计算机科学
COMPUTER SCIENCE

基于超图网络嵌入的蛋白质复合体识别算法

王杰, 杨贤灿, 赵兴旺

引用本文

王杰, 杨贤灿, 赵兴旺. 基于超图网络嵌入的蛋白质复合体识别算法[J]. 计算机科学, 2025, 52(12): 102-114.

WANG Jie, YANG Xiancan, ZHAO Xingwang. Protein Complex Identification Algorithm Based on Hypergraph Network Embedding [J]. Computer Science, 2025, 52(12): 102-114.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.250900062

基于超图网络嵌入的蛋白质复合体识别算法

王　杰[1]　杨贤灿[1]　赵兴旺[2]
1 山西财经大学信息学院　太原 030006
2 山西大学计算机与信息技术学院　太原 030006
(20191031@sxufe.edu.cn)

摘　要　蛋白质复合体在细胞生物学过程中起着关键作用,对理解细胞功能和生物过程的识别至关重要。在蛋白质-蛋白质相互作用(Protein-Protein Interaction,PPI)网络中采用网络聚类识别蛋白质复合体已经成为数据挖掘与生物信息学的研究热点,各种计算方法被提出用于识别蛋白质复合体。然而,大多数方法仅利用原始网络来挖掘密集子图或子网络,未能突破传统图结构对多节点交互关系的局限。针对生物网络中普遍存在的多对多复杂交互特性问题,提出基于超图网络嵌入的蛋白质复合体识别算法(Protein Complex Identification Method Based on Hypergraph Network Embedding,PCIHNE)。该算法首先利用超图网络对多元关系的直接建模能力,将原始PPI网络转换为超图网络。其次,对超图网络采用分层压缩策略递归地压缩为多个不同层次的较小超图,以此构建多尺度分析框架。再次,将超图卷积应用于不同层次,得到每个节点在不同尺度下的表示。将这些节点表示进行连接,得到完整的节点嵌入表示。基于节点嵌入表示,在低阶原始网络上构建加权PPI网络。最后,在加权PPI网络上采用基于核心附属策略,得到预测的蛋白质复合体。在多个酵母和人类真实的数据集上将所提算法与其他蛋白质复合体识别算法进行比较,实验结果表明,所提方法在F-measure和Accuracy指标上取得了较好的蛋白质复合体识别性能。

关键词:蛋白质相互作用网络;蛋白质复合体;超图;网络嵌入;网络聚类
中图分类号　TP399

Protein Complex Identification Algorithm Based on Hypergraph Network Embedding

WANG Jie[1],YANG Xiancan[1] and ZHAO Xingwang[2]
1 School of Information,Shanxi University of Finance and Economics,Taiyuan 030006,China
2 School of Computer and Information Technology,Shanxi University,Taiyuan 030006,China

Abstract　Protein complexes are crucial for understanding cellular functions and identifying biological processes,playing critical roles in cell biology. The use of network clustering in PPI networks to identify protein complexes has become a hot research topic in data mining and bioinformatics. A variety of computational methods have emerged to identify protein complexes. However,most existing algorithms primarily use original network to detect dense subnetworks and fail to break through the limitations of traditional graph structures for multi-node interactions. Aiming at the issue of many-to-many complex interaction characteristics prevalent in biological networks,this paper proposes a novel protein complex identification method based on hypergraph network embedding(PCIHNE). Through the ability of hypergraph networks,it firstly converts the original PPI network into a hypergraph network. Then a hierarchical compression strategy recursively compresses the hypergraph into multiple smaller hypergraphs at different levels,thereby constructing a multi-scale analysis framework. Next,hypergraph convolution is performed on each levels to generate node representations at different granularities. These node representations are concatenated to obtain the complete node representation. Based on the representations obtained from hypergraph learning,a weighted PPI network is constructed by similarity on the original network. Finally,a core-attachment based strategy is used to obtain predicted protein complexes in the weighted PPI network. It evaluates the effectiveness of PCIHNE by comparing it with other protein complex algorithms on multiple yeast and human datasets. Experimental results demonstrate that PCIHNE is better than comparison protein complex identification methods regarding F-measure and Accuracy metrics.

