Prediction of protein–protein interaction using graph neural networks

Kanchan Jha, Sriparna Saha & Hiteshi Singh

摘要

蛋白质是执行几乎所有生物过程和细胞功能所需的基本生物大分子。蛋白质很少单独执行它们的任务，而是与周围环境中存在的其他蛋白质相互作用（称为蛋白质-蛋白质相互作用）以完成生物活动。蛋白质-蛋白质相互作用 (PPI) 的知识揭示了细胞行为及其功能。计算方法自动预测 PPI，并且在资源和时间方面比实验方法更便宜。到目前为止，大多数关于 PPI 的工作主要集中在序列信息上。在这里，我们使用图卷积网络 (GCN) 和图注意力网络 (GAT) 通过利用蛋白质的结构信息和序列特征来预测蛋白质之间的相互作用。我们根据蛋白质的 PDB 文件构建蛋白质图，其中包含原子的 3D 坐标。蛋白质图表示氨基酸网络，也称为残基接触网络，其中每个节点都是一个残基。如果两个节点在阈值距离内有一对原子（每个节点一个），则它们是连接的。为了提取节点/残基特征，我们使用蛋白质语言模型。语言模型的输入是蛋白质序列，输出是基础序列的每个氨基酸的特征向量。获得的结果证明了所提出方法的有效性，因为它优于以前的领先方法。

关键词：蛋白质相互作用; 图神经网络; 语言模型;

1 引言

蛋白质是由二十种标准氨基酸组成的有机大分子。它们负责执行生物体内几乎所有的生物过程和细胞功能。DNA 转录和复制、激素调节、新陈代谢、分子细胞信号转导和信号转导是涉及蛋白质相互作用的生命活动的一些例子。此外，PPI 的知识已被证明有助于新药的发现以及疾病的预防和诊断。然而，蛋白质很少单独执行它们的功能，而是与周围环境中的其他蛋白质相互作用并完成它们的任务。有几种高通量实验方法，例如酵母双杂交筛选 (Y2H)、串联亲和纯化 (TAP) 和质谱蛋白复合物鉴定 (MS-PCI)，用于鉴定蛋白质之间的相互作用。这些实验方法有助于创建不同物种的 PPI 数据集，但速度较慢。此外，由于实验环境和设备分辨率影响输出，这些方法收集的 PPI 数据具有很高的误报率和误报率。当与实验方法结合使用时，高通量计算方法可提高 PPI 预测的准确性和质量。

2 相关工作

迄今为止，已经有大量的工作采用传统的机器学习 (ML) 技术来解决计算生物学领域的几个问题，例如蛋白质分类、结构预测、蛋白质相互作用预测等。支持向量机 (SVM)，随机森林 (RF)、神经网络 (NN) 是广泛使用的 ML 算法来对蛋白质相互作用进行分类。这些算法的输入是手工设计的特征，这些特征主要来自使用物理化学特性、进化信息和氨基酸分布模式的基础蛋白质序列。例如，Shen 等[1]采用联合三元组特征提取方法，将 20 个氨基酸根据其性质分为 7 组，以核函数 SVM 作为学习算法。Guo 等人[2]使用自协方差 (AC) 方法对蛋白质序列进行编码，并使用 SVM 作为分类器来预测 PPI。AC 方法通过捕获序列中相隔一定距离的残基的物理化学性质来考虑相邻效应。Li 等人[3]开发了一种方法，该

方法结合了基于特定位置评分矩阵 (PSSM) 和物理化学特性的进化特征。判别向量机 (DVM) 用作分类器来预测蛋白质相互作用。Huang 等人[4]采用了蛋白质的全局编码表示和基于加权稀疏表示的分类器来对 PPI 进行分类。Li 等人[5] 提出了一种基于序列的方法来预测自相互作用蛋白质。首先，将已知的蛋白质序列转换为位置特异性评分矩阵（PSSM），然后使用低秩近似（LRA）从 PSSM 中获取特征向量。最后，提取的向量是旋转森林分类器的输入，它区分自相互作用和非自相互作用蛋白质。Zhou 等人[6] 引入了梯度提升决策树算法来预测 PPI。该算法基于几个蛋白质描述符，例如频率、组成、变换、分布和自协方差来编码蛋白质序列。除了基于序列的特征，我们还有其他信息来源来获取 PPI 模型的输入特征，包括基因融合、蛋白质结构、功能、基因表达谱等。Ding 和 Kihara[7] 回顾并分类了几种基于上面提到的输入特征。在所有蛋白质来源中，序列衍生特征最常用于预测 PPI。

