dset_186_72_PDB164 数据集上 PPIS-MFH 与其他同类方法的性能对比结果

Table 4 Performance comparison results of PPIS-MFH with other similar methods on Dset_186_72_PDB164 dataset

方法	*Acc*	*Precision*	*Recall*	*F-measure*	*MCC*	*AUC PR*
PSIVER	0.653	0.253	0.468	0.328	0.138	0.250
SPPIDER	0.622	0.209	0.459	0.287	0.089	0.230
SPRINGS	0.631	0.248	0.598	0.350	0.181	0.280
ISIS	0.694	0.211	0.362	0.267	0.097	0.240
RF_PPI	0.598	0.173	0.512	0.258	0.118	0.210
DeepPPISP	0.655	0.303	0.577	0.397	0.206	0.320
Attention-CNN	0.657	0.313	0.611	0.414	0.229	0.359
StackingPPINet	**0.705**	0.309	0.612	0.414	0.222	0.339
PPIS-MFH	0.658	**0.317**	**0.625**	**0.420**	**0.238**	**0.362**

需要特别指出的是，由于 PPI 位点预测属于不平衡分类问题，在这类问题中，F-measure 和 MCC 是更为准确的性能衡量标准。PPIS-MFH 在这两项指标上均达到了最佳水平，这一结果充分说明了其相较于其他方法的优越性。因此，可以得出结论：PPIS-MFH 在整体性能上占据了领先地位，特别是在处理不平衡数据方面展现了显著的效果。

为了进行更深入的比较，本文在 Dset_331 数据集上对 PPIS-MFH 模型与 DeepPPISP 模型的性能进行了对比，如图 2 所示。

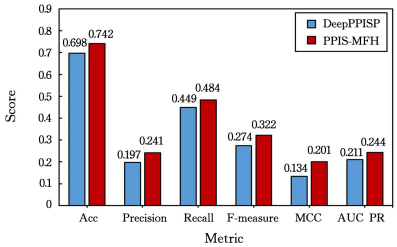


图 2 PPIS-MFH 和 DeepPPISP 在 Dset_331 数据集上的性能对比

Fig. 2 Performance comparison between PPIS-MFH and DeepPPISP on Dset_331 dataset

结果表明，PPIS-MFH 在各项评估指标上均实现了显著提升，具体包括：准确率增长 6.3%，精度提升 22.3%，召回率增长 7.8%，F 值增长 17.5%，马修斯相关系数提升 50%以及 AUC PR 增长 15.6%。这些指标的提升验证了本文提出的 PPIS-MFH 方法在 PPI 位点预测上的有效性。

在对比不同数据集的表现时，可以看到尽管 PPIS-MFH 模型在 Dset_331 数据集上有所提升，但其 F 值相对较低。我们认为这一现象可能与 Dset_331 数据集中相互作用位点的占比较小有关。此外，考虑到 Dset_331 数据集中的蛋白质序列长度大多数集中在 200～300 区间，而 Dset_186_72_PDBset_164 数据集的蛋白质序列主要分布在 1～200 区间，较长的蛋白质序列可能更难以有效预测。

4.3.2 不同滑动窗口的影响

在探索序列特征对 PPI 位点预测的影响时，本文不仅考虑了输入特征的多样性，还对局部特征的尺寸进行了分析。通过实验，本文设置不同尺寸的滑动窗口以评估它们对 PPIS-MFH 模型性能的影响。如表 5 所列，结果表明使用长度为 7 的滑动窗口时，模型达到了最优的 F-measure，MCC 与 AUC PR 性能。

表 5 PPIS-MFH 中应用不同尺寸滑动窗口的性能对比结果

Table 5 Performance comparison results of applying different sizes of sliding windows in PPIS-MFH

滑动窗口尺寸	*Acc*	*Precision*	*Recall*	*F-measure*	*MCC*	*AUC PR*
7	**0.658**	**0.317**	0.625	**0.420**	**0.238**	**0.362**
9	0.591	0.280	**0.680**	0.397	0.198	0.337
11	0.636	0.297	0.609	0.399	0.205	0.340
13	0.599	0.281	0.655	0.393	0.192	0.334
15	0.650	0.303	0.590	0.401	0.209	0.337

