小 型 微 型 计 算 机 系 统
Journal of Chinese Computer Systems
DOI:10.20009/j.cnki.21-1106/TP.2025-0096
2026 年 4 月 第 4 期
Vol.47 No.4 2026

融合图注意力网络的蛋白质相互作用位点预测方法

张 晴[1],张 洋[1],周晓根[1],张 彪[1],胡 俊[1,2]

[1](浙江工业大学 信息工程学院,杭州 310023)
[2](苏州系统医学研究院 人工智能与计算生物学研究中心,江苏 苏州 215123)
Email:junh_cs@126.com

摘 要: 识别蛋白质-蛋白质相互作用位点对解析生命过程和药物设计至关重要.近年来,基于深度学习的方法在蛋白质相互作用位点预测中取得了显著进展.本文提出了一种新型深度学习模型 GL-GAT,该模型整合了卷积神经网络(CNN)、双向长短时记忆网络(Bi-LSTM)以及图注意力网络(GAT)以充分挖掘蛋白质序列与三维结构信息.具体而言,首先利用 CNN 提取蛋白质语言模型 ESM-2 生成的序列特征,并采用 Bi-LSTM 提取位置特异性得分矩阵、二级结构和位置信息;然后,将上述特征拼接生成节点特征,并基于蛋白质三维结构构建蛋白质图;最后,通过 GAT 聚合邻域信息,并通过全连接神经网络输出残基位点预测概率.实验结果表明,相较于现有主流方法,GL-GAT 在多个关键性能指标上均表现出更高的预测精度.

关 键 词: 蛋白质-蛋白质相互作用位点预测;蛋白质语言模型;图注意力网络
中图分类号: TP301 **文献标识码:** A **文 章 编 号:** 1000-1220(2026)04-0886-08

Protein-protein Interaction Site Prediction Method Integrating Graph Attention Network

ZHANG Qing[1],ZHANG Yang[1],ZHOU Xiaogen[1],ZHANG Biao[1],HU Jun[1,2]

[1](College of Information Engineering,Zhejiang University of Technology,Hangzhou 310023,China)
[2](Center for AI and Computational Biology,Suzhou Institution of Systems Medicine,Suzhou 215123,China)

Abstract: Identifying protein-protein interaction (PPI) sites is crucial for understanding biological processes and drug design. In recent years, deep learning-based methods have made significant progress in predicting PPI sites. This paper proposes a novel deep learning model, GL-GAT, which integrates Convolutional Neural Networks (CNN), Bidirectional Long Short-Term Memory networks (Bi-LSTM), and Graph Attention Networks (GAT) to effectively leverage both protein sequence and three-dimensional structural information. Specifically, CNN is first used to extract sequence features generated by the protein language model ESM-2, and Bi-LSTM is employed to capture position-specific scoring matrix, secondary structure, and spatial location information. These features are then concatenated to form node representations, and a protein graph is constructed based on the protein's three-dimensional structure. Finally, GAT is used to aggregate neighborhood information, and a fully connected neural network is applied to predict residue-level binding probabilities. Experimental results demonstrate that, compared to current state-of-the-art methods, GL-GAT achieves superior predictive accuracy across multiple critical performance metrics.

Keywords: protein-protein interaction site prediction; protein language modeling; graph attention network

0 引 言

蛋白质是生命活动的核心物质,是细胞的主要有机成分之一,广泛参与基因转录、蛋白质合成、免疫应答、细胞信号传导及物质运输等生物学过程[1].蛋白质的生物学功能通常依赖于与其他蛋白质分子间的特异性相互作用,这种相互作用被称为蛋白质-蛋白质相互作用(Protein-Protein Interaction, PPI)[2].PPI 位点是蛋白质在相互作用过程中参与结合的残基,其准确识别不仅对揭示蛋白质功能机制具有重要理论意义,还为靶向药物设计提供了关键理论依据.近年来,实验生物学在蛋白质相互作用研究方面取得了重大进展,开发了包括下拉测定[3]、共免疫沉淀[4]、表面等离子体共振[5]、细菌双杂交[6]和细胞学双杂交[7]等多种实验技术.这些方法在 PPI 及其作用位点鉴定方面发挥了重要作用,并推动大量 PPI 复合物结构的解析,许多蛋白质结构已被收录于蛋白质数据库[8].然而,传统实验方法普遍存在通量低、耗时较长、成本较高等局限性,难以满足大规模 PPI 位点预测的需求.因此,开发高效、自动化的计算预测方法已成为生物信息学的重要方向,为大规模 PPI 位点提供了新方案.

随着机器学习和深度学习技术的快速发展,过去 20 年间涌现了多种计算方法用于蛋白质相互作用位点的预测[8].这些方法主要可分为 3 类:基于结构的方法、基于序列的方法以

收稿日期:2025-03-07 收修改稿日期:2025-04-15 基金项目:国家自然科学基金项目(61902352,62203389,62201506)资助;浙江省自然科学基金项目(LY21F020025)资助;浙江省属高校基本科研业务费项目(RF-A20200012)资助. 作者简介:张 晴,女,1999 年生,硕士研究生,研究方向为生物信息学、计算智能及深度学习;张 洋,男,2002 年生,硕士研究生,研究方向为生物信息学、计算智能及深度学习;周晓根(通信作者),男,1987 年生,博士,教授,博士生导师,CCF 会员,研究方向为生物信息学、计算智能及深度学习;张 彪,男,1989 年生,博士,讲师,研究方向为生物信息学、计算智能及深度学习;胡 俊,男,1989 年生,博士,副研究员,研究方向为生物信息学、计算智能及深度学习.

及结合结构和序列信息的混合方法. 早期研究主要集中在基于结构的方法,这类方法利用蛋白质的三维结构信息进行PPI 位点预测. 例如,Bradford 等人[9]通过表面补丁分析技术,从蛋白质三维结构数据中提取局部结构特征,实现 PPI 位点预测. Chen 等人[10]通过分析蛋白质表面残基的非共价相互作用原子三维概率密度图,挖掘关键判别信息. 然而,基于结构的方法依赖于蛋白质的真实结构信息,但由于结构数据的获取成本较高,且已知的蛋白质结构数量有限,这限制了此类方法在大规模 PPI 位点预测中的应用[11]. 随着高通量测序技术的快速发展,蛋白质序列数据的获取变得更加便捷和经济,推动了基于序列的方法的兴起. 这类方法在缺乏结构信息的情况下取得了重要进展,并逐渐成为 PPI 位点预测领域的主流研究方向.