研究人员最近使用最新的深度学习技术对 PPI 进行了大量工作，从而提高了模型的性能。这些技术允许他们使用高维和复杂的输入特征。Sun 等人[8] 使用堆叠自动编码器来获得基于序列的输入特征的紧凑且相关的表示。Du 等人[9] 提出了一种基于深度学习的 PPI 模型，称为 DeepPPI，它从蛋白质描述符中学习高级特征，优于传统的 ML 算法。Hashemifar 等人[10] 开发了一种称为 DPPI 的模型，该模型由三个模块组成：siamese-like CNN、随机投影和预测。DPPI 的输入是一对蛋白质的序列图，输出是表示它们是否相互作用的二进制值。Gonzalez-Lopez 等人[11]提出了一种 PPI 模型，其中使用 NLP 技术学习特征，例如从原始蛋白质序列嵌入和递归神经网络。他们表明，仅使用原始数据即可获得最先进的结果，而无需依赖特征工程来预测蛋白质相互作用。EnsDNN（集成深度神经网络）是 Zhang 等人[12]提出的一个非常复杂的 PPI 模型。它使用自协方差（AC）、局部描述符（LD）和多尺度连续和不连续局部描述符（MCD）方法来获得蛋白质序列的不同特征表示。然后将这些特征向量分别馈送到九个独立的神经网络，这些神经网络的层数和神经元数量不同，总共产生 27 个神经网络。然后将每个神经网络的输出馈送到具有两个隐藏层的多层感知器模型以预测标签。Jha 等人[13] 提出了一个深度多模式框架，它利用基于结构和本体的特征来预测蛋白质相互作用并优于现有工作。

大多数上述深度学习方法只考虑了基于序列的特征。基于蛋白质结构信息的 PPI 模型的探索明显较少。如今，最新的深度学习技术，如深度卷积神经网络 (CNN)，已被广泛使用，因为它们具有从结构数据中无缝提取特征的潜力。例如，本论文作者在之前的工作[14] 中，我们使用具有基于序列的特征的结构信息来预测蛋白质之间的相互作用。作者利用预训练的 ResNet50 模型从蛋白质的二维体积表示中提取结构特征。获得的结果表明，图像相关任务的技术可以扩展到蛋白质结构。但这些分析分子结构的方法存在计算成本高和可解释性高等问题。

3　本文方法

本文提出了一个结合基于图形的技术和语言模型（SeqVec 和 ProtBert）来预测 PPI 的框架。蛋白质的分子图具有代表构成蛋白质的氨基酸的节点。这里使用语言模型直接从蛋白质序列中获取每个残基（图中的节点）的特征。使用基于语言模型的特征向量的优点是它不需要领域知识来编码序列。获得的结果表明其在预测蛋白质相互作用方面的适用性。我们使用基于图的方法，例如图卷积网络 (GCN) 和图注意力网络 (GAT)，从结合结构和序列信息的蛋白质表示中学习特征。最后，将成对的蛋白质特征向量连接起来并馈送到具有两个隐藏层和一个输出层的分类器。以下是这项工作的主要贡献：

1. 使用带有残基的蛋白质图形表示作为节点。我们相信它将提高模型的性能，因为它涵盖了空间结构/低级属性。

2. 我们使用预训练的语言模型（SeqVec 和 ProtBert）来获取构建图中每个残基/节点的特征向量。我们表明该特征向量比其他特征（例如残基的物理化学性质、氨基酸的单热编码等）更有用。