4.3.3 消融实验

1）为了探究不同类型输入特征在 PPIS-MFH 模型中的具体贡献，本文逐一移除各输入特征并观察模型性能的变化。

通过比较完整模型与去除特定特征模型的性能差异，可以评估每种特征的重要性。如表 6 所列，其中“P”表示 PSSM，“D”表示二级结构，“S”表示蛋白质序列。通过数据可以得出结论，原始蛋白质序列与 PSSM 对于 PPIS-MFH 模型的性能至关重要。从模型中移除原始蛋白质序列后，Acc，F-measure 和 MCC 分别从 0.658，0.420 和 0.238 显著下降至 0.582，0.385 和 0.177。同样，从模型中移除 PSSM 后，Acc，F-measure 和 MCC 分别下降到 0.589，0.390 和 0.178。相比之下，二级结构信息的缺失对模型性能的影响较小。没有二级结构信息的情况下，F-measure 和 MCC 分别下降到 0.399 和 0.210。这些结果表明，尽管 3 种输入特征各有其作用，但在 PPIS-MFH 模型中，原始蛋白质序列的贡献最为显著，而 PSSM 和二级结构信息则起到了辅助作用。

表 6 PPIS-MFH 中应用不同输入特征的性能对比结果

Table 6 Comparative performance results of applying different input features in PPIS-MFH

选取特征	*Acc*	*Precision*	*Recall*	*F-measure*	*MCC*	*AUC PR*
PDS	0.658	**0.317**	**0.625**	**0.420**	**0.238**	**0.362**
PD	0.582	0.271	0.661	0.385	0.178	0.326
PS	**0.668**	0.311	0.556	0.399	0.210	0.349
SD	0.589	0.276	0.664	0.390	0.177	0.330

2）通过删除 ViT 网络来证明本文方法的有效性。

表 7 展示了 PPIS-MFH 中应用 ViT 网络前后的模型性能对比结果，其中“Y”表示应用，“N”表示未应用。可以看到，在没有应用 ViT 网络的情况下，全部性能都有所下降，ACC，Precision，Recall，F-measure，MCC 和 AUC PR 分别从 0.676，0.325，0.587，0.418，0.237 和 0.361 下降到 0.607，0.272，0.585，0.371，0.159 和 0.287。

表 7　PPIS-MFH 中应用 ViT 网络前后的模型性能对比结果

Table 7　Comparison results of model performance before and after applying ViT network in PPIS-MFH

应用 ViT	*Acc*	*Precision*	*Recall*	*F-measure*	*MCC*	*AUC PR*
Y	**0.658**	**0.317**	**0.625**	**0.420**	**0.238**	**0.362**
N	0.607	0.272	0.585	0.371	0.159	0.287

此外，本文采用 Patch Embedding 线性映射层的不同维度[128，256，512]对嵌入特征进行分析。由表 8 可以看出，当线性映射层表示维度为 256 时，ACC，Precision，F-measure，MCC 与 AUC PR 这 5 个评价指标的表现最佳。

表 8　PPIS-MFH 中应用 ViT 网络中不同维度线性映射层的性能对比结果

Table 8　Performance comparison results of different dimensional linear mapping layers in the applied ViT network in PPIS-MFH

维度	*Acc*	*Precision*	*Recall*	*F-measure*	*MCC*	*AUC PR*
128	0.621	0.292	0.645	0.403	0.209	0.345
256	**0.658**	**0.317**	0.625	**0.420**	**0.238**	**0.362**
512	0.603	0.288	**0.685**	0.406	0.214	0.344

3）通过对模块采用不同组合方式以验证多特征交叉网络各部分的重要性。

如表 9 所列，这些模块在单独和组合使用时表现出了独特的性能特点。可以发现，单独使用 Multi-Head Attention TextCNN 模块时，模型在负面类别的辨识以及整体的分类准确度上展现了显著优势。相比之下，TextRNN-Attention 单独使用时模型更擅长识别正样本。图 3 的对比进一步印证了这两种模块的互补性，揭示了它们共同作用时能够带来更全面的提升，从而实现了模型性能的整体最优化。

表 9　特征交叉网络中应用不同模块的性能对比结果

Table 9　Performance comparison results of applying different modules in feature crossing networks

