doi：10.13413/j.cnki.jdxblxb.2023227

基于特征图网络和多种生物信息预测关键蛋白质的深度学习框架

刘桂霞，曹心恬，赵 贺
（吉林大学 计算机科学与技术学院，符号计算与知识工程教育部重点实验室，长春 130012）

摘要：针对生物实验识别关键蛋白质费时费力，使用计算方法预测关键蛋白质无法有效整合生物信息的问题，提出一个深度学习框架. 首先利用网络拓扑结构、基因表达数据和 GO (gene ontology)注释数据构建加权蛋白质相互作用网络；然后分别使用特征图网络和双向长短期记忆细胞从亚细胞定位数据、蛋白质复合物数据和基因表达数据中提取特征向量；最后将这些特征向量输入到任务学习层预测关键蛋白质. 实验结果表明，相比于现有的计算方法，该方法预测性能更好.

关键词：关键蛋白质；特征图网络；亚细胞定位；基因表达；GO 注释；蛋白质复合物
中图分类号：TP391 **文献标志码**：A **文章编号**：1671-5489(2024)03-0593-13

Deep Learning Framework for Predicting Essential Proteins Based on Feature Graph Network and Multiple Biological Information

LIU Guixia, CAO Xintian, ZHAO He
(*Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, College of Computer Science and Technology, Jilin University, Changchun* 130012, *China*)

Abstract: Aiming at the problem that identifying essential proteins in biological experiments was time-consuming and laborious, and using computational methods to predict essential proteins could not effectively integrate biological information, we proposed a deep learning framework. Firstly, a weighted protein interaction network was constructed by using network topology structure, gene expression data and gene ontology (GO) annotated data. Secondly, feature vectors were extracted from subcellular localization data, protein complex data and gene expression data by using feature graph network and bi-directional long short-term memory cells, respectively. Finally, these feature vectors were input into the task learning layer to predict essential proteins. The experimental results show that, compared with existing computational methods, the proposed method has better predictive performance.

Keywords: essential protein; feature graph network; subcellular localization; gene expression; GO annotation; protein complex

关键蛋白质对许多生命过程至关重要，预测关键蛋白质对理解生物功能、识别致病基因和药物发

收稿日期：2023-06-06.

第一作者简介：刘桂霞(1963—)，女，汉族，博士，教授，博士生导师，从事机器学习和计算生物学的研究，E-mail：liugx@jlu.edu.cn.

基金项目：国家自然科学基金(批准号：62372208；61772226)和吉林省科技发展规划重点项目(批准号：20210204133YY).

现有重大意义[1]. 目前，鉴定关键蛋白质的生物方法主要是利用RNA(ribonucleic acid)干扰[2]、条件性敲除[3]和单基因敲除[4]等实验方式，但这些方法既费时又费力. 随着高通量技术的快速发展，积累了大量的生物数据，为计算方法提供了技术支持.

目前的计算方法主要分为以下两类：基于复杂网络的方法和基于机器学习的方法. 基于复杂网络的方法起源于Jeong等[5]提出的中心性原则，即一个蛋白质在蛋白质相互作用(protein-protein interaction，PPI)网络中的拓扑连接越紧密，它就越关键. 受此启发，研究人员使用PPI网络的拓扑结构识别关键蛋白质，如度中心性(degree centrality，DC)、介数中心性(betweenness centrality，BC)、局部平均连通性(local average connectivity，LAC)[6]和局部相互作用密度(local interaction density，LID)[7]等. 但这些方法忽略了关键蛋白质固有的生物学特征影响，因此研究人员考虑结合不同的生物学数据发现关键蛋白质. PeC[8]和WDC[9]方法将PPI网络与基因表达数据相结合，以提高关键蛋白质的识别准确率. Lei等[10]提出了基于亚细胞定位数据、RNA-Seq数据和GO(gene ontology)注释数据的关键蛋白预测方法RSG. 文献[11]提出了一种基于PPI网络的局部密度、BC以及蛋白质复合物中的度中心性(in-degree centrality of complex，IDC)的线性组合新方法LBCC.

目前，已提出了许多预测关键蛋白质的传统机器学习方法和深度学习框架. 机器学习方法遵循以下步骤预测关键蛋白质：选择代表性特征，构建训练集和测试集，选择合适的算法或框架，最后评估性能. 朴素Bayes、随机森林、支持向量机(support vector machine，SVM)、Adaboost、决策树和逻辑回归属于识别关键蛋白质的传统机器学习方法. DeepEP是由Zeng等[12]提出的一个深度学习框架，利用多尺度卷积神经网络(convolutional neural network，CNN)从基因表达数据中提取生物特征，采用node2vec[13]从PPI网络中学习拓扑特征，然后将它们拼接预测关键蛋白质. DeepEP还采用了一种抽样策略减轻不平衡学习的影响. Zeng等[14]考虑到基因表达数据的顺序属性，还提出了一个深度学习的框架，它利用双向长短期记忆细胞(bi-directional long short-term memory，BiLSTM)[15]捕捉其特征，也采用node2vec从PPI网络中学习拓扑特征，但增加了对亚细胞定位数据的利用. Yue等[16]提出了一种深度学习方法，将node2vec提取的PPI网络拓扑特征、亚细胞定位数据和基因表达数据相结合，该方法将深度可分离卷积的概念应用于基因表达数据，以在不同的实验设置中随时间提取特征. DeepCellEss[17]是一种基于序列的可解释性深度学习框架，利用CNN和BiLSTM从蛋白质序列中学习潜在信息，使用多头自注意力机制提供残差级的模型可解释性，用于细胞系特异性关键蛋白质预测. CTF[18]基于h-quasi-cliques和uv-triangle图等边缘特征以及多源信息识别关键蛋白质.

虽然上述方法效果良好，但仍存在一些缺点：1) 实验数据问题，实验方法得到的PPI网络数据中存在假阴性和假阳性的问题；2) 实验方法问题，基于复杂网络的方法通过设计一个函数计算中心性指标，用其评估一个蛋白质的重要性，但研究人员很难设计一个好的计算函数，因为需要大量的先验知识. 计算函数只产生标量，易受PPI网络中噪声的干扰，并且PPI网络非常复杂，标量不能充分描述PPI网络的拓扑结构信息. 基于传统机器学习的方法依赖于人工选择特征. 特征的代表性越强，这些方法的预测效果越好. 在现有的深度学习框架中，PPI网络中蛋白质之间的拓扑属性没有得到充分利用，主要体现在node2vec提取的特征中.

为解决上述问题，进一步提高关键蛋白质的预测精度，本文提出一个基于特征图网络(feature graph network，FGN)和多种生物信息预测关键蛋白质的深度学习框架. 在酵母菌数据集BioGRID和DIP上的对比实验结果表明，本文方法优于目前主流的基于复杂网络方法和机器学习方法. 消融实验结果表明，FGN[19]和亚细胞定位数据的使用显著提高了关键蛋白质的预测性能，蛋白质复合物数据也有助于提高预测效果. 通过给PPI网络加权，能降低数据中噪声的影响，丰富PPI网络的边缘信息，从而进一步提高本文提出的深度学习框架的性能.

1 算法设计

本文提出一个预测关键蛋白质的深度学习框架，其主要思想如下.

1) 基于PPI网络拓扑特征、基因表达数据和GO注释数据构建加权PPI网络，以减小PPI网络中

噪声数据的影响，并丰富 PPI 网络的边缘信息.