3. 提出了两种基于图的架构：基于 GCN 和基于 GAT，以从集成空间结构和序列特征的蛋白质表示中学习特征。所获得的结果证明了所提出的方法优于现有工作的优越性。

所提出的预测 PPI 的方法包括三个模块：蛋白质图构建、特征提取和分类器以预测它们之间的相互作用。本节详细讨论每个模块和用于实现它们的深度学习技术。

3.1 利用图表示蛋白质

我们使用 PDB 文件构建了蛋白质的分子图，也称为氨基酸/残基接触网络。PDB 文件是包含结构信息（例如 3D 原子坐标）的文本文件。设 G(V, E) 是表示蛋白质的图，其中每个节点 (v ∈ V) 是残基，残基之间的边由 (e ∈ E) 描述。如果两个残基具有欧氏距离小于阈值距离的任何一对原子（每个残基一个），则两个残基是相连的。

蛋白质图中的每个节点都有一些与之相关的属性。这些节点特征主要是从蛋白质序列和结构中获得的。在这项工作中，使用两种预训练语言模型（基于 LSTM[15]和基于 BERT[16]）从蛋白质序列中提取节点特征。此外，我们还考虑了其他一些获取节点特征的方法，例如 20 种标准氨基酸的 one-hot 编码和残基的理化性质。在单热编码方法的情况下，每个节点表示为长度为 20 的向量。Meiler 等人[17]提供的氨基酸的七种物理化学性质被假设通过产生疏水力或氢来影响蛋白质之间的相互作用他们之间的纽带。因此，这些特征被用作图中每个节点的其他特征向量。在这些节点特征中（基于获得的结果），使用预训练的基于 LSTM 的语言模型（LM）提取节点特征的蛋白质图优于基于其他特征向量的方法。

图 1 描述了从 PDB 文件生成蛋白质图的所有步骤。第一步是获取每个蛋白质的 PDB 序列和分子图结构。

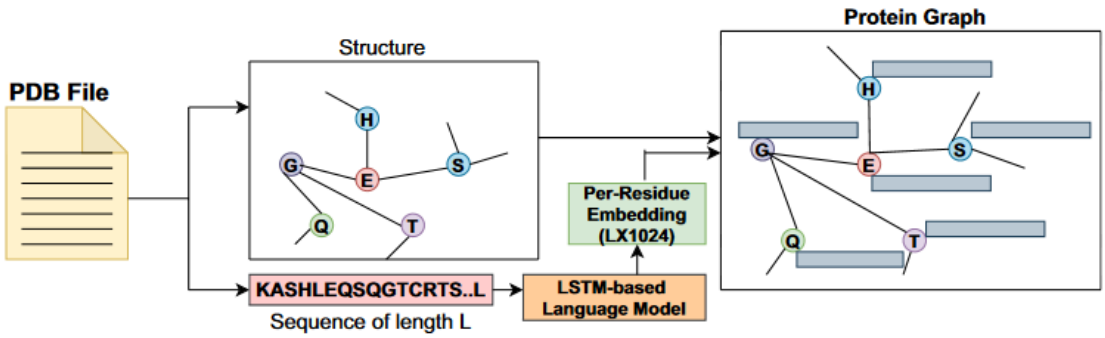

图 1: 从 PDB 文件中获取蛋白质并用图表示

3.2 图神经网络模块

GCN 用于从表示为图形的数据中学习特征。设 P = (V , E) 是蛋白质的图形表示，其中顶点数是 PDB 序列的长度（即 $|V| = L$ ）。每对节点之间的连接被编码在邻接矩阵中，$A \in R^{L,L}$ 。矩阵 $X \in R^{L,F}$ 包含给定图的所有节点的残基级特征（F：节点特征的维度）。GCN 的每一层都将邻接矩阵 (A) 和上一层的节点嵌入 ($H^{(l)} \in R^{L,F_i}$) 作为输入，并输出下一层的节点级嵌入 ($H^{(l+1)} \in R^{L,F_{l+1}}$)。

$$H^{(l+1)} = GC(H^{(l)}, A) \tag{1}$$

这里，$H^{(0)} = X \in R^{L,F}$, F_l 和 F_{l+1} 分别表示层 l 和 $l+1$ 的节点级嵌入的维度。方程式（1）的更具体的表达，定义的是：

$$H^{(l+1)} = ReLU(\hat{D}^{-0.5}\hat{A}\hat{D}^{-0.5}H^{(l)}W^{(l+1)}) \tag{2}$$