模块选取	*Acc*	*Precision*	*Recall*	*F-measure*	*MCC*	*AUC PR*
Multi-Head AttentionTextCNN 模块	**0.687**	**0.324**	0.533	0.403	0.219	0.344
TextRNN-Attention 模块	0.596	0.285	**0.689**	0.403	0.209	0.337
Both	0.658	0.317	0.625	**0.420**	**0.238**	**0.362**

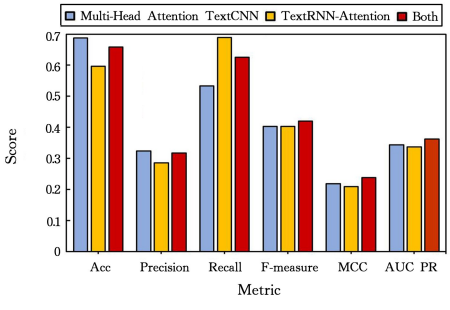


图 3　特征交叉网络不同组成成分的性能表现

Fig. 3　Performance of different components of feature crossing networks

4）在评估全局特征对 PPIS-MFH 模型性能的影响时，本文去除了全局特征，并观察了这一变化对模型性能的影响。

根据表 10 的数据对比可以得出结论：全局序列特征对 PPIS-MFH 模型的性能起到了关键作用。在没有全局序列特征的情况下，ACC，Precision，F-measure，MCC 与 AUC PR 分别从 0.676，0.325，0.418，0.237 和 0.361 下降到 0.472，0.239，0.364，0.136 和 0.287。这些性能的下降表明了全局特征在提升分类性能方面的有效性。

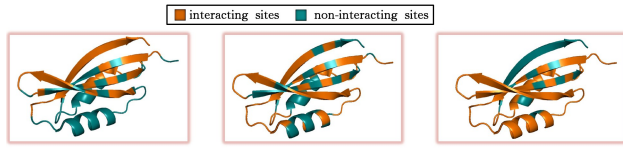
表 10　PPIS-MFH 中应用全局特征前后性能对比结果

Table 10　Performance comparison results before and after applying global features in PPIS-MFH

特征	*Acc*	*Precision*	*Recall*	*F-measure*	*MCC*	*AUC PR*
局部特征	0.472	0.239	**0.765**	0.364	0.136	0.287
全局特征＋局部特征	**0.658**	**0.317**	0.625	**0.420**	**0.238**	**0.362**

4.3.4　可视化分析

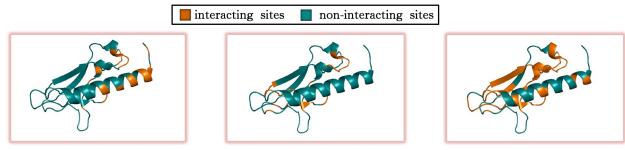
为了更加直观地展现 PPIS-MFH 模型在 PPI 位点预测方面的性能，本文在 Dset_186_72_PDB164 数据集的测试集上选取了 5 个蛋白质样本进行结果可视化。如图 4－图 8 所示，可以清晰观察到 PPIS-MFH 成功识别出了大多数的相互作用位点以及非相互作用位点。同时，本文还在相同的测试集上将 PPIS-MFH 模型与 DeepPPISP[19] 方法进行了性能比对。图中用橘色标出了结合位点，而绿色则代表非结合位点。从结果可以看出，PPIS-MFH 在 PPI 位点预测方面更具优势。



(a) True(90 sites)　(b) PPIS-MFH (60 Correct sites)　(c) DeepPPISP (50 Correct sites)

图 4　PDB 编号为 1F60_Chain B 的蛋白质实例可视化（电子版为彩图）

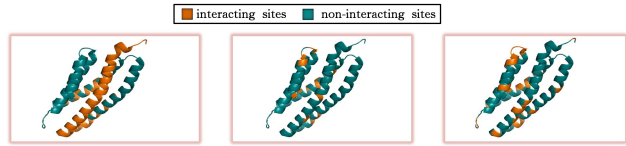
Fig. 4　Visualization of protein instance with PDB number 1F60_Chain B



(a) True(104 sites)　(b) PPIS-MFH (78 Correct sites)　(c) DeepPPISP (54 Correct sites)