2）本文提出的网络结构如图 1 所示，包含两部分. 上半部分先从亚细胞定位数据和蛋白质复合物数据中分别提取一个初始特征向量，再分别接入一个输出维度为 1 024 的全连接层和激活层，以进一步提取特征. 将得到的两个 1 024 维的特征向量拼接，形成一个 1 024×2 的特征向量，输入到两层的 FGN 中，以学习更丰富的节点表示. 从每个 FGN 层输出后，再先后输入到批量归一化(batch normalization，BN)层、激活层和 Dropout(随机失活)层，以防止过拟合，从而提高网络的泛化能力. FGN 通过将边缘信息(PPI 网络中蛋白质之间相连的边)编码到特征邻接矩阵中，从而能更好地保存利用边缘信息.

3）网络结构的下半部分利用 BiLSTM 从基因表达数据中提取特征向量，以捕捉蛋白质随时间变化的表达状态. 本文使用的基因表达数据集有 3 个连续的代谢周期，一个周期中有 12 个时间点，其中每个蛋白质(例如 G_i)对应一个 36 维的基因表达数据，T_j 表示第 j 个时间点.

4）拼接上述所有的特征向量，并将其输入到任务分类层(即一层全连接层)预测关键蛋白质，是一个二分类任务. 由于关键蛋白质的比例较小，因此本文采用抽样方法缓解不平衡学习的问题.

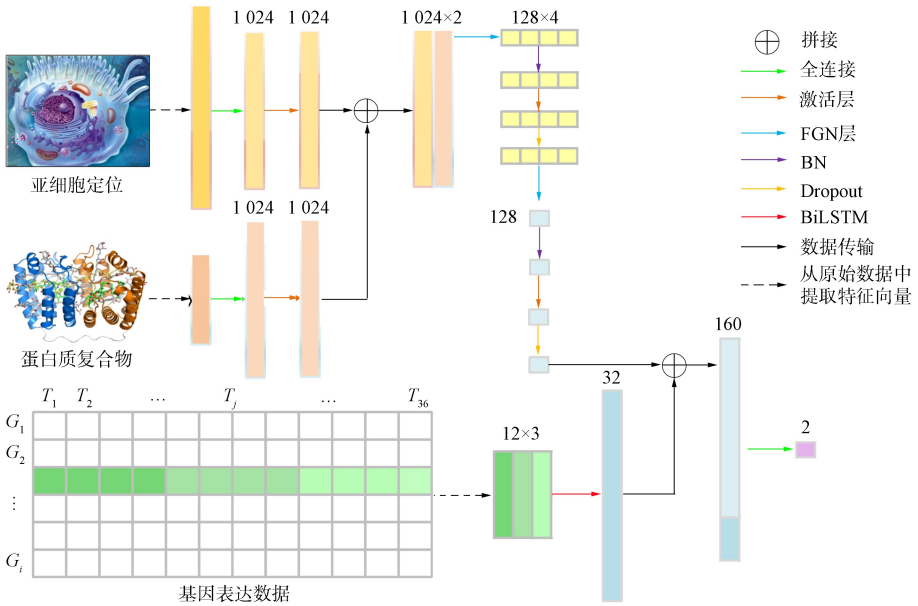

图 1　网络结构

Fig. 1　Network structure

1.1　构建加权的 PPI 网络

PPI 网络通常被抽象为无向无权图 $G(V,E)$，其中 V 表示蛋白质的集合，E 表示蛋白质-蛋白质相互作用的集合. 生物实验获得的 PPI 网络中，蛋白质之间的相互作用通常被认为是等价的，并被赋予相同的权重. 但现有的 PPI 网络数据集中通常存在噪声，对预测性能产生不利影响. 根据蛋白质相互作用的强度为其分配不同的权重，能减少噪声数据的影响，并丰富 PPI 网络的边缘信息，从而有利于发现关键蛋白质. 本文采用 Pearson 相关系数(Pearson correlation coefficient，PCC)[20]、GO 语义相似度(GO semantic similarity，GSS)[21]和边缘聚集系数(edge clustering coefficient，ECC)[22]衡量蛋白质之间的相互作用强度.

关键蛋白质常会相互作用，因此，本文通过计算蛋白质对应的基因表达数据的 PCC 评估两个蛋白质相互作用的强度. 蛋白质 i 和 j 的 PCC 计算公式如下：

$$\mathrm{PCC}_{ij}=\frac{\sum_{k=1}^{n}(x_k-\bar{\boldsymbol{v}}_i)(y_k-\bar{\boldsymbol{v}}_j)}{\sqrt{\sum_{k=1}^{n}(x_k-\bar{\boldsymbol{v}}_i)^2}\sqrt{\sum_{k=1}^{n}(y_k-\bar{\boldsymbol{v}}_j)^2}},\tag{1}$$

其中 $\boldsymbol{v}_i=(x_1,x_2,\cdots,x_n)$ 和 $\boldsymbol{v}_j=(y_1,y_2,\cdots,y_n)$ 是蛋白质 i 和 j 在 n 个时间点上的基因表达值，$\bar{\boldsymbol{v}}_i$ 和 $\bar{\boldsymbol{v}}_j$

分别是其平均值. 由于PCC的取值范围是[-1,1],所以本文用(PCC+1)/2替代上述定义的PCC,使其取值范围为[0,1]. PCC值越大,两个蛋白质的相互作用越强.

一般使用GSS评估蛋白质对的功能相似性. 由于GO术语代表了基因的功能特征,所以蛋白质 i 和 j 共同的GO术语越多,它们的功能越相似. 蛋白质 i 和 j 的GSS计算公式如下:

$$\mathrm{GSS}_{ij}=\begin{cases}\dfrac{|\mathrm{GO}_i\cap\mathrm{GO}_j|^2}{|\mathrm{GO}_i|\times|\mathrm{GO}_j|}, & |\mathrm{GO}_i|>0,\quad |\mathrm{GO}_j|>0,\\ 0, & \text{其他},\end{cases}\tag{2}$$

其中 GO_i 和 GO_j 分别表示蛋白质 i 和 j 的GO术语集合.

本文使用ECC衡量两个蛋白质在网络结构上的相似程度. ECC值越高,说明这两个蛋白质的拓扑结构越相似,它们之间的相互作用越强. 蛋白质 i 和 j 的ECC计算公式如下:

$$\mathrm{ECC}_{ij}=\begin{cases}\dfrac{|N_i\cap N_j|}{\min\{|N_i|-1,|N_j|-1\}}, & |N_i|>1,\quad |N_j|>1,\\ 0, & \text{其他},\end{cases}\tag{3}$$

其中 N_i 和 N_j 分别表示蛋白质 i 和 j 的直接邻居节点集,$N_i\cap N_j$ 表示蛋白质 i 和 j 的公共邻居集合. 式(3)表明,两个蛋白质的公共邻居越多,ECC值越高,表示两个蛋白质的拓扑结构越相似.

最后,利用

$$W_{ij}=\mathrm{ECC}_{ij}\times(\mathrm{GSS}_{ij}+\mathrm{PCC}_{ij})\tag{4}$$

对蛋白质 i 和 j 的边进行加权[23].

1.2 亚细胞定位特征向量

在机器学习方法中,亚细胞定位数据经常用于构建特征向量. 文献[14]的方法使用11种亚细胞定位构建特征向量. 但如果在本文的PPI网络中使用上述方法,某些蛋白质并不存在于11种亚细胞定位中的任何一种,因此它们的特征向量为0. 为使更多的特征向量非0并充分利用亚细胞定位信息,本文利用COMPARTMENTS数据库[24]中集成通道所提供的所有亚细胞定位.