在这里，$\hat{A}$ 是添加了单位矩阵 $I_L \in R^{L,L}(\hat{A} = A + I_L)$ 的邻接矩阵。将单位矩阵添加到邻接矩阵强制图中的自循环，确保在卷积运算期间将节点自身的特征包含在和中。$\hat{D}$ 是对角节点度矩阵，计算如下：$\hat{D}_{ii} = \sum_{j=1}^{L} \hat{A}_{ij}$。它用于以对称方式（$\hat{D}^{-0.5}\hat{A}\hat{D}^{-0.5}$）对邻接矩阵进行归一化，以便我们在每个卷积层之后得到归一化的残差特征。在 GCN 层之后，每个残差特征向量被更新为图中相邻节点特征的加权和，包括残差自身的特征（等式 2）。$W^{(l+1)}$ 是可训练的权重矩阵。

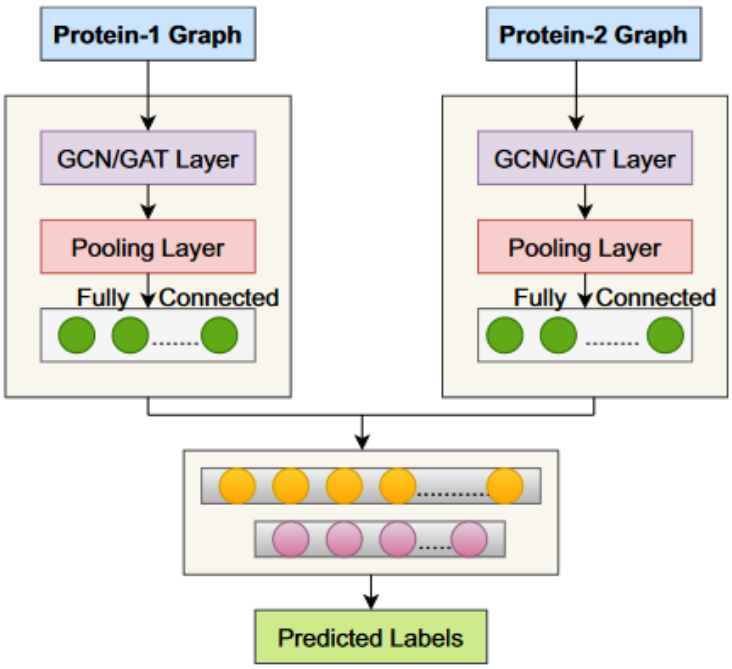

图 2: 本文提出的图神经网络框架

3.3 语言模型

蛋白质是氨基酸的长链，其中每个氨基酸（残基）都可以看作一个词，每个序列都可以看作一个句子。最近，研究人员开始使用自然语言处理 (NLP) 的语言模型 (LM) 来编码蛋白质序列。他们训练了许多蛋白质 LM 来生成蛋白质嵌入，可以用作深度学习模型的输入特征。这些输入特征已被证明比早期的方法（替代矩阵、捕获生物物理特性）更有价值，可以执行蛋白质亚细胞定位和结构预测等任务。在这项工作中，我们探索了两种蛋白质 LM 模型：SeqVec[15]（基于 LSTM）和 ProtBert[16]（基于 BERT）。使用这些语言模型获取特征向量的过程如图 3 所示。