图 5　PDB 编号为 2P7V_Chain A 的蛋白质实例可视化（电子版为彩图）

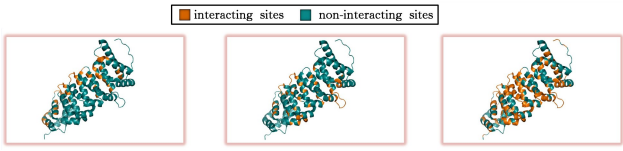
Fig. 5　Visualization of protein example with PDB number 2P7V_Chain A



(a) True(151 sites)　(b) PPIS-MFH (129 Correct sites)　(c) DeepPPISP (107 Correct sites)

图 6　PDB 编号为 3C5T_Chain A 的蛋白质实例可视化（电子版为彩图）

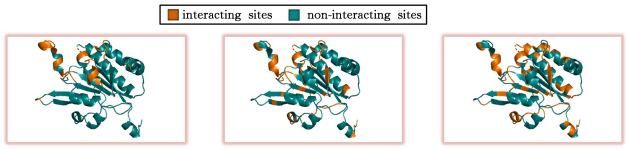
Fig. 6　Visualization of protein example with PDB number 3C5T_Chain A



(a) True(335 sites) (b) PPIS-MFH (263 Correct sites) (c) DeepPPISP (180 Correct sites)

图 7 PDB 编号为 3GNI_Chain A 的蛋白质实例可视化(电子版为彩图)

Fig. 7 Visualization of protein example with PDB number 3GNI_Chain A



(a) True(215 sites) (b) PPIS-MFH (165 Correct sites) (c) DeepPPISP (136 Correct sites)

图 8 PDB 编号为 3VU9_Chain A 的蛋白质实例可视化(电子版为彩图)

Fig. 8 Visualization of protein example with PDB number 3VU9_Chain A

结束语 本文提出了多特征融合模型 PPIS-MFH,旨在通过整合全局与局部特征来提升 PPI 位点的预测性能。具体来说,本文使用 ViT 网络解析局部特征及其上下文信息。此外,本文还引入了一个特征交叉网络,用于深挖二级结构、位置特异性得分矩阵以及整个蛋白质序列的潜在特征。在公开的 PPI 位点预测数据集上进行验证时,本文模型展现出了相较于其他方法的优势。

在本文的后续工作中,将持续研究如何有效结合进化信息、二级结构和原始蛋白序列的特征,探索它们之间的结构关系,并进一步简化模型结构,以提升模型的运行效率,并拓展其适用性。

参 考 文 献

[1] DAS S, CHAKRABARTI S. Classification and prediction of protein-protein interaction interface using machine learning algorithm [J]. Scientific Reports, 2021, 11(1): 1761.

[2] BUTLAND G, PEREGRÍN-ALVAREZ J M, LI J, et al. Interaction network containing conserved and essential protein complexes in Escherichia coli [J]. Nature, 2005, 433(7025): 531-537.

[3] LI X, LI W, ZENG M, et al. Network-based methods for predicting essential genes or proteins: a survey [J]. Briefings in Bioinformatics, 2020, 21(2): 566-583.

[4] DE LAS RIVAS J, FONTANILLO C. Protein-protein interactions essentials: key concepts to building and analyzing interactome networks [J]. PLoS Computational Biology, 2010, 6(6): e1000807.

[5] BRETTNER L M, MASEL J. Protein stickiness, rather than number of functional protein-protein interactions, predicts expression noise and plasticity in yeast [J]. BMC Systems Biology, 2012, 6: 1-10.

[6] TERENTIEV A A, MOLDOGAZIEVA N T, SHAITAN K V. Dynamic proteomics in modeling of the living cell. Protein-protein interactions [J]. Biochemistry (Moscow), 2009, 74: 1586-1607.

[7] WODAK S J, VLASBLOM J, TURINSKY A L, et al. Protein-protein interaction networks: the puzzling riches [J]. Current Opinion in Structural Biology, 2013, 23(6): 941-953.

[8] LI Y, GOLDING G B, ILIE L. DELPHI: accurate deep ensemble model for protein interaction sites prediction [J]. Bioinformatics, 2021, 37(7): 896-904.

[9] HOU Q, DE GEEST P F G, VRANKEN W F, et al. Seeing the trees through the forest: sequence-based homo-and heteromeric protein-protein interaction sites prediction using random forest [J]. Bioinformatics, 2017, 33(10): 1479-1487.