对于蛋白质 i 和亚细胞定位 L,COMPARTMENTS数据库为亚细胞定位-蛋白质对 L-i 提供了最终的置信度 $w_{L\text{-}i}$. 置信度可表明亚细胞定位类型和来源的可靠性. 置信度越高,亚细胞定位-蛋白质对的关系越可靠. 一种蛋白质会出现在多种不同的亚细胞定位中,一种亚细胞定位也会包含多种不同的蛋白质. 亚细胞定位 L 的分数 N_L 定义为包含的蛋白质数量. 对 N_L 进行归一化处理,使其取值范围为[0,1],归一化处理表示为

$$N_L=\frac{N_L-\min}{\max-\min},\tag{5}$$

其中min和max分别表示所有亚细胞定位分数中的最小值和最大值. 本文将蛋白质 i 的亚细胞定位信息编码成一个一维向量 $\boldsymbol{Y}_i=(y_{1\text{-}i},y_{2\text{-}i},\cdots,y_{L\text{-}i},\cdots)$,其中 $y_{L\text{-}i}$ 是 L-i 对应的最终分数,计算公式如下:

$$y_{L\text{-}i}=N_L\times w_{L\text{-}i},\tag{6}$$

1.3 蛋白质复合物特征向量

蛋白质在同一时间和地点共同发挥作用形成蛋白质复合物. 在众多复合物中发现的蛋白质可能是关键的,因此蛋白质复合物有利于预测关键蛋白质. 在本文的深度学习框架中,利用蛋白质复合物构建特征向量,以预测关键蛋白质.

将蛋白质 i 的蛋白复合物信息编码成一个一维向量 $\boldsymbol{Z}_i=(z_1,z_2,\cdots,z_g,\cdots)$,其中 g 表示一种类型的蛋白复合物. 若蛋白质 i 属于蛋白质复合物 g,则 $z_g=1$,否则 $z_g=0$.

本文将每个蛋白质对应的亚细胞定位特征向量和蛋白质复合物特征向量分别接入一个输出维度为1 024的全连接层和激活层,以进一步提取特征. 将得到的两个1 024维的特征向量拼接,形成一个1 024×2维的特征向量. 此时一个蛋白质对应一个1 024×2维的特征向量,将其输入到两层的FGN中,以学习更丰富的节点表示.

1.4　特征图网络

文献[19]提出了特征图网络，它能直接对特征“交互”进行建模. 本文给 PPI 网络中每个节点都赋予一个初始特征向量，由于特征向量通过节点的拓扑结构进行交互，因此对特征“交互”进行建模能更好地保存并利用节点的边缘信息，从而提高关键蛋白质的预测性能.

1.4.1　特征图

在无向图 $G(V,E)$中，每个节点 v 都对应一个特征向量 $\boldsymbol{X}=(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_i,\cdots,\boldsymbol{x}_F)^{\mathrm{T}}$，其中 F 是维度. 特征图 Q 对应于 v，将 v 的特征向量的分量作为节点特征，即 $\boldsymbol{x}_1,\cdots,\boldsymbol{x}_i,\cdots,\boldsymbol{x}_F$ 分别是 Q 中的节点 $1,\cdots,i,\cdots,F$的特征向量，并且表示 Q 共有 F 个节点. Q 可以描述为 $G_{Q^F}=(V^F,E^F)$，其中每个节点 $v_i^F\in V^F$ 对应于特征向量 $\boldsymbol{x}_i$，如图 2 所示. 如果 $\boldsymbol{X}$ 是一维向量，则 $\boldsymbol{x}_i$ 就是标量；如果 $\boldsymbol{X}$ 是多通道向量，则 $\boldsymbol{x}_i$ 就是一维向量.

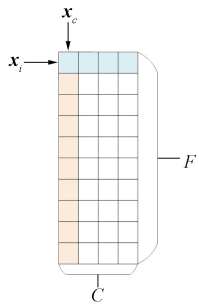

图 2　无向图中节点的特征向量

Fig. 2　Feature vector of node in undirected graph

PPI 网络中的每个节点都对应一个特征图. 本文将 1 024×2 维的特征向量输入到两层 FGN 中，则对于 PPI 网络中的特征图，其节点数即为 1 024，特征维度即为 2.

1.4.2　特征邻接矩阵

特征邻接矩阵 $\boldsymbol{A}_c^F$ 对应于节点 v 的特征图 G_{Q^F}，其对应的特征向量为 $\boldsymbol{X}=(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_i,\cdots,\boldsymbol{x}_F)^{\mathrm{T}}$. 节点 v与其直接邻居 $N(v)$之间的边表明他们的特征之间存在关联性. FGN 将 v 和 $N(v)$上的相关性建模为特征邻接矩阵，并对多个通道中的每个通道 c 独立建模，$c=1,2,\cdots,C$. 特征邻接矩阵 $\boldsymbol{A}_c^F(\boldsymbol{x})$ 定义为

$$\boldsymbol{A}_c^F(\boldsymbol{x})\triangleq \mathrm{sgnroot}(E_{y\sim N(v)}[w_y(\boldsymbol{x}_{[:,c]}\boldsymbol{y}_{[:,c]}^{\mathrm{T}}+\boldsymbol{y}_{[:,c]}\boldsymbol{x}_{[:,c]}^{\mathrm{T}})]), \tag{7}$$

$$E_{y\sim N(v)}[w_y(\boldsymbol{x}_{[:,c]}\boldsymbol{y}_{[:,c]}^{\mathrm{T}}+\boldsymbol{y}_{[:,c]}\boldsymbol{x}_{[:,c]}^{\mathrm{T}})]=\frac{\sum\limits_{y\in N(v)} w_y(\boldsymbol{x}_{[:,c]}\boldsymbol{y}_{[:,c]}^{\mathrm{T}}+\boldsymbol{y}_{[:,c]}\boldsymbol{x}_{[:,c]}^{\mathrm{T}})}{|N(v)|}, \tag{8}$$

其中 $\mathrm{sgnroot}(x)=\mathrm{sign}(x)\sqrt{|x|}$，$y$ 为 v 的直接邻居，$\boldsymbol{y}$ 为对应的特征向量，$\boldsymbol{x}$ 为节点 v 对应的特征向量，$\boldsymbol{x}_{[:,c]}$为 $\boldsymbol{x}_c$(见图 2)，w_y 为对应边的权重. $\boldsymbol{A}_c^F$ 是通过式(7)从邻域样本动态独立构建的，将 PPI 网络中的连通信息(蛋白质之间的连接以及权重)编码到特征邻接矩阵中. 对于每个节点 v，将生成 C 个大小为 $F\times F$ 的特征邻接矩阵.

1.4.3　特征图层

特征图层会改变 G_{Q^F} 中节点的数量，所以需要基于转变后的邻居重新计算 $\boldsymbol{A}_c^F(\boldsymbol{x})$. 特征图网络第 l 层的定义和转换公式如下：

$$\boldsymbol{A}_{(l)}^F(\boldsymbol{x})\triangleq \boldsymbol{A}^F(\boldsymbol{x}_{(l)},\boldsymbol{y}_{(l)}),\quad \forall y\in N(v), \tag{9}$$

$$\boldsymbol{x}_{(l+1)}^F=\sigma(\boldsymbol{W}^F\cdot\hat{\boldsymbol{A}}_{(l)}^F(\boldsymbol{x})\cdot\boldsymbol{x}_{(l)}^F), \tag{10}$$

$$\boldsymbol{y}_{(l+1)}^F=\sigma(\boldsymbol{W}^F\cdot\hat{\boldsymbol{A}}_{(l)}^F(\boldsymbol{x})\cdot\boldsymbol{y}_{(l)}^F), \tag{11}$$

其中 $\boldsymbol{W}^F\in\mathbb{R}^{F_{(l+1)}\times F_{(l)}}$ 为可学习参数，$\sigma(\cdot)$为非线性激活函数，$\hat{\boldsymbol{A}}_{(l)}^F(\boldsymbol{x})\in\mathbb{R}^{F_{(l)}\times F_{(l)}}$ 为 $\boldsymbol{A}_{(l)}^F(\boldsymbol{x})$的归一化. 在式(9)～式(11)中省略了通道 c，每个通道独立转换.