SeqVec 是双向语言模型 ELMo（来自语言模型的嵌入) 的改编，并且依赖于上下文。SeqVec 模型与标准 ELMo 架构具有相同的配置，但有一些小的变化，例如将标记减少到 28 个（20：标准氨基酸，2：稀有氨基酸，3：未知或不明确的氨基酸，2：标记来表示序列的开始和结束，1：掩蔽）。为了处理增加的蛋白质序列长度，增加了展开步骤的数量。该语言模型由一个字符卷积 (CharCNN) 层和两个双向 LSTM 层组成。CharCNN 层将每个氨基酸映射到固定长度 (1024) 的潜在空间。该层不考虑来自相邻单词的信息。该层的输出是第一个 bi-LSTM 层的输入。每个 bi-LSTM 单元的维度为 1024（前 512 用于前向传播，最后 512 用于后向传播）。基于 ELMo 的 SeqVec 模型是在 UniRef50 上训练的，UniRef50 是一个包含 3300 万个蛋白质序列的数据集。对于长度为 L 的蛋白质序列，此预训练嵌入器生成大小

为 (3, L, 1024) 的嵌入。我们通过连接 3 层（1 个 CharCNN 和 2 个 bi-LSTM）的输出来获得这种大小的特征向量，每个层对每个单词或氨基酸都有 1024 维嵌入。通过对 3 层的嵌入求和来生成每个残基嵌入。

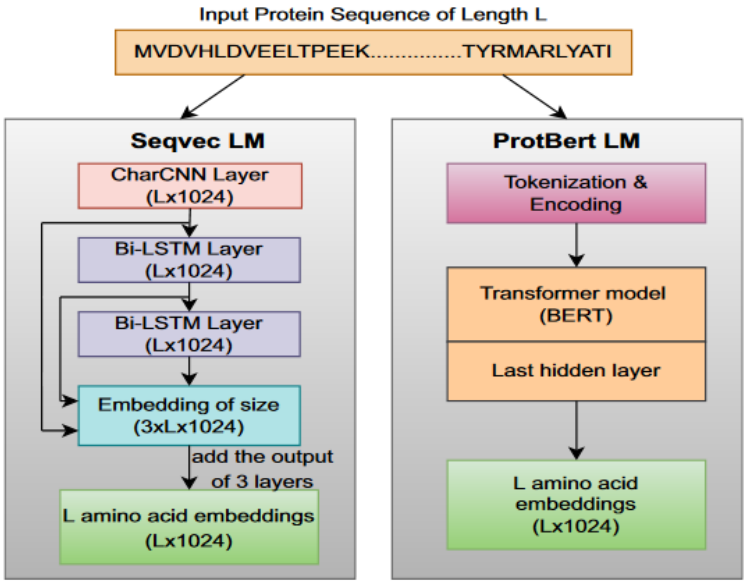

图 3: 使用语言模型从蛋白质序列中提取特征的图示

我们用来获取每个残基嵌入（蛋白质图中的节点特征）的第二个嵌入器是 ProtBert，它是在 BFD-100 数据集上训练的。BFD 拥有 21 亿条蛋白质序列，超过 3930 亿个氨基酸。ProtBert LM 使用 BERT 模型为每个残基生成描述性特征。BERT 是一种基于 Transformer 的深度学习模型，已广泛用于 NLP 中的迁移学习。对于长度为 L 的任何蛋白质序列，ProtBert LM 首先执行标记化并为每个标记添加位置编码。该层的输出然后通过堆叠的自注意层传递，这些层生成上下文感知嵌入。每个基于自我注意的编码层为蛋白质序列的每个残基输出长度为 1024 的嵌入。维度为 (L, 1024) 的注意力堆栈的最后一个隐藏状态被用作特征矩阵。由于与蛋白质序列相比，自然语言句子的长度要小得多，因此对蛋白质 LM 进行了一些更改，例如增加层数和展开步骤。

4 复现细节

4.1 与已有开源代码对比

本次复现工作代码主要参考原文提供的开源代码https://github.com/JhaKanchan15/PPI_GNN.git. 改进或创新点如下：

1. 在图卷积网络模型架构方面，增加了 2 到 3 层的卷积层进行实验，研究图卷积层数对于最后预测的影响。

2. 整合利用语言模型（SeqVec 和 BERT）对蛋白质序列进行特征提取的代码，作者提供的代码中该步骤需要手动输入蛋白质序列进行特征提取。在整合后可以直接对数据文件夹中所有蛋白质文件统一进行结构信息和特征的提取，并且直接转换为由 PyG 表示的图结构，转换后的图结构可直接使用图神经网络进行训练。