[10] HOU Q, LENSINK M F, HERINGA J, et al. Club-martini: selectingfavourable interactions amongst available candidates, a coarse-grained simulation approach to scoring docking decoys [J]. PloS One, 2016, 11(5): e0155251.

[11] ZHOU Y, JIANG Y, YANG Y. AGAT-PPIS: A novel protein-protein interaction site predictor based on augmented graph attention network with initial residual and identity mapping [J]. Briefings in Bioinformatics, 2023, 24(3): bbad122.

[12] PITRE S, DEHNE F, CHAN A, et al. PIPE: a protein-protein interaction prediction engine based on the re-occurring short polypeptide sequences between known interacting protein pairs [J]. BMC Bioinformatics, 2006, 7: 1-15.

[13] OFRAN Y, ROST B. Predicted protein-protein interaction sites from local sequence information [J]. FEBS Letters, 2003, 544(1/2/3): 236-239.

[14] MURAKAMI Y, MIZUGUCHI K. Applying the Naïve Bayes classifier with kernel density estimation to the prediction of protein-protein interaction sites [J]. Bioinformatics, 2010, 26(15): 1841-1848.

[15] YOUSEF A, CHARKARI N M. A novel methodbased on new adaptive LVQ neural network for predicting protein-protein interactions from protein sequences [J]. Journal of Theoretical Biology, 2013, 336: 231-239.

[16] SINGH G, DHOLE K, PAI P P, et al. SPRINGS: prediction of protein-protein interaction sites using artificial neural networks [R]. PeerJ PrePrints, 2014.

[17] WANG B, CHEN P, WANG P, et al. Radial basis function neural network ensemble for predicting protein-protein interaction sites in heterocomplexes [J]. Protein and Peptide Letters, 2010, 17(9): 1111-1116.

[18] KOIKE A, TAKAGI T. Prediction of protein-protein interaction sites using support vector machines [J]. Protein Engineering Design and Selection, 2004, 17(2): 165-173.

[19] WANG X, YU B, MA A, et al. Protein-protein interaction sites prediction by ensemble random forests with synthetic minority oversampling technique [J]. Bioinformatics, 2019, 35(14): 2395-2402.

[20] ZENG M, ZHANG F, WU F X, et al. Protein-protein interaction site prediction through combining local and global features with deep neural networks [J]. Bioinformatics, 2020, 36(4): 1114-1120.

[21] ZHANG B,LI J,QUAN L,et al. Sequence-based prediction of protein-protein interaction sites by simplified long short-term memory network [J]. Neurocomputing,2019,357:86-100.

[22] LU S,LI Y,NAN X,et al. Attention-based convolutional neural networks for protein-protein interaction site prediction [C]// 2021 IEEE International Conference on Bioinformatics and Biomedicine(BIBM). IEEE,2021:141-144.

[23] CONG H,LIU H,CAO Y,et al. Protein-protein interaction site prediction by modelensembling with hybrid feature and self-attention [J]. BMC Bioinformatics,2023,24(1):456.

[24] WANG X,YU B,MA A,et al. Protein-protein interaction sites prediction by ensemble random forests with synthetic minority oversampling technique [J]. Bioinformatics,2019,35(14):2395-2402.

[25] JOOSTEN R P,TE BEEK T A H,KRIEGER E,et al. A series of PDB related databases for everyday needs [J]. Nucleic Acids Research,2010,39(suppl_1):D411-D419.

[26] KABSCH W,SANDER C. Dictionary of protein secondary structure:pattern recognition of hydrogen-bonded and geometrical features [J]. Biopolymers:Original Research on Biomolecules,1983,22(12):2577-2637.

[27] WANG J,YANG B,REVOTE J,et al. POSSUM:a bioinformatics toolkit for generating numerical sequence feature descriptors based onPSSM profiles [J]. Bioinformatics,2017,33(17):2756-2758.

[28] WODAK S J,VLASBLOM J,TURINSKY A L,et al. Protein-protein interaction networks:the puzzling riches[J]. Current Opinion in Structural Biology,2013,23(6):941-953.

HU Zhaolong, born in 2000, postgraduate. His main research interests include deep learning and bioinformatics.

HU Chunling, born in 1970, Ph.D, professor, is a member of CCF (No. 18622M). Her main research interests include machine learning and bioinformatics.