上述内容每个节点 v 只考虑了直接邻居，为使节点 v 考虑的节点更丰富，使特征图网络能更好地学习节点表示，本文将引入 top-k intimacy.

1.4.4　top-k intimacy

对于图亲密度矩阵 $\boldsymbol{S}\in\mathbb{R}^{|V|\times|V|}$，其中$|V|$表示无向图的节点个数，$S(i,j)$表示节点 v_i 和 v_j 之间的亲密度[25]. 目前，存在不同的指标衡量图中节点之间的亲密度，如 Pagerank 算法，Adamic/Adar，Katz 等. Jaccard 系数[26]是一种衡量两个集合之间相似度的方法，本文定义基于 Jaccard 系数的图亲密

度矩阵 $\boldsymbol{S}$. 节点 v_i 和 v_j 之间的 Jaccard 系数，即 $S(i,j)$定义为

$$S(i,j)=\frac{|N_i \cap N_j|}{|N_i \cup N_j|}, \tag{12}$$

其中 N_i 和 N_j 分别表示节点 v_i 和 v_j 的邻居节点集合. 对于节点 v_i，定义亲密节点集合为 $\Gamma(v_i)=\{v_j \mid v_j \in V \backslash \{v_i\} \wedge S(i,j) \geqslant \theta_i\}$，对 $S(i,:)$从大到小排序，第 k 个值为 θ_i. 对于节点 v，用 $\Gamma(v) \bigcup N(v)$ 代替 1.4.2 中的 $N(v)$. 本文中 $k=10$，表示特征图网络将从节点 v 的直接邻居和前 10 个亲密节点中学习节点 v 的特征向量，从而预测关键蛋白质.

1.5 基因表达特征向量

蛋白质是基因表达的产物，文献[8]研究表明，利用基因表达数据可提高识别关键蛋白质的准确率. 本文采用的基因表达数据存在时间上的顺序关系，考虑到 BiLSTM 应用在序列数据中效果较好，所以采用 BiLSTM 从基因表达数据中捕捉蛋白质随时间变化的表达状态，得到基因表达特征向量.

BiLSTM 网络结构模型由两个独立的 LSTM 构成，基因表达数据分别以正序和逆序输入到两个 LSTM 网络中进行特征提取，输出向量拼接后形成最终的特征向量. 因此，BiLSTM 模型中每个时间步的输出都取决于过去和未来的数据.

2 实验结果与分析

2.1 实验数据

实验采用酵母菌数据集，包括 PPI 网络数据集、关键蛋白质数据集、基因表达数据集、亚细胞定位数据集、GO 注释数据集和蛋白质复合物数据集.

为评估本文方法在预测关键蛋白质方面的性能，选择两种不同的 PPI 网络：BioGRID 和 DIP. 关键蛋白质数据是从 MIPS,SGD,DEG 和 SGDP 数据库中收集的，预处理后包含 1 285 种关键蛋白质.

基因表达数据集来自数据库 GEO(登录号：GSE3431). 该数据集包含了 7 134 个基因表达数据，每个基因表达数据有 3 个连续的代谢周期，每个周期有 12 个时间点. 亚细胞定位数据和 GO 注释数据是从数据库 COMPARTMENT 的集成通道中提取的. 蛋白质复合物数据是从 MIPS,SGD,ALOY 和 CYC2008 数据集中收集的，经过预处理后共包含 745 种蛋白质复合物. 数据集 BioGRID 和 DIP 的详细信息列于表 1. 对于没有基因表达数据的蛋白质，本文将采用基因表达数据的均值作为其基因表达数据.

表 1 数据集 BioGRID 和 DIP 的信息

Table 1 Information of BioGRID and DIP datasets

酵母菌数据集	蛋白质	相互作用边	关键蛋白质	拥有基因表达数据的蛋白质
BioGRID	5 686	104 000	1 200	5 320
DIP	4 850	21 592	1 146	4 780

2.2 不平衡学习

现有数据集中存在样本不平衡的问题，例如：数据集 BioGRID 中非关键蛋白质和关键蛋白质的数量之比为 3.73∶1，数据集 DIP 中非关键蛋白质和关键蛋白质的数量之比为 3.23∶1. 为减小样本不平衡的影响，本文将采用文献[12]提出的抽样方法缓解训练过程中的不平衡学习问题.

对于原始数据集，首先进行随机打乱，然后将关键蛋白质的 20%和非关键蛋白质的 20%组合在一起作为独立的测试集，余下的作为训练集. 用 M 和 N 分别表示训练集中关键蛋白质的数量和非关键蛋白质的数量. 在每轮训练中，先从训练集的非关键蛋白质中采样 M 个蛋白质，然后将其与训练集中的所有关键蛋白质(M 个)组合为一个集合训练网络，该集合共有 $2M$ 个蛋白质，从而可保证训练过程中结果不偏向任何类别(关键蛋白质类和非关键蛋白质类).

2.3 评价指标

对于不平衡学习，通过比较 AP(average precision)值和 AUC(area under curve)值评估本文提出的方法与其他方法的性能，这两个值分别表示 PR(precision-recall)曲线下方的面积和 ROC(receiver

operating characteristic)曲线下方的面积. PR 曲线和 ROC 曲线都是在各种阈值设置下绘制的关系图. PR 曲线是召回值(Recall)与精度值(Precision)的关系图，ROC 曲线是真阳性率(true postive rate)与假阳性率(false postive rate)的关系图. 此外，本文还使用其他指标评估模型性能，计算公式分别为

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

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

$$\text{Accuracy}=\frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}}, \tag{15}$$

$$F_1\_\text{score}=\frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}, \tag{16}$$

其中：TP,TN,FP 和 FN 分别表示真阳性、真阴性、假阳性和假阴性的蛋白质数量；Precision 表示预测为正的样本中正确预测的比率，表示模型对正预测判断的可信度；Recall 表示所有正样本中被正确预测的比率；Accuracy 表示正确预测的样本在所有样本中的比率；F_1_score 是精确率和召回率的调和平均值，F_1_score 越接近 1，说明模型在 Precision 和 Recall 这两个指标上的综合表现越好. 在不平衡学习中综合评估分类器性能方面，AP 值、AUC 值和 F_1_score 比其他指标更重要.

由于在数据集 BioGRID 和 DIP 上的参数实验和消融实验结论类似，所以下面仅在 BioGRID 数据集上讨论问题.

2.4　输入特征的不同维度

本文提出的深度学习框架，首先从亚细胞定位数据和蛋白质复合物数据中分别提取一个初始特征向量，再分别接入一个输出维度为 λ 的全连接层，然后将得到的两个 λ 维的特征向量拼接，形成一个 $\lambda\times2$ 维的特征向量，做为 FGN 的输入. 下面在数据集 BioGRID 上讨论 λ 分别取 1 024,512,256,128,64,32 时的实验结果，从而选取最佳值，实验结果列于表 2.