基于序列的 PPI 位点预测方法主要通过特征选择和深度学习模型的优化来提升预测精度. 在特征选择方面,现有方法主要采用序列衍生特征和蛋白质语言模型特征. 序列衍生特征主要包括物理化学性质、进化信息以及结构信息等. 例如,位置特异性得分矩阵(Position Specific Scoring Matrix, PSSM)通过计算进化背景下残基的替换概率,能够有效刻画位点的保守性,被广泛应用于 PPI 位点预测. 此外,基于 DSSP[12](Dictionary of Secondary Structure of Proteins)算法提供的二级结构信息可揭示蛋白质折叠状态,与相互作用位点密切相关. 蛋白质语言模型特征基于自监督学习,能够从大规模蛋白质序列中提取深层次的表示向量,无需额外的特征工程. 例如,MSA Transformer[13]生成的 768 维嵌入和 Prontans[14]生成的 1024 维嵌入已被应用于 PPI 位点预测. 与序列衍生特征相比,蛋白质语言模型能够捕捉全局上下文信息,提供更丰富的表征能力. 除了特征选择,深度学习模型的优化也在不断提升 PPI 位点预测的精度. 近年来,多种基于深度学习的方法结合不同特征,构建了卷积神经网络(CNN)、循环神经网络(RNN)和 Transformer 等模型,例如:DLPred[15]通过结合简化长短时记忆网络(SLSTM)与序列衍生特征(PSSM 和理化性质)进行 PPI 位点预测;DELPHI[16]采用 CNN、RNN 和微调技术,并整合多种序列衍生特征(如 PSSM、理化性质等),提升预测性能;EnsemPPIS[17]结合 Transformer 与门控卷积网络,并引入 Prontans 嵌入,进一步增强了模型的表征能力;Seq-Insite[11]则融合 MSA Transformer 和 Prontans 嵌入,并结合多层感知机和长短期记忆网络(LSTM),提高了基于序列的 PPI 位点预测能力. 此外,滑动窗口机制在 PPI 位点预测中常用于提取序列的局部上下文信息. DeepPPISP[18]、HN-PPISP[19]、ProB-Site[20]和 D-PPIsite 等方法均采用该策略,在蛋白质序列上使用滑动窗口提取固定长度的局部片段,以增强模型对局部环境的感知能力,提升预测性能. 尽管上述基于序列的 PPI 位点预测方法在一定程度上提升了预测精度,但它们主要依赖序列信息,而忽略了蛋白质的空间结构信息. 由于蛋白质折叠的影响,序列相邻的残基在三维结构中可能距离较远,而非相邻残基可能在空间上形成关键的相互作用. 这种空间信息的缺失,可能影响预测的准确性. 因此,近年来研究者开始探索结合序列与结构信息的方法,以进一步提升 PPI 位点预测性能.

近年来,随着图神经网络的快速发展,研究人员逐渐将 PPI 位点预测问题视为图节点分类任务. 该方法的核心在于使用蛋白质的三维结构信息构建蛋白质图,并将序列信息作为节点特征嵌入,以融合蛋白质序列与结构信息,通过图神经网络提取残基之间的空间关系. GraphPPIS[21]首次将图神经网络应用于 PPI 位点预测,采用图卷积网络(Graph Convolutional Network,GCN)框架,将 PSSM、HMM 和 DSSP 特征等序列衍生特征整合为节点特征,并基于残基间的构建蛋白质图. RGN[22]在此基础上引入了图注意力网络(Graph Attention Network,GAT),结合 GCN-GAT 混合架构,以提取更深层次的特征,并整合 PSSM、HMM、氢键估计算以及 ProtTrans 嵌入作为节点特征. 随后,EGRET[23]提出基于图注意力的边缘聚合网络,将 ProtTrans 嵌入作为节点特征,并引入结构信息作为边特征,以增强模型对残基空间关系的学习能力. 尽管基于序列与结构的混合方法在 PPI 位点预测中取得了一定进展,但仍存在局限性. 首先,大多数方法依赖 PDB(Protein Data Bank)库的高精度蛋白质结构,而 PDB 库的覆盖范围有限,导致训练数据不足,影响模型学习能力和预测精度. 其次,现有方法未能充分利用 TrRosetta[24]、I-TASSER[25]、AlphaFold3[26]等结构预测工具,扩充训练数据,限制了模型的泛化能力. 此外,在序列信息融合方面,未能有效结合包含序列上下文信息的局部特征和反映整体序列信息的全局特征,这进一步限制了模型的预测性能.

本文提出了一种基于全局与局部特征的图注意力网络的蛋白质相互作用位点预测方法 GL-GAT. 为提升模型的特征学习能力,本研究针对输入序列特征,采用不同策略分别提取全局特征和局部特征. 全局特征由蛋白质语言模型 ESM-2 生成,并输入一维卷积网络(Conv1d)进行特征提取;局部特征则由滑动窗口提取位置特异性得分矩阵、蛋白质二级结构及位置信息中提取,并输入双向长短时记忆网络(Bi-LSTM)进行特征学习. 随后,将融合后的全局与局部特征作为节点特征,并结合蛋白质结构信息构建的蛋白质图,输入至图注意力网络学习. 最终,模型采用全连接层进行分类,输出每个残基作为 PPI 结合位点的概率. 考虑到 PDB 库结构数据的覆盖范围有限,本研究利用 AlphaFold2 预测的蛋白质结构补充训练数据,以提升模型的结构学习能力. 基准测试结果表明,GL-GAT 在多个关键性能指标上均优于现有主流方法.