Keywords　Protein-protein interaction network,Protein complexes,Hypergraphs,Network embedding,Network clustering

到稿日期:2025-09-09　返修日期:2025-11-10
基金项目:国家自然科学基金(62006145);山西省基础研究计划(202303021212169)
This work was supported by the National Natural Science Foundation of China(62006145) and Fundamental Research Program of Shanxi Province (202303021212169).
通信作者:杨贤灿(937741973@qq.com)

1　引言

蛋白质是由氨基酸通过肽键连接形成的大分子有机物，是细胞结构与功能的核心执行者。从酶催化、信号转导到免疫防御，蛋白质几乎参与了所有生命活动。然而，绝大多数蛋白质并非孤立的发挥作用，而是通过与其他生物分子（如蛋白质、核酸等）形成复合体来实现复杂功能[1]。作为细胞过程中的关键功能单位，蛋白质复合体参与了细胞周期调控、信号传导及基因表达调控等多种生物过程，已有大量证据表明，蛋白质复合体与多种疾病密切相关[2]，在生物医学研究方面具有重要意义。

识别蛋白质复合体的实验方法（如核磁共振（NMR）[3]和质谱技术[4]）由于成本高、耗时长，难以全面解析所有蛋白质复合体。挖掘蛋白质-蛋白质相互作用（Protein-Protein Interaction，PPI）网络中的紧密连通节点集合，是利用计算方法识别蛋白质复合体的主要途径[5]。PPI 可以通过基于序列的预测方法[6]、基于深度学习的预测方法[7-8]和基于遗传进化关系的预测方法[9]等来获取，进而构建丰富的 PPI 网络。以 PPI 网络为对象，基于网络拓扑来设计蛋白质复合体识别计算方法成为研究热点之一，这项研究有助于深入理解细胞内复杂的生物过程，推动疾病研究和药物开发，具有重要的研究价值和应用前景[10]。

现有蛋白质复合体识别算法主要是在传统 PPI 网络上利用二元交互关系进行复合体识别，这种基于低阶原始网络的建模方式难以表征蛋白质群体间真实存在的多节点协同作用。为了弥补这一不足，可以通过超图特有的多元关系建模能力探索生物网络中的复杂结构。超图可以表示蛋白质节点中的高阶非成对的复杂关系，相较于只能表示节点间成对连接的原始图结构，它允许单条超边同时连接多个节点，而不仅限于一对一[11]。超图的这种一对多关系建模特性适用于描述多个蛋白质节点组合形成蛋白质复合体的过程。

为提高蛋白质复合体识别的精度，本文突破传统图模型的维度限制，采用高阶超图（Hypergraph）网络模型来建模蛋白质相互作用关系，提出了基于超图网络嵌入的蛋白质复合体识别算法（PCIHNE）。PCIHNE 算法将原始 PPI 网络转换为超图，并采用分层压缩策略提取不同层次超图的拓扑特征，从而得到蛋白质节点的嵌入表示。在此基础上，通过核心附属策略识别蛋白质复合体。该算法不仅突破了传统图模型在二元关系表达上的限制，而且有效利用了 PPI 网络的局部和全局拓扑信息。

2　相关工作

目前，利用网络拓扑信息识别蛋白质复合体的算法主要包括基于图划分算法、基于启发式算法和基于图嵌入算法等。

2.1　基于图划分算法

基于图划分算法的核心思想是通过优化特定的目标函数，将图划分为若干个互不相交的子图，每个子图包含图中一定数量的节点或边，使得子图内部的连接关系紧密，不同子图之间的连接关系稀疏[12]。基于此思想设计的算法，通常将子图作为蛋白质复合体。例如，马尔可夫