表 2　输入特征的不同维度的性能

Table 2　Performance of different dimensions of input features

λ	AP	AUC	F_1_score	Accuracy	Recall	Precision
1 024	0.784 9	0.903 5	0.755 9	0.900 7	0.729 2	0.784 8
512	0.778 0	0.898 1	0.728 0	0.885 8	0.725 0	0.731 1
256	0.753 3	0.880 0	0.720 2	0.886 6	0.691 7	0.751 1
128	0.787 4	0.898 8	0.728 7	0.874 3	0.800 0	0.669 0
64	0.767 9	0.887 4	0.716 8	0.870 8	0.775 0	0.666 7
32	0.757 3	0.893 8	0.708 3	0.889 3	0.637 5	0.796 9

由表 2 可见，在 λ 从 256 增加到 1 024 的过程中，随着 λ 的增加，AP,AUC 和 F_1_score 值也不断增加，综合可见，实验结果越来越好；在 λ 从 32 增加到 128 的过程中，随着 λ 的增加，AP,F_1_score 和 Recall 值也不断增加，但 $\lambda=256$ 时，AP,AUC,F_1_score 和 Recall 指标不如 $\lambda=128$ 的效果好. 因此 λ 在一定范围内增加，能使模型获取更多的信息，从而提升预测效果，但 λ 并不是越大越好，因为本文使用的生物数据含有噪声，λ 越大，模型获取的有用信息增加，但同时噪声信息也会增加. 对比 $\lambda=128$ 和 $\lambda=1\,024$，在 AUC,F_1_score 和 Accuracy 指标上，$\lambda=1\,024$ 的实验结果大于 $\lambda=128$ 的实验结果；$\lambda=1\,024$ 的 AP 值略小于 $\lambda=128$ 的 AP 值；$\lambda=1\,024$ 的 Recall 值比 $\lambda=128$ 小 0.0708，$\lambda=1\,024$ 的 Precision 值比 $\lambda=128$ 大 0.115 8. 综合可见，$\lambda=1\,024$ 时实验结果更好. 图 3 为输入不同维度特征的 ROC 和 PR 曲线. 由图 3 可见，不同 λ 对应的 ROC 和 PR 曲线没有显著差异，这主要是由于数据噪声问题导致的.

2.5　邻居节点的不同组合

在最初的 FGN 中，对于每个节点 v 只考虑了它的直接邻居，并且设置一个阈值 γ，表示每个节点 v 最多考虑 γ 个直接领居. 为使节点 v 通过 FGN 能学习更丰富的节点表示，本文利用图亲密度矩阵为节点 v 增加了 η 个亲密节点，$\gamma+\eta=60$. 本文在数据集 BioGRID 上讨论 γ 和 η 的不同组合，实验结果

列于表3.

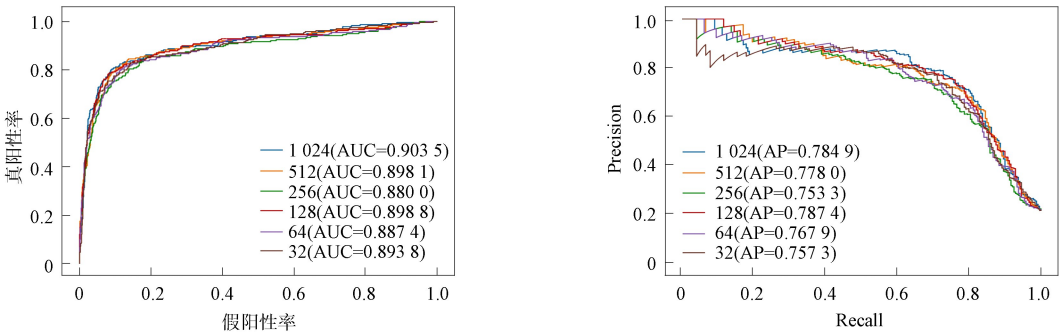

图3 输入不同维度特征的ROC和PR曲线

Fig. 3 ROC and PR curves of different dimensions of input features

表3 邻居节点不同组合的性能

Table 3 Performance of different combinations of neighbors

$\gamma:\eta$	AP	AUC	F_1_score	Accuracy	Recall	Precision
50∶10	0.784 9	0.903 5	0.755 9	0.900 7	0.729 2	0.784 8
50∶0	0.778 9	0.900 3	0.711 5	0.856 8	0.837 5	0.618 5
40∶20	0.746 9	0.891 4	0.708 8	0.866 4	0.770 8	0.656 0
30∶30	0.752 7	0.880 4	0.569 8	0.864 7	0.425 0	0.864 4
20∶40	0.769 4	0.891 0	0.639 0	0.877 9	0.512 5	0.848 3
10∶50	0.775 5	0.891 6	0.604 4	0.873 5	0.458 3	0.887 1

由表3可见，$\gamma:\eta=50:10$在AP,AUC,F_1_score和Accuracy 4个指标上都取得了最佳结果. 虽然$\gamma:\eta=50:10$对应的Recall和Precision不是最优值，但最佳Recall值对应的Precision值过小，最佳Precision值对应的Recall值也过小，又考虑到前4个指标在不平衡学习中更重要，所以认为$\gamma:\eta=50:10$时，实验结果最佳. 表3中$\gamma:\eta=50:0$表示只有直接邻居节点，对比第一行数据可见，在指标AP,AUC,F_1_score和Accuracy上，第一行均大于第二行数据. 虽然第一行的Recall值比第二行小0.108 3，但第一行的Precision值比第二行大0.166 3，综合认为第一行数据的实验结果更好，因此增加η个亲密节点有意义. γ对应于直接邻居节点，η对应于亲密节点，从两个不同的角度丰富了节点v的特征向量，从而提高了模型性能. 观察表3中第二行到第六行对应的Recall和Precision数据可见，η的值与Recall值大致呈负相关，与Precision值大致呈正相关. 从$\gamma:\eta=50:0$变化到$\gamma:\eta=50:10$，η增加了10，由表3可见，对应的Recall值下降，而Precision值上升了，从而验证了上述结论.

图4为邻居节点不同组合的ROC和PR曲线. 由图4可见，$\gamma:\eta=50:10$对应的ROC曲线和PR曲线大致包围了$\gamma:\eta=50:0$对应的曲线，再次验证了增加η个亲密节点有利于提升模型性能. 总

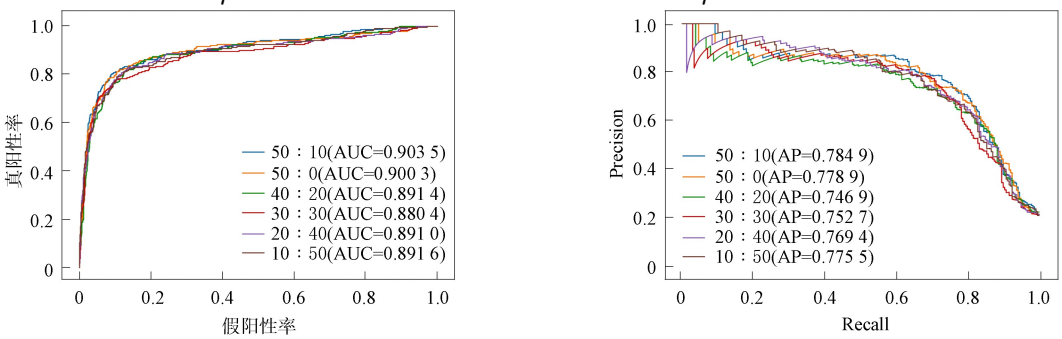

图4 邻居节点不同组合的ROC和PR曲线

Fig. 4 ROC and PR curves of different combinations of neighbors

体上看，邻居节点的不同组合对应的 ROC 和 PR 曲线没有显著差异. 本文认为是因为 $\gamma+\eta=60$，无论 γ 和 η 的哪种组合，FGN 都能学习到一个较好的节点表示.

2.6　消融实验

下面进行一系列的消融实验以验证每个部分都是必不可少的，并总结各部分对模型的贡献. 本文依次进行以下 6 个消融实验：实验 1 是本文提出的深度学习框架的完整实验，用作对照组，实验 2～6 都是实验组. 实验 2 没有对 PPI 网络进行加权；实验 3 缺少基因表达数据，因此输入到任务分类层的特征向量只有从 FGN 中输出的部分；实验 4 缺少 FGN 层，因此将得到的 1 024×2 维的特征向量展开成 2 048 维的特征向量，依次与输出维度为 512 和输出维度为 128 的全连接层相连，再与基因表达特征向量拼接得到最终的特征向量；实验 5,6 分别缺少了亚细胞定位数据和蛋白质复合物数据，因此都只能得到一个 1 024×1 维的特征向量送入 FGN 层. 由于生物信息的减少，相应地将第一个 FGN 层的输出由 128×4 改为 128×2，第二个 FGN 层的输出由 128×1 改为 64×1. 消融实验在数据集 BioGRID 上进行，实验结果列于表 4.