1　材料和方法

1.1　基准数据集

本研究采用 7 个独立的测试数据集对 GL-GAT 的预测性能进行系统性评估,包括 Dset_186[27]、Dset_164[28]、Dset_72[27]、Dset_70[29]、Dset_60[30]、Dset_448[31]和 Dset_355[16]. Dset_186 包含 186 条蛋白质序列,这些序列源自 105 个异源二聚体蛋白复合物,且序列相似性均低于 25%. Dset_72 由 72 条蛋白质序列组成,通过 BLASTClust[32]筛除与 Dset_186 序列相似性≥25% 的序列后构建而成. Dset_164 包含 164 条蛋白质序列,数据来源于 2010 年 6 月 ~2013 年 11 月期间新增注释的 PDB 数据库条目,其构建遵循与 Dset_186 和 Dset_72 相同的严格筛选标准,以确保数据集的非冗余性和代表性. 为提升数据集的多样性和覆盖范围,GraphPPIS 将 Dset_186、

Dset_72 和 Dset_164 整合为一个统一数据集,并利用 BLAST-Clust 去除序列相似性超过 25% 或重叠区域超过 90% 的冗余蛋白质,最终获得 395 条非冗余蛋白链. 其中,335 条用于训练集,剩余 60 条作为独立测试集(Dset_60). 同样,DeepPPISP 也将 Dset_186、Dset_72 和 Dset_164 组合成一个数据集,并按照 83.3% 的比例随机划分为训练集和测试集,最终得到的测试集包含 70 条蛋白链(Dset_70). Dset_448 包含 448 条蛋白质序列,数据来源于 BioLiP 数据库,且任意两条序列的序列相似性均低于 25%. 为进一步降低数据冗余度,Dset_355 由 DELPHI 构建,通过从 Dset_448 中删除与 DLPred 训练集中序列相似性超过 40% 的 93 条蛋白质序列而来.

在模型训练阶段,本研究采用了 Dset_9982 数据集,该数据集最初由 DELPHI 收集并整理,包含 9982 条非冗余蛋白质序列,其中包括 427687 个结合残基与 3826511 个非结合残基. 所有蛋白质序列之间的序列相似性均不超过 25%,此外,该数据集与 7 个独立测试集之间的序列相似性低于 25%,以确保模型评估的严谨性和泛化能力. 在训练过程中,本研究采用固定比例的训练-验证拆分策略,将训练数据集按 9∶1 的比例随机划分,其中 90% 的数据用于模型训练,剩余 10% 用于验证. 相关数据集的详细统计信息见表1. 为保证数据输入的一致性,在训练过程中对蛋白质序列长度进行了标准化处理,将所有序列长度调整至 1000 个残基. 对于超过 1000 个残基的序列,采用截断处理;对于短于 1000 个残基的序列,则通过零填充补齐至 1000.

表 1 数据集信息

Table 1 Information of datasets

数据集	蛋白数	残基数	结合残基	非结合残基
Dset_9982	9982	4254198	427687	3826511
Dset_186	186	36219	5517	30702
Dset_164	164	33681	6096	27585
Dset_72	72	18140	1923	16217
Dset_60	60	13144	2075	11069
Dset_70	70	11791	2332	9459
Dset_448	448	116500	15810	100690
Dset_355	355	95940	11467	84473

1.2 全局序列特征

近年来,蛋白质语言模型生成的特征被用于 PPI 位点预测研究,包括 Prontans 和 MSA Transformer 嵌入等. 本文采用基于 Transformer 结构的蛋白质语言模型 ESM-2(Evolutionary Scale Modeling 2,ESM-2)[33] 生成的嵌入作为全局序列特征. ESM-2 通过大规模预训练学习氨基酸残基间的远程依赖关系,生成高维度的序列嵌入表示. 与传统的多序列比对(MSA)方法不同,ESM-2 无需同源序列输入,而是通过自注意力机制从单序列中提取丰富的信息. 该模型在 UniRef50 蛋白质数据集上训练,并为每个残基生成一个 1280 维的特征向量,为 PPI 位点预测提供高质量的全局序列表征.

1.3 局部序列特征

受 DeepPPIS、HN-PPISP、ProB-Site 和 D-PPIsite 方法的启发,本研究采用局部上下文编码技术生成局部序列特征. 具体而言,选取 PSSM、DSSP 和位置信息(Position Information,PI) 3 种特征,并结合滑动窗口机制提取局部序列特征. PSSM 特征是一个 20 维的向量,由 PSI-BLAST[34] 工具生成. 该工具采用 3 次迭代搜索非冗余数据库,并 0.001 作为 E 值阈值进行多序列比对,以获取每个氨基酸残基的进化信息. DSSP 特征是一个 14 维的向量,由 DSSP 程序提取,该程序基于氨基酸原子的空间排列和其键连接信息,预测蛋白质的二级结构. 位置信息特征为 1 维向量,表示每个氨基酸残基的相对位置信息. 蛋白质中第 i 个残基的 PI 值通过公式 i/L 计算,其中 L 是蛋白质的总残基数. 这些特征被拼接成一个维度为 $L\times35$ 的特征矩阵,并通过滑动窗口机制(窗口大小为 W)进行处理,最终得到维度为 $L\times W\times35$ 的局部序列特征矩阵,作为后续模型训练与预测的输入. 在特征融合过程中,PSSM、DSSP 和 PI 这 3 类特征被直接拼接,未施加额外的加权因子,因此在输入层的初始权重相同. 训练过程中,模型通过梯度更新动态调整各特征的权重分布.

1.4 蛋白质的图形表示

本研究将蛋白质的三维结构表示为图 G,其中每个节点对应于一个氨基酸残基. 为了捕捉残基之间的空间关系,每个节点连接 k 个(其中 $k=21$)个相邻节点,邻接关系的确定基于残基间的平均原子距离. 具体而言,对于给定的节点 i,通过计算该残基与其他残基之间所有原子对的欧几里得距离,并取其均值(即平均原子距离),以此衡量残基间的整体空间接近程度. 其中,蛋白质的原子坐标从蛋白质三维结构文件中获得. 随后,选取与节点 i 平均原子距离最近的 k 个残基作为其邻接节点. 这种图形构建方式确保充分保留蛋白质三维构象中的空间信息,为后续的图神经网络建模提供有效支持.