3. 优化代码结构，去除了一些冗余的代码，并且修改了一些主要的代码块，使得这些代码更加可读。

4.2 实验环境搭建

本试验所使用的主要环境列表如下：

python == 3.8.13

bio-embeddings == 0.2.2

biographs == 0.1

biopython == 1.80

networkx == 2.8.7

numpy == 1.23.4

scipy == 1.9.2

torch == 1.8.0

torch-cluster == 1.5.9

torch-geometric == 2.1.0.post1

torch-optimizer == 0.3.0

torch-scatter == 2.0.8

torch-sparse == 0.6.9

torch-spline-conv == 1.2.1

5 实验结果分析

我们使用开源深度学习框架 PyTorch 来构建预测 PPI 的图神经网络模型。为了构建我们在这项工作中探索的 GNN 模型，我们使用 PyTorch geometric（PyG）。该工作的分类器模块具有两个完全连接 (FC) 层，分别具有 256 和 64 个单元。使用学习率为 0.001 的 Adam 优化器来最小化训练模型时的 MSE 损失。在所提出的 PPI 模型中的每个全连接层之后，我们使用一个 dropout 层，dropout 率为 0.2。

在本次复现工作中，将每种蛋白质表示为一个图，并采用不同的图神经网络（GCN 和 GAT）对这些图结构的数据进行操作。两种图神经网络实现数据由表 1 给出。的可以看出，在测试的 6 个指标中，其中 GCN 有 5 项指标比 GAT 略好。

表 1: 基于 LSTM 特征提取的 GCN 与 GAT 网络对比

Model	ACC	PR	F1-Score	MCC	AUROC	AUPRC
GCN 1 layer	97.68	98.18	98.41	94.11	97.86	98.55
GAT 1 layer	97.54	98.33	98.31	93.79	97.68	98.50

影响深度学习模型性能的参数之一是层数。一般来说，假设 GNN 层数越多，可以从边和节点特征中收集到的信息就越多。然而，实际上，由于梯度消失和过度平滑，过多的层会导致性能下降。平滑被定义为图中节点之间的相似性。它被认为是 GNN 的本质，因为图中的节点相互交换消息。当堆叠多层时，它们之间的这种消息交互会使节点的表示相似并导致过度平滑。为了证明 GNN 层数对 PPI 任务的影响，我们在所提出的 PPI 框架中试验了多达 3 层的 GNN。实现数据在表 2 中。很明显，当我们将 GNN 层数从一层增加到两层时，PPI 模型（基于 GCN）在评估指标方面的性能是相当的。对于三层，性能指标的值低于具有一层或两层的指标。

表 2: 多个 GCN 层对比

Model	ACC	PR	F1-Score	MCC	AUROC	AUPRC
GCN 1 layers	97.68	98.18	98.41	94.11	97.86	98.55
GCN 2 layers	97.41	97.41	98.22	93.43	97.57	98.34
GCN 3 layers	96.59	97.01	97.15	92.22	95.98	97.61

6 总结与展望

在这项工作中，提出了一种结合图神经网络 (GNN) 和语言模型 (LM) 来预测蛋白质之间相互作用的方法。首先，我们从包含结构信息的 PDB 文件构建分子蛋白质图（氨基酸/残基作为节点）。然后我们使用 LM 从 PDB 序列生成每个残基嵌入，用作蛋白质图的节点特征。然后，基于 GNN 的模型从蛋白质的图形表示中提取特征（结合结构和序列信息）。最后，我们将每个蛋白质对的基于 GNN 的模型的输出连接起来，然后将生成的向量馈送到 PPI 分类器。这个分类器有两个全连接层和一个输出层。我们已经在流行的数据集上评估了方法的性能，获得的结果证明了它的有效性。并且该工作优于以前的方法，包括基于 PPI 网络的图形自动编码器模型。然而，基于 PPI 网络的模型在样本数量方面优于其基于蛋白质结构的对应模型，因为结构信息不适用于所有现有蛋白质。未来将探索其他基于深度学习的方法来从蛋白质表示（序列和结构）中学习特征，例如多尺度表示学习和用于学习 3D 蛋白质结构的内在-外在卷积和池化。