表 4　消融实验的结果

Table 4　Results of ablation experiments

实验序号	方法	AP	AUC	F_1_score	Accuracy	Recall	Precision
1	完整框架	0.784 9	0.903 5	0.755 9	0.900 7	0.729 2	0.784 8
2	PPI 网络没有加权	0.771 8	0.894 7	0.720 7	0.884 9	0.704 2	0.738 0
3	缺少基因表达数据	0.761 7	0.891 4	0.713 3	0.884 9	0.679 2	0.751 2
4	缺少 FGN 层	0.604 8	0.777 8	0.622 1	0.855 9	0.562 5	0.695 9
5	缺少亚细胞定位数据	0.578 9	0.832 6	0.436 7	0.487 7	0.941 7	0.284 3
6	缺少蛋白质复合物数据	0.743 6	0.893 6	0.685 4	0.842 7	0.812 5	0.592 7

由表 4 可见，本文提出的深度学习框架(即实验 1)，在除 Recall 外的所有指标上都取得了最好的结果. 虽然实验 1 的 Recall 值比最佳 Recall 值(实验 5)小，但实验 1 远大于实验 5 的 Precision 值，因此实验 1 的效果最佳. 实验 1 是对照组并且结果最好，实验 2～6 都是实验组，所以可以证明实验 1 中的每个部分都是必不可少的. 虽然实验 2,3 的结果略逊于实验 1，但仍能说明给 PPI 网络加权、利用基因表达数据对关键蛋白质的预测效果有提升作用. 实验 6 在关键指标 AP 和 F_1_score 上明显低于实验 1，Precision 值也远小于实验 1，指标 AUC 和 Accuracy 值也低于实验 1，因此可认为蛋白质复合物信息有助于提升预测性能. 实验 4,5 在关键指标 AP,AUC 和 Accuracy 上都显著低于实验 1，表明 FGN 的使用可以显著提高预测效果，亚细胞定位数据在关键蛋白质的预测中有至关重要的作用.

图 5 为消融实验的 ROC 和 PR 曲线. 由图 5 可见，实验 1 的 ROC 曲线和 PR 曲线明显包围了实验 4,5，再次验证了 FGN 和亚细胞定位数据的使用能显著提升关键蛋白质的预测效果. 实验 1 的 PR 曲线明显包围了其他实验的曲线. 而 ROC 曲线图中除实验 4,5 外，其他曲线没有显著差异.

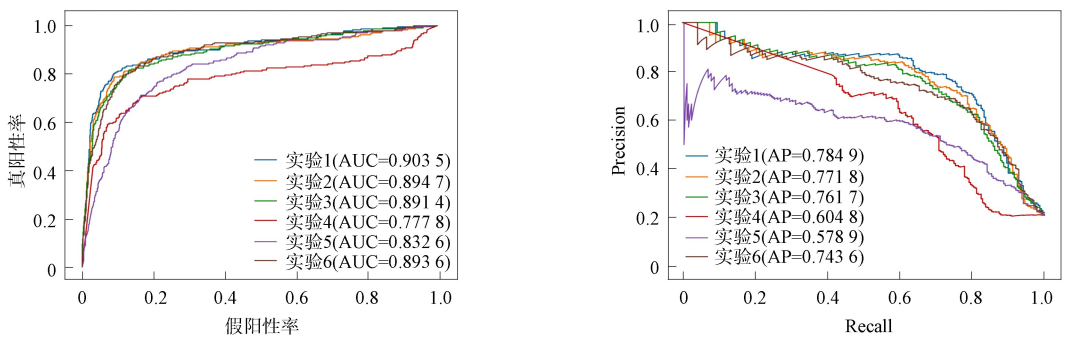

图 5　消融实验的 ROC 和 PR 曲线

Fig. 5　ROC and PR curves of ablation experiments

2.7　与基于复杂网络方法的对比

为证明本文提出的深度学习框架的有效性，本文在数据集 BioGRID 和 DIP 上将其与目前几种流行的基于复杂网络的关键蛋白质预测方法进行比较，其中 DC,BC,CC,EC,NC 和 LAC 是基于 PPI

网络的拓扑结构识别关键蛋白质的方法，而 WDC 和 PeC 则将 PPI 网络的拓扑结构与基因表达数据相结合.

首先，本文使用每种方法计算 PPI 网络中蛋白质对应的分数；其次，按降序排列蛋白质的分数，对数据集 BioGRID 和 DIP 分别选择前 1 200 和前 1 146 个蛋白质作为候选关键蛋白质；最后，根据蛋白质的真实标签计算 F_1_score，Recall，Accuracy 和 Precision. 数据集 BioGRID 和 DIP 上的实验结果分别列于表 5 和表 6. 由表 5 和表 6 可见，本文方法在两个数据集上都远优于其他方法，这是由于上述方法主要基于 PPI 网络的拓扑结构预测关键蛋白质，并且结果是一个标量. 随着网络规模和数据噪声的增加，标量不能充分捕捉有效信息. 实验结果证明了本文方法在识别关键蛋白质方面的有效性.

表 5 数据集 BioGRID 上基于复杂网络方法的实验结果

Table 5 Experimental results based on complex network methods on BioGRID dataset

方法	F_1_score	Accuracy	Recall	Precision
本文	0.755 9	0.900 7	0.729 2	0.784 8
DC	0.446 7	0.766 4	0.446 7	0.446 7
BC	0.407 5	0.749 9	0.407 5	0.407 5
CC	0.321 7	0.713 7	0.321 7	0.321 7
EC	0.411 7	0.751 7	0.411 7	0.411 7
NC	0.455 8	0.770 3	0.455 8	0.455 8
LAC	0.450 0	0.767 9	0.450 0	0.450 0
WDC	0.466 7	0.774 9	0.466 7	0.466 7
PeC	0.423 3	0.756 6	0.423 3	0.423 3

表 6 数据集 DIP 上基于复杂网络方法的实验结果

Table 6 Experimental results based on complex network methods on DIP dataset

方法	F_1_score	Accuracy	Recall	Precision
本文	0.715 9	0.873 3	0.673 9	0.763 5
DC	0.408 4	0.720 4	0.408 4	0.408 4
BC	0.361 3	0.698 1	0.361 3	0.361 3
CC	0.377 0	0.705 6	0.377 0	0.377 0
EC	0.377 8	0.706 0	0.377 8	0.377 8
NC	0.452 0	0.741 0	0.452 0	0.452 0
LAC	0.459 0	0.744 3	0.459 0	0.459 0
WDC	0.463 4	0.746 4	0.463 4	0.463 4
PeC	0.445 0	0.737 7	0.445 0	0.445 0

2.8 与基于机器学习方法的对比

基于机器学习方法已广泛应用于预测关键蛋白质. 为证明本文方法的优越性，本文采用 6 种传统机器学习方法和 3 种深度学习框架(DeepEP，文献[14]框架和文献[16]框架)进行对比实验. 实验中传统机器学习方法 SVM、AdaBoost、逻辑回归、朴素 Bayes、随机森林和决策树是由带有默认参数的 scikit-learn 库实现的. DeepEP 和文献[16]中提供了源代码. 本文根据文献[14]实现了该模型.