1.5 卷积双向 LSTM 图注意力架构

本研究提出了一种融合卷积和双向长短时记忆的图注意力网络架构,命名为 CBGAN(Convolutional BiLSTM on Graph Attention Network). CBGAN 的整体架构如图 1 所示. 如图 1 所示,在特征输入阶段,该方法的特征输入包括全局序列特征和局部序列特征. 全局序列特征由预训练的 ESM-2 模型生成,初始形状为 $L\times1280$. 为降低特征维度并提取深层次信息,使用卷积核大小为 21 的 Conv1d 层对全局序列特征进行处理,得到形状为 $L\times64$ 的全局特征表示. 局部序列特征则由 PSSM、DSSP 和 PI 通过滑动窗口方法生成,构成形状为 $L\times W\times35$ 的特征矩阵. 该矩阵被输入到一个包含 16 个单元的

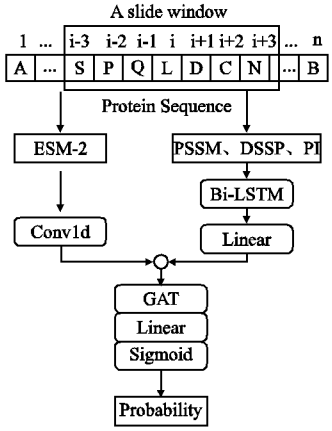

图 1 CBGAN 的架构

Fig. 1 Architecture of CBGAN

Bi-LSTM 层,以捕获序列中的上下文依赖关系,并输出形状为 $L\times W\times 32$ 的张量. 随后,线性层沿窗口维度进聚合特征,进一步降维至 $L\times 32$. 经过上述处理,全局序列特征的最终表示为 $\boldsymbol{H}_{global}\in R^{L\times 64}$,局部序列特征的最终表示为 $\boldsymbol{H}_{local}\in R^{L\times 32}$,二者沿着特征维度拼接,形成节点特征 $\boldsymbol{H}=\{\boldsymbol{h}_i\mid i=1,2,\cdots,L\}\in R^{L\times 96}$.

在图神经网络建模阶段,节点特征 $\boldsymbol{H}$ 和 1.4 节所述的蛋白质图结构一同输入至 GAT 层. 该 GAT 结构由两层网络组成,每层包含 3 个独立的注意力头,以计算不同的信息权重,捕捉更加丰富的特征表示. 各注意力头的输出在每一层末端进行拼接,生成形状为 $L\times 32$ 的节点表示 z_i,以增强模型的表达能力. 随后,节点表示 $\boldsymbol{z}_i$ 与初始节点特征 $\boldsymbol{h}_i$ 进行拼接,得到最终的融合特征 $\hat{\boldsymbol{h}}_i$. 最终,融合特征 $\hat{\boldsymbol{h}}_i$ 通过线性变换层映射至概率分布空间,输出结合位点的概率 P_i.

1.6　CBGAN 的实施

本研究采用 PyTorch 框架(1.10.2 版),在 GPU(NVIDIA GeForce RTX 3090)实现训练以及预测模型. 优化器采用自适应矩估计优化器(Adam),损失函数使用二元交叉熵损失函数(BCE Loss),如公式(1)所示:

$$BCE=-\frac{1}{N}\sum_{i}^{N}\left[l_i\,log(y_i)+(1-l_i)\,log(1-y_i)\right] \quad (1)$$

其中,N 是总样本数,l_i 是第 i 个样本的真实标签,y_i 是模型的预测值. 训练过程中,学习率、批次大小和训练轮数分别设置为 0.001、100 和 100.

1.7　评估指标

本研究采用 5 个常用的评估指标:准确度(Accuracy,*ACC*)、精确度(Precision,*PRE*)、灵敏度(Sensitivity,*SEN*)、特异性(Specificity,*SPE*)、*F*1 评分(*F*1)以及马修斯相关系数(Matthews Correlation Coefficient,*MCC*). 由于 *F*1 评分在精确度与灵敏度之间寻求平衡,适用于处理类别分布不均匀的情况,而 *MCC* 作为衡量分类质量的综合指标,能够更全面的评估模型性能. 因此,在 PPI 位点预测任务中,*F*1、*MCC* 对于评估模型的稳定性与可靠性尤为重要. 5 个评估指标计算见公式(2)~公式(7):

$$ACC=\frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

$$PRE=\frac{TP}{TP+FP} \quad (3)$$

$$SEN=\frac{TP}{TP+FN} \quad (4)$$

$$SPE=\frac{TN}{TN+FP} \quad (5)$$

$$F1=\frac{2\times SEN\times PRE}{SEN+PRE} \quad (6)$$

$$MCC=\frac{TP\times TN-FN\times FP}{\sqrt{(TP+FP)\times(TP+FN)\times(TN+FP)\times(TN+FN)}} \quad (7)$$

其中,真阳性(True Positive,*TP*)指被正确预测为交互位点的氨基酸残基;真阴性(True Negative,*TN*)指被正确预测为非交互位点的氨基酸残基;假阳性(False Positive,*FP*)指被错误预测为交互位点的氨基酸残基;假阴性(False Negative,*FN*)指被错误预测为非交互位点的氨基酸残基. 通常,在模型评估过程中会设定一个阈值,以使预测的交互位点数与实际位点数保持一致,确保 *SEN*、*PRE* 和 *F*1 相等. 此外,为了更全面地评估模型性能,本研究采用受试者工作特征曲线下面积(Area Under the Receiver Operating Characteristic Curve,*AUC*)和精确召回曲线下面积(Area Under the Precision-Recall Curve,*AUPR*)作为补充评价指标. *AUC* 反映模型在不同阈值下的整体分类能力,而 *AUPR* 更适用于不平衡分类问题,能够更准确地评估模型在正样本上的预测能力.

2　实验结果与分析

2.1　不同类型的局部特征对模型的影响

为了深入探讨局部特征中不同类型输入特征(PSSM、DSSP 和 PI)在模型中的贡献,本节设计并实施了一系列消融实验. 在实验过程中,滑动窗口大小固定为 1,以系统性地分析不同类型的输入特征作为局部特征对模型性能的影响. 实验结果表明,当模型同时整合所有特征时,其在验证集及测试集上的各项评估指标均达到最优状态. 具体来说,通过分别去除了 PSSM、DSSP 以及 PI,评估了每种特征的独立贡献.

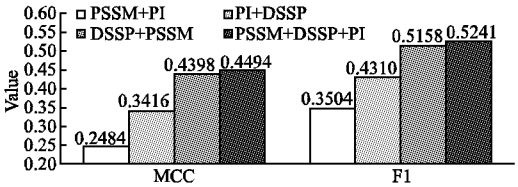

图 2　GL-GAT 在 Dest_448 数据集上不同类型的局部特征消融结果

Fig. 2　Ablation study of different types of local features for GL-GAT on the Dset_448 dataset

实验结果表明,如图 2 所示,当 3 种特征同时使用时,*MCC* 和 *F*1 均达到最高值,分别为 0.449 和 0.524. 然而,去除任何单一特征后,模型性能均出现下降:去除 PSSM,*MCC* 和 *F*1 分别降低 10.7%、9.3%;去除 DSSP,*MCC* 和 *F*1 分别降低 20.1%、17.4%;去除 PI,*MCC* 和 *F*1 分别降低 1%、0.8%. 结果表明,PSSM、DSSP 和 PI 特征在模型中携带互补的信息,并通过协同作用提升了预测性能. 值得注意的是,当 DSSP 特征被去除时,各项指标的下降幅度最为明显. 这一现象表明,DSSP 特征在基于局部特征的 PPI 位点预测中对模型性能起着关键作用. 综上所述,PSSM、DSSP 和 PI 特征的结合能够有效提升模型的预测性能.

2.2　不同长度的滑动窗口对模型的影响

本节系统地探讨了不同滑动窗口大小对模型性能的影响,并进行了详细分析. 具体而言,采用滑动窗口方法提取局部序列特征,同时去除全局序列特征,以确保实验结果能够准确反映不同滑动窗口尺寸对局部序列信息及模型表现的贡献.

对于给定滑动窗口大小 W,定义局部窗口大小 w,其中 $W=2w+1$. 对 w 在 1~9 之间的不同取值(即 $w=1,2,3,\cdots,9$)时 *MCC* 和 *F*1 的数值进行了评估,以分析滑动窗口大小对模型性能的影响. 实验结果表明,如图 3 所示,当 $w=3$ 时,*MCC* 和 *F*1 值分别为 0.459 和 0.533 高于其他窗口尺寸的 *MCC* 和

$F1$ 值. 此外,当 $w<3$ 时,MCC 和 $F1$ 值随着窗口大小的增大而增加,当 $w>3$ 时,MCC 和 $F1$ 值出现波动并逐渐下降,其值

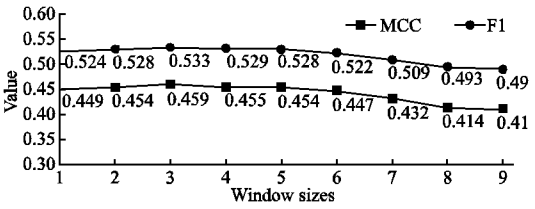

图 3 不同滑动窗口长度在验证集上对模型预测性能比较

Fig. 3 Comparison of model prediction performance with different sliding window lengths on the validation set

始终低于 $w=3$ 时的数值. 结果表明,滑动窗口大小的选取对模型预测至关重要.

2.3 不同全局序列特征对模型的影响

为探究全局序列特征的最优选择,本节对比了两种特征表示:1)ESM-2 生成的嵌入表示;2)传统特征组合 TSF(PSSM + DSSP + PI). 本节实验采用控制变量法,固定 GL-GAT 的局部特征模块,仅更换全局特征输入,并在多个独立测试集上进行验证. 如表 2 所示,在 Dset_448 和 Dset_164 数据集上,ESM-2 作为全局序列特征时,模型在所有评估指标上均优于 TSF,表明 ESM-2 具备更强的全局序列信息编码能力.

表 2 不同全局序列特征 ESM-2 和传统序列特征组合(TSF)对模型的影响

Table 2 Impact of different global sequence features (ESM-2) and traditional sequence features (TSF) on model performance

数据集	特征	F1	MCC	AUC	AUPR
Dest_448	ESM-2	0.582	0.516	0.880	0.597
	TSF	0.524	0.450	0.851	0.528
Dest_164	ESM-2	0.477	0.361	0.783	0.466
	TSF	0.448	0.326	0.766	0.430

在 Dset_448 数据集上,与 TSF 相比,ESM-2 特征使 MCC 提升 6.6%,$F1$ 提高 5.7%;在 Dset_164 数据集上,MCC 提升 3.5%,$F1$ 提高 2.9%. 此外,在 Dset_448 和 Dset_164 数据集上,ESM-2 还实现了更高的 AUC 和 $AUCPR$. 实验结果表明,ESM-2 能够有效编码关键序列信息,弥补 TSF 的不足,提升 PPI 位点预测性能.

2.4 不同网络框架之间的性能比较

为验证 GL-GAT 网络架构 PPI 位点预测任务中的有效性,本节设计了两种对比模型进行实验:1)GL-GCN,采用 GCN 代替 GAT 进行信息聚合,保持相同的蛋白质图结构输入,验证 GAT 中注意力机制相较于传统图卷积的特征提取能力;2)GL-BiLSTM,仅输入序列特征,移除蛋白质图结构信息,评估空间结构信息在 PPI 位点预测中的作用.

在 Dset_448 数据集上的实验结果表明,如图 4 所示,GL-GAT 展现出良好性能,其 MCC、$F1$、AUC 和 $AUPR$ 指标分别达到 0.516、0.582、0.880 和 0.597. 具体而言:相较于 GL-BiLSTM,GL-GAT 通过引入图结构信息使 4 项指标分别提升 3.5%、3.1%、1.7% 和 3.5%,表明空间结构信息对于 PPI 位点识别至关重要. 而与 GL-GCN 的对比显示,相较于传统图卷积,GL-GAT 通过注意力机制进一步提升了模型性能,4 项指标分别提升 5.3%、4.6%、3.1% 和 6.1%,这一结果表明,GAT 的注意力机制能够在建模残基间关联关系时赋予不同

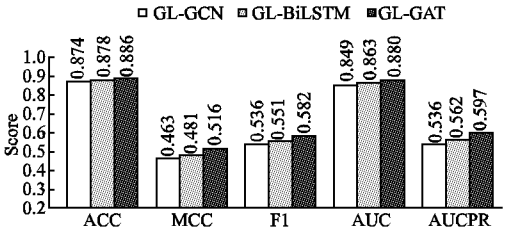

图 4 不同模型框架对模型预测性能比较

Fig. 4 Comparison of model prediction performance across different model architectures

邻居节点不同的重要性权重,从而增强关键位点的识别能力. 此外,GL-GAT 在 PPI 位点预测任务中的优越性能进一步证明了其能够有效融合蛋白质的空间结构信息与序列特征,提高预测精度.

2.5 与其他方法的预测能力比较

本节对 GL-GAT 模型的预测性能进行了系统性评估,并与 9 种基于序列的 PPI 位点预测方法(DELPHI、ProNA2020、EnsemPPIS、PIPENN、ProB-Site、D-PPIsite、SENSDeep、PITHIA 和 Seq-Insite)进行了比较. 同时,还与 4 种基于结构的方法(EGRET、GraphPPIS、RGN 和 MaSIF-site)进行了对比,以验证其泛化能力.

实验结果表明,如表 3 所示,GL-GAT 在多个数据集上均取得最佳表现,尤其是在 MCC、$F1$、AUC 和 $AUPR$ 等关键指标上展现出明显优势. 例如,在 Dset_448 数据集上,GL-GAT 的 MCC、$F1$、AUC 和 $AUPR$ 分别为 0.516、0.582、0.880 和 0.597,较表现最优的 Seq-Insite 分别提升 5.4%、4.6%、2.1% 和 4.5%. 在 Dset_72 数据集上,GL-GAT 相比表现最佳的 D-PPIsite,在 MCC、$F1$、AUC 和 $AUPR$ 指标上分别提升 2.8%、2.5%、3.8% 和 2.0%. 此外,在结构数据集 Dset_70 上,GL-GAT 的 MCC、$F1$、AUC 和 $AUPR$ 分别为 0.327、0.460、0.740 和 0.426,相比于结构方法中表现最佳的 EGRET,分别提升 5.7%、2.2%、2.1% 和 2.1%,进一步验证了 GL-GAT 在蛋白质相互作用位点预测任务中的有效性.

2.6 案例分析

本节选择 Dset_355 和 Dset_60 数据集中的两个蛋白质(B4E5Y6 和 3cqcA)作为案例研究对象. 首先,将 GL-GAT 与两种基于序列的方法 PITHIA 和 Seq-InSite 在 B4E5Y6 蛋白质上的表现进行比较. B4E5Y6 蛋白质包含 17 个结合位点和 151 个非结合位点. 如图 5 所示,GL-GAT 预测得到 17 个真阳性、140 个真阴性、11 个假阳性以及 0 个假阴性. 与 PITHIA 和 Seq-InSite 相比,GL-GAT 在识别真实结合位点方面表现更优,且假阴性数量更少表明其在预测的稳定性和整体准确度上有明显优势.

接下来,分析 GL-GAT 在 3cqcA 蛋白质上的预测性能,并与两种基于结构的预测方法 GraphPPIS 和 RGN 进行比较. 3cqcA 蛋白质包含 22 个结合位点和 99 个非结合位点. 如图 6

所示,GL-GAT 预测得到 19 个真阳性、82 个真阴性、17 个假阳性和 3 个假阴性. 相比于 RGN,GL-GAT 在真阳性预测方面表现更优,并有效降低了预测假阴性的数量. 尽管在真阴性和假阴性方面略低于 GraphPPIS,但 GL-GAT 的值为 0.583,

表 3　GL-GAT 与基于序列和基于结构的预测方法的性能比较

Table 3　Performance comparison of GL-GAT with sequence-based and structure-based prediction method

Dataset	Model	ACC	PRE	SPE	SEN	MCC	F1	AUC	AUPR
Dset_448	PITHIA[a]	0.840	0.408	0.907	0.408	0.315	0.408	0.778	0.387
	PIPENN *	0.860	0.470	0.870	0.470	0.254	0.385	0.729	0.357
	D-PPIsite *	0.859	0.480	0.919	0.480	0.399	0.480	0.824	0.479
	Seq-Insite *	0.874	0.535	0.927	0.535	0.462	0.535	0.859	0.552
	GL-GAT	**0.886**	**0.581**	**0.934**	**0.581**	**0.516**	**0.581**	**0.880**	**0.597**
Dset_335	PITHIA[a]	0.793	0.444	0.873	0.444	0.317	0.444	0.768	0.442
	D-PPIsite *	0.871	0.460	0.927	0.460	0.387	0.460	0.822	0.448
	Seq-Insite *	0.841	**0.573**	0.902	**0.573**	0.476	**0.573**	0.853	**0.617**
	Seq-Insite[f]	0.886	0.525	0.935	0.525	0.460	0.525	0.859	0.532
	GL-GAT	**0.897**	0.571	**0.941**	0.5571	**0.512**	0.571	**0.880**	0.573
Dset_164	DELPHI[a]	0.765	0.274	0.914	0.274	0.189	0.274	0.711	0.237
	D-PPIsite *	0.851	0.299	0.917	0.299	0.216	0.299	0.740	0.254
	PITHIA *	0.815	0.360	0.892	0.360	0.252	0.360	0.731	0.340
	GL-GAT	**0.810**	**0.477**	**0.884**	**0.477**	**0.361**	**0.4477**	**0.782**	**0.466**
Dset_186	SENSDeep *	0.776	0.355	0.866	0.355	0.223	0.358	0.685	0.338
	PITHIA *	0.815	0.360	0.892	0.360	0.252	0.360	0.731	0.340
	D-PPIsite *	0.778	0.386	0.864	0.386	0.250	0.386	0.710	0.364
	GL-GAT	**0.829**	**0.440**	**0.899**	**0.440**	**0.339**	**0.440**	**0.778**	**0.410**
Dset_72	DELPHI *	0.847	0.274	0.914	0.274	0.189	0.274	0.711	0.237
	SENSDeep *	0.788	0.258	0.832	**0.448**	0.224	0.327	0.714	0.264
	D-PPIsite *	0.851	0.299	0.917	0.299	0.216	0.299	0.740	0.254
	GL-GAT	**0.857**	**0.324**	**0.920**	0.324	**0.244**	**0.324**	**0.778**	**0.274**
Dset_70	ProNA2020 *	0.741	0.297	N/A	0.229	0.106	0.258	N/A	N/A
	EnsemPPIS *	0.732	0.375	N/A	0.532	0.277	0.440	0.719	0.405
	EGRET *†	0.715	0.358	N/A	0.591	0.270	0.438	0.719	0.405
	Seq-Insite *	0.781	0.447	0.864	0.447	0.311	0.447	**0.766**	**0.440**
	GL-GAT	**0.787**	**0.460**	**0.867**	**0.460**	**0.327**	**0.460**	0.740	0.426
Dset_60	DELPHI[a]	0.792	0.343	0.877	0.343	0.219	0.343	0.699	0.319
	SENSDeep *	0.768	0.344	N/A	0.370	0.199	0.339	0.686	0.371
	ProB-Site *	0.799	0.407	N/A	**0.612**	**0.368**	**0.517**	**0.844**	**0.467**
	Seq-Insite *	0.826	0.448	0.897	0.448	0.345	0.448	0.798	0.430
	MaSIF-site *†	0.780	0.370	N/A	0.561	0.326	0.446	0.775	0.439
	GraphPPIS[a†]	0.816	0.417	0.891	0.417	0.308	0.417	0.776	0.406
	RGN[a†]	0.824	0.443	0.896	0.443	0.338	0.443	0.783	0.427
	GL-GAT	**0.826**	**0.449**	**0.897**	0.449	0.345	0.449	0.778	0.420

标有[a] 的结果是从 Seq-Insite 获得的,标有[†] 的方法是基于结构的;其余的都是基于序列的. 标有 * 的摘自各自的论文,标有[f] 的结果来自 Seq-Insite[11] 提供的补充材料.

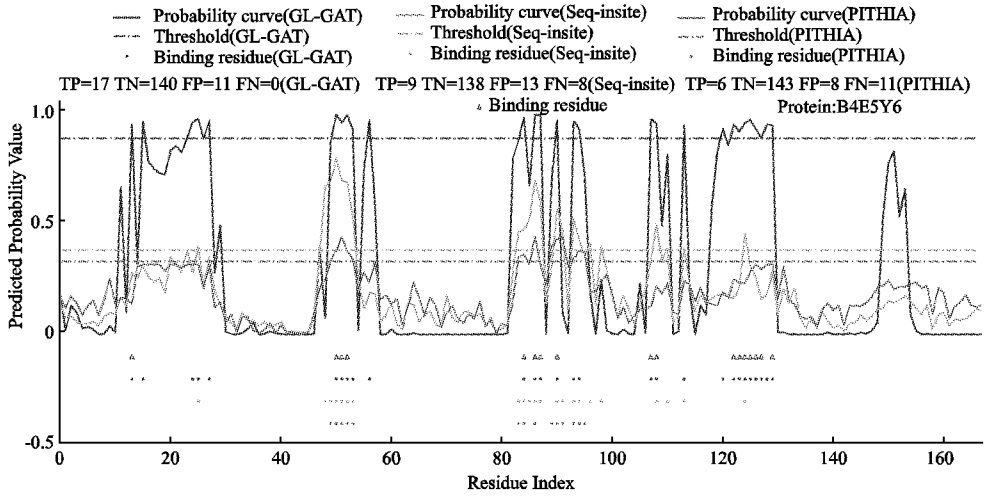

图 5　GL-GAT、Seq-insite 和 PITHIA 在 B4E6Y6 上的 PPI 预测性能

Fig. 5　PPI prediction performance of different predictors, i. e. , GL-GAT, Seq-insite and PITHIA on B4E6Y6

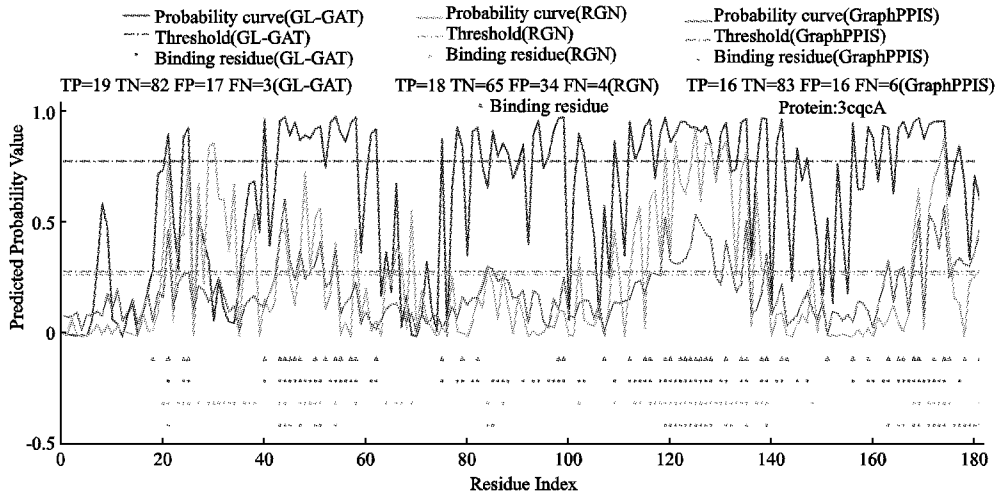

图 6 GL-GAT、RGN 和 GraphPPIS 在 3cqcA 上的 PPI 预测性能

Fig. 6 PPI prediction performance of different predictors, i. e. , GL-GAT, RGN and GraphPPIS on 3cqcA

高于 GraphPPIS 的 0.495,表明其整体预测性能更优. 案例分析结果表明,GL-GAT 在 PPI 位点预测中表现出较高的准确性和稳定性,尤其在真阳性预测方面具有优势.

3 总 结

本研究提出了一种结合全局与局部特征的图注意力网络深度学习框架 GL-GAT,以用于蛋白质相互作用位点的预测. GL-GAT 通过整合蛋白质的序列信息与结构特征,充分挖掘空间约束,提高了 PPI 位点预测性能. 实验结果表明,与现有的基于结构和基于序列的方法相比,GL-GAT 在多个基准数据集上的预测精度均有所提升.

未来研究可以从以下几个方向进一步改进本研究的方法:首先,引入更多蛋白质结构特征,如二面角和残基距离信息等,更全面地刻画蛋白质的空间特性. 其次,优化 GL-GAT 的模型架构,例如改进注意力机制,提升预测精度和泛化能力. 最后,拓展 GL-GAT 的应用范围,例如用于预测蛋白质-蛋白质结合残基对. 尽管 GL-GAT 仍有进一步优化的空间,但实验结果表明,其在蛋白质位点预测任务中具有竞争力,并有望成为该领域中高精度的预测工具之一.

References:

[1] Jones S, Thornton J M. Principles of protein-protein interactions [J]. Proceedings of the National Academy of Sciences, 1996, 93 (1): 13-20.

[2] Han J D, Bertin N, Hao T, et al. Evidence for dynamically organized modularity in the yeast protein-protein interaction network[J]. Nature, 2004, 430(6995): 88-93.

[3] Louche A, Salcedo S P, Bigot S. Protein-protein interactions: pull-down assays[J]. Methods in Molecular Biology, 2017, 1615: 247-255, doi: 10.1007/978-1-4939-7033-9_20.

[4] Lin J S, Lai E M. Protein-protein interactions: co-immunoprecipitation[J]. Methods in Molecular Biology, 2017, 1615: 211-219, doi: 10.1007/978-1-4939-7033-9_17.

[5] Douzi B. Protein-protein interactions: surface plasmon resonance [J]. Methods in Molecular Biology, 2017, 1615: 257-275, doi: 10.1007/978-1-4939-7033-9_21.

[6] Karimova G, Gauliard E, Davi M, et al. Protein-protein interaction: bacterial two hybrid [J]. Methods in Molecular Biology, 2024, 2715: 207-224, doi: 10.1007/978-1-0716-3445-5_13.

[7] Atmakuri K. Protein-protein interactions: cytology two-hybrid [J]. Methods in Molecular Biology, 2017, 1615: 189-197, doi: 10.1007/978-1-4939-7033-9_15.

[8] Hu J, Dong M, Tang Y X, et al. Improving protein-protein interaction site prediction using deep residual neural network[J]. Analytical Biochemistry, 2023, 670: 115132. doi: 10.1016/j. ab. 2023. 115132.

[9] Bradford J R, Westhead D R. Improved prediction of protein-protein binding sites using a support vector machines approach[J]. Bioinformatics, 2005, 21(8): 1487-1494.

[10] Chen C T, Peng H P, Jian J W, et al. Protein-protein interaction site predictions with three-dimensional probability distributions of interacting atoms on protein surfaces [J]. PLoS One, 2012, 7(6): e37706. doi: 10.1371/journal. pone. 0037706.

[11] Hosseini S, Golding G B, Ilie L. Seq-InSite: sequence supersedes structure for protein interaction site prediction[J]. Bioinformatics, 2024, 40(1), doi: 10.1093/bioinformatics/btad738.

[12] Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features[J]. Biopolymers, 1983, 22(12): 2577-2637.

[13] Rao R M, Liu J, Verkuil R, et al. MSA transformer[J]. bioRxiv 2021. 02. 12. 430858, doi: https://doi. org/10. 1101/2021. 02. 12. 430858.

[14] Elnaggar A, Heinzinger M, Dallago C, et al. ProtTrans: toward understanding the language of life through self-supervised learning [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(10): 7112-7127.

[15] 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, doi: 10. 1016/j. neucom. 2019. 05. 013.

[16] Li Y, Golding G B, Ilie L. DELPHI: accurate deep ensemble model for protein interaction sites prediction[J]. Bioinformatics, 2021, 37(7): 896-904.

[17] Mou M, Pan Z, Zhou Z, et al. A transformer-based ensemble framework for the prediction of protein-protein interaction sites[J]. Research (Wash D C), 2023, 6: 0240. doi: 10. 34133/research. 0240.

[18] 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.

[19] Kang Y, Xu Y, Wang X, et al. HN-PPISP: a hybrid network based on MLP-Mixer for protein-protein interaction site prediction[J]. Brief Bioinform, 2023, 24(1): bbac48, doi: 10. 1093/bib/bbac480.

[20] Khan S H, Tayara H, Chong K T. ProB-site: protein binding site prediction using local features[J]. Cells, 2022, 11(13): 2117, doi: 10. 3390/cells11132117.

[21] Yuan Q, Chen J, Zhao H, et al. Structure-aware protein-protein interaction site prediction using deep graph convolutional network [J]. Bioinformatics, 2021, 38(1): 125-132.

[22] Wang S, Chen W, Han P, et al. RGN: residue-based graph attention and convolutional network for protein-protein interaction site prediction[J]. Journal of Chemical Information and Modeling, 2022, 62(23): 5961-5974.

[23] Mahbub S, Bayzid M S. EGRET: edge aggregated graph attention networks and transfer learning improve protein-protein interaction site prediction[J]. Bioinformatics, 2022, 23(2): bbab578, doi: 10. 1093/bib/bbab578.

[24] Du Z, Su H, Wang W, et al. The trRosetta server for fast and accurate protein structure prediction[J]. Nature Protocols, 2021, 16(12): 5634-5651.

[25] Zhou X, Zheng W, Li Y, et al. I-TASSER-MTD: a deep-learning-based platform for multi-domain protein structure and function prediction[J]. Nature Protocols, 2022, 17(10): 2326-2353.

[26] Abramson J, Adler J, Dunger J, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3[J]. Nature, 2024, 630(8016): 493-500.

[27] 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.

[28] Singh G, Dhole K, Pai P, et al. SPRINGS: prediction of protein-protein interaction sites using artificial neural networks[J]. PeerJ PrePrints 2: e266v2, doi: 10. 7287/peerj. preprints. 266v2.

[29] 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.

[30] Yuan Q, Chen J, Zhao H, et al. Structure-aware protein-protein interaction site prediction using deep graph convolutional network [J]. Bioinformatics, 2021, 38(1): 125-132.

[31] Zhang J, Kurgan L. SCRIBER: accurate and partner type-specific prediction of protein-binding residues from proteins sequences[J]. Bioinformatics, 2019, 35(14): i343-i353.

[32] Altschul S F, Gish W, Miller W, et al. Basic local alignment search tool[J]. Journal of Molecular Biology, 1990, 215(3): 403-410.

[33] Lin Z, Akin H, Rao R, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model[J]. Science, 2023, 379(6637): 1123-1130.

[34] Altschul S F, Madden T L, Schäffer 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.