参考文献

[1] SHEN J, ZHANG J, LUO X, et al. Predicting protein–protein interactions based only on sequences information[J]. Proceedings of the National Academy of Sciences, 2007, 104(11): 4337-4341.

[2] GUO Y, YU L, WEN Z, et al. Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences[J]. Nucleic acids research, 2008, 36(9): 3025-3030.

[3] LI Z W, YOU Z H, CHEN X, et al. Highly accurate prediction of protein-protein interactions via incorporating evolutionary information and physicochemical characteristics[J]. International journal of molecular sciences, 2016, 17(9): 1396.

[4] HUANG Y A, YOU Z H, CHEN X, et al. Sequence-based prediction of protein-protein interactions using weighted sparse representation model combined with global encoding[J]. BMC bioinformatics, 2016, 17(1): 1-11.

[5] LI J Q, YOU Z H, LI X, et al. PSPEL: in silico prediction of self-interacting proteins from amino acids sequences using ensemble learning[J]. IEEE/ACM transactions on computational biology and bioinformatics, 2017, 14(5): 1165-1172.

[6] ZHOU C, YU H, DING Y, et al. Multi-scale encoding of amino acid sequences for predicting protein interactions using gradient boosting decision tree[J]. PLoS One, 2017, 12(8): e0181426.

[7] DING Z, KIHARA D. Computational methods for predicting protein-protein interactions using various protein features[J]. Current protocols in protein science, 2018, 93(1): e62.

[8] SUN T, ZHOU B, LAI L, et al. Sequence-based prediction of protein protein interaction using a deep-learning algorithm[J]. BMC bioinformatics, 2017, 18(1): 1-8.

[9] DU X, SUN S, HU C, et al. DeepPPI: boosting prediction of protein–protein interactions with deep neural networks[J]. Journal of chemical information and modeling, 2017, 57(6): 1499-1510.

[10] HASHEMIFAR S, NEYSHABUR B, KHAN A A, et al. Predicting protein–protein interactions through sequence-based deep learning[J]. Bioinformatics, 2018, 34(17): i802-i810.

[11] GONZALEZ-LOPEZ F, MORALES-CORDOVILLA J A, VILLEGAS-MORCILLO A, et al. End-to-end prediction of protein-protein interaction based on embedding and recurrent neural networks[C]//2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). 2018: 2344-2350.

[12] ZHANG L, YU G, XIA D, et al. Protein–protein interactions prediction based on ensemble deep neural networks[J]. Neurocomputing, 2019, 324: 10-19.

[13] JHA K, SAHA S, SAHA S. Prediction of Protein-Protein Interactions using Deep Multi-Modal Representations[C]//2021 International Joint Conference on Neural Networks (IJCNN). 2021: 1-8.

[14] JHA K, SAHA S. Amalgamation of 3D structure and sequence information for protein–protein interaction prediction[J]. Scientific Reports, 2020, 10(1): 1-14.

[15] HEINZINGER M, ELNAGGAR A, WANG Y, et al. Modeling aspects of the language of life through transfer-learning protein sequences[J]. BMC bioinformatics, 2019, 20(1): 1-17.

[16] ELNAGGAR A, HEINZINGER M, DALLAGO C, et al. ProtTrans: towards cracking the language of Life's code through self-supervised deep learning and high performance computing[J]. arXiv preprint arXiv:2007.06225, 2020.

[17] MEILER J, MÜLLER M, ZEIDLER A, et al. Generation and evaluation of dimension-reduced amino acid parameter representations by artificial neural networks[J]. Molecular modeling annual, 2001, 7(9): 360-369.