传统机器学习算法将每个蛋白质对应的亚细胞定位特征向量、蛋白质复合物特征向量和基因表达数据拼接成一个长的一维向量作为输入. DeepEP 仅使用 PPI 网络和基因表达数据. 文献[14]方法使用了 PPI 网络、基因表达数据和亚细胞定位数据. 亚细胞定位通常分为 11 个类别，文献[14]的方法使用 11 维向量编码亚细胞定位数据，然后作为深度学习框架的输入. 文献[16]方法使用了本文中的 PPI 网络和亚细胞定位数据，但基因表达数据与本文不同，来自数据库 GEO(登录号：GSE7645). 本文按文献[16]的方式对亚细胞定位数据进行处理，然后作为其框架的输入. 为确保不同方法比较的公平性，本文在 BioGRID 和 DIP 的相同测试数据集上进行对比实验，实验结果列于表 7 和表 8.

由表 7 和表 8 可见，在酵母菌数据集 BioGRID 和 DIP 上，本文提出的框架在除 Recall 指标外的各指标上都明显高于其他方法. 虽然逻缉回归的 Recall 值略高于本文方法，但其他指标远低于本文方法

的值，综上表明本文方法优于目前主流基于机器学习的关键蛋白质预测方法. 4 种深度学习方法在大多数指标上优于其他传统机器学习方法，表明在数据集 BioGRID 和 DIP 上深度学习方法比传统机器学习方法能学习到更多信息. 随机森林的效果在数据集 BioGRID 上略差于 DeepEP，但是在数据集 DIP 上高于 DeepEP. DeepEP 是深度学习框架，随机森林是传统机器学习，但 DeepEP 只利用了 PPI 网络和基因表达数据，而随机森林比 DeepEP 多利用了亚细胞定位数据和蛋白质复合物数据，生物信息更丰富，因此丰富的生物信息弥补了结构上的缺陷，使得随机森林和 DeepEP 取得了相似的结果，甚至在数据集 DIP 上超越了 DeepEP. 文献[14]方法比 DeepEP 多了一种亚细胞定位数据，除指标 Recall 外，所有指标都比 DeepEP 更好. 进一步表明融合更多有效生物信息有利于提高关键蛋白质的预测准确率.

表 7　数据集 BioGRID 上基于机器学习方法的实验结果

Table 7　Experimental results based on machine learning methods on BioGRID dataset

方法	AP	AUC	F_1_score	Accuracy	Recall	Precision
本文	0.784 9	0.903 5	0.755 9	0.900 7	0.729 2	0.784 8
DeepEP	0.614 5	0.828 7	0.600 7	0.804 9	0.695 8	0.528 5
文献[14]	0.655 4	0.842 9	0.628 9	0.833 0	0.670 8	0.591 9
文献[16]	0.747 5	0.880 7	0.664 1	0.848 9	0.708 3	0.625 0
AdaBoost	0.455 4	0.779 0	0.524 2	0.714 4	0.745 8	0.404 1
决策树	0.350 1	0.695 5	0.507 3	0.762 7	0.579 2	0.451 3
朴素 Bayes	0.327 2	0.667 8	0.481 7	0.701 2	0.658 3	0.379 8
逻辑回归	0.466 0	0.788 8	0.540 1	0.723 2	0.770 8	0.415 7
SVM	0.459 5	0.759 7	0.477 0	0.660 8	0.733 3	0.353 4
随机森林	0.579 4	0.825 3	0.549 6	0.740 8	0.750 0	0.433 7

表 8　数据集 DIP 上基于机器学习方法的实验结果

Table 8　Experimental results based on machine learning methods on DIP dataset

方法	AP	AUC	F_1_score	Accuracy	Recall	Precision
本文	0.776 2	0.902 9	0.727 3	0.873 3	0.713 0	0.742 1
DeepEP	0.543 8	0.737 5	0.532 5	0.755 9	0.587 0	0.487 4
文献[14]	0.578 6	0.783 4	0.542 6	0.789 9	0.526 1	0.560 2
文献[16]	0.725 4	0.873 2	0.646 2	0.798 1	0.778 3	0.552 5
AdaBoost	0.425 5	0.713 0	0.451 4	0.692 1	0.534 8	0.390 5
决策树	0.322 4	0.659 2	0.477 7	0.614 8	0.743 5	0.351 9
朴素 Bayes	0.383 3	0.689 5	0.525 7	0.762 1	0.556 5	0.498 1
逻辑回归	0.477 6	0.792 2	0.569 6	0.719 9	0.782 6	0.447 8
SVM	0.375 5	0.706 8	0.308 1	0.699 3	0.282 6	0.338 5
随机森林	0.589 9	0.770 6	0.543 6	0.816 7	0.460 9	0.662 5

如图 6 和图 7 所示，本文方法在数据集 DIP 上的 ROC 曲线和 PR 曲线明显包围了其他方法的曲线，在数据集 BioGRID 上的 ROC 曲线也明显包围了其他方法的曲线. 虽然本文方法在数据集 BioGRID 数据集上的 PR 曲线与文献[16]方法的曲线略有交叉，但本文方法的 AP 值大于文献[16]方法的 AP 值，因此本文方法总体上优于目前主流基于机器学习的关键蛋白质预测方法.

综上所述，针对生物实验识别关键蛋白质费时费力，使用计算方法预测关键蛋白质无法有效整合生物信息的问题，本文提出了一个基于特征图网络和多种生物信息预测关键蛋白质的深度学习框架. 该框架考虑如何更好利用 PPI 网络中的边缘信息，从而更好地提取特征向量，最终提高关键蛋白质的预测准确率. 首先，利用基因表达数据、GO 注释数据和 PPI 网络拓扑特征对 PPI 网络进行加权，使 PPI 网络中的边缘信息更丰富；其次，通过使用 FGN 将边缘信息编码到特征邻接矩阵中，从而能更好地保存并利用边缘信息. 在酵母菌数据集 BioGRID 和 DIP 上的实验结果表明，本文方法优于目前主流的复杂网络方法和机器学习方法. 消融实验结果表明，本文框架中的每部分都是必不可少的，其中 FGN 和亚细胞定位数据的使用显著提高了关键蛋白质的预测性能，蛋白质复合物数据也有助于提高

预测效果. 通过给PPI网络加权，能减少数据中噪声的影响，从而进一步提高对关键蛋白质的预测性能.

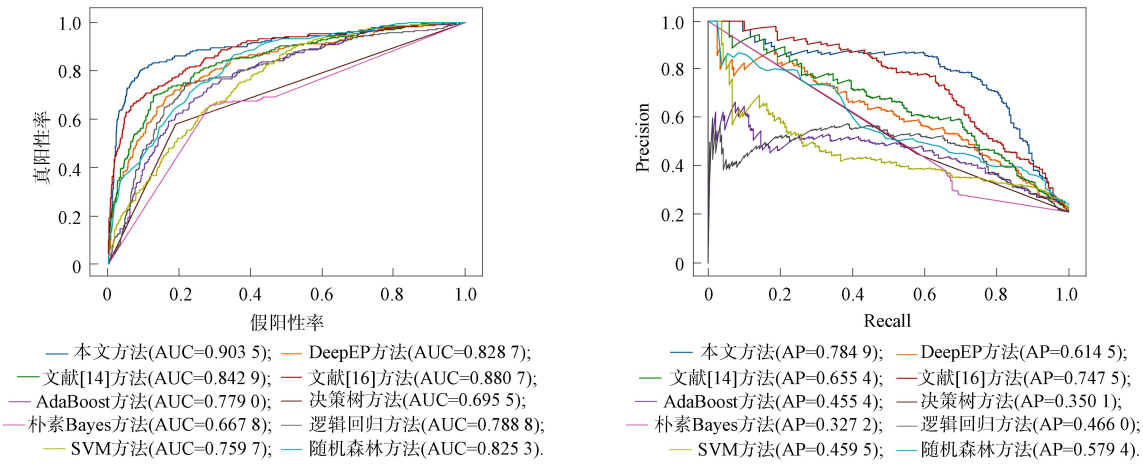

图6 数据集BioGRID上对比实验的ROC和PR曲线

Fig. 6 ROC and PR curves of comparative experiments on BioGRID dataset

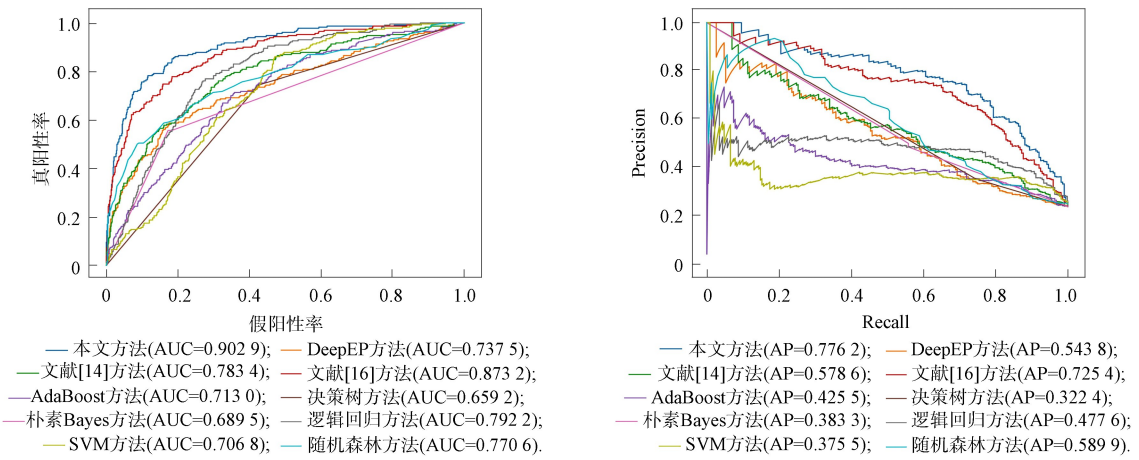

图7 数据集DIP上对比实验的ROC和PR曲线

Fig. 7 ROC and PR curves of comparative experiments on DIP dataset

参 考 文 献

[1] WINZELER E A, SHOEMAKER D D, ASTROMOFF A, et al. Functional Characterization of the *S. cerevisiae* Genome by Gene Deletion and Parallel Analysis [J]. Science, 1999, 285: 901-906.

[2] CULLEN L M, ARNDT G M. Genome-Wide Screening for Gene Function Using RNAi in Mammalian Cells [J]. Immunology and Cell Biology, 2005, 83(3): 217-223.

[3] ROEMER T, JIANG B, DAVISON J, et al. Large-Scale Essential Gene Identification in Candida Albicans and Applications to Antifungal Drug Discovery [J]. Molecular Microbiology, 2003, 50(1): 167-181.

[4] GIAEVER G, CHU A M, NI L, et al. Functional Profiling of the *Saccharomyces cerevisiae* Genome [J]. Nature, 2002, 418: 387-391.

[5] JEONG H, MASON S P, BARABÁSI A L, et al. Lethality and Centrality in Protein Networks [J]. Nature, 2001, 411: 41-42.

[6] LI M, WANG J X, CHEN X, et al. A Local Average Connectivity-Based Method for Identifying Essential Proteins from the Network Level [J]. Computational Biology and Chemistry, 2011, 35(3): 143-150.

[7] QI Y, LUO J W. Prediction of Essential Proteins Based on Local Interaction Density [J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2015, 13(6): 1170-1182.

[8] LI M, ZHANG H H, WANG J X, et al. A New Essential Protein Discovery Method Based on the Integration of Protein-Protein Interaction and Gene Expression Data [J]. BMC Systems Biology, 2012, 6: 1-9.

[9] TANG X W, WANG J X, ZHONG J C, et al. Predicting Essential Proteins Based on Weighted Degree Centrality [J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2013, 11(2): 407-418.

[10] LEI X J, ZHAO J, FUJITA H, et al. Predicting Essential Proteins Based on RNA-Seq, Subcellular Localization and GO Annotation Datasets [J]. Knowledge-Based Systems, 2018, 151: 136-148.

[11] QIN C, SUN Y Q, DONG Y D. A New Method for Identifying Essential Proteins Based on Network Topology Properties and Protein Complexes [J]. PloS One, 2016, 11(8): e0161042-1-e0161042-30.

[12] ZENG M, LI M, WU F X, et al. DeepEP: A Deep Learning Framework for Identifying Essential Proteins [J]. BMC Bioinformatics, 2019, 20: 1-10.

[13] GROVER A, LESKOVEC J. Node2vec: Scalable Feature Learning for Networks [C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2016: 855-864.

[14] ZENG M, LI M, FEI Z H, et al. A Deep Learning Framework for Identifying Essential Proteins by Integrating Multiple Types of Biological Information [J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2019, 18(1): 296-305.

[15] HUANG Z H, XU W, YU K. Bidirectional LSTM-CRF Models for Sequence Tagging [EB/OL]. (2015-08-15) [2023-02-11]. https://arxiv.org/abs/1508.01991.

[16] YUE Y, YE C, PENG P Y, et al. A Deep Learning Framework for Identifying Essential Proteins Based on Multiple Biological Information [J]. BMC Bioinformatics, 2022, 23(1): 318-1-318-27.

[17] LI Y M, ZENG M, ZHANG F H, et al. DeepCellEss: Cell Line-Specific Essential Protein Prediction with Attention-Based Interpretable Deep Learning [J]. Bioinformatics, 2023, 39(1): btac779-1-btac779-9.

[18] LIU P Q, LIU C, MAO Y Y, et al. Identification of Essential Proteins Based on Edge Features and the Fusion of Multiple-source Biological Information [J]. BMC Bioinformatics, 2023, 24(1): 203-1-203-24.

[19] WANG C, QIU Y H, GAO D S, et al. Lifelong Graph Learning [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 13719-13728.

[20] KERETSU S, SARMAH R. Weighted Edge Based Clustering to Identify Protein Complexes in Protein-Protein Interaction Networks Incorporating Gene Expression Profile [J]. Computational Biology and Chemistry, 2016, 65: 69-79.

[21] LEI X J, ZHANG Y C, CHENG S, et al. Topology Potential Based Seed-Growth Method to Identify Protein Complexes on Dynamic PPI Data [J]. Information Sciences, 2018, 425: 140-153.

[22] RADICCHI F, CASTELLANO C, CECCONI F, et al. Defining and Identifying Communities in Networks [J]. Proceedings of the National Academy of Sciences, 2004, 101(9): 2658-2663.

[23] LEI X J, YANG X Q, FUJITA H. Random Walk Based Method to Identify Essential Proteins by Integrating Network Topology and Biological Characteristics [J]. Knowledge-Based Systems, 2019, 167: 53-67.

[24] BINDER J X, PLETSCHER-FRANKILD S, TSAFOU K, et al. COMPARTMENTS: Unification and Visualization of Protein Subcellular Localization Evidence [J]. Database, 2014, 2014: bau012-1-bau012-9.

[25] ZHANG J W, ZHANG H P, XIA C Y, et al. Graph-BERT: Only Attention Is Needed for Learning Graph Representations [EB/OL]. (2020-01-15)[2023-01-15]. https://arxiv.org/abs/2001.05140.

[26] JACCARD P. Étude Comparative de la Distribution Florale Dans une Portion des Alpes et des Jura [J]. Bulletin de la Societe Vaudoise des Sciences Naturelles, 1901, 37: 547-579.

（责任编辑：韩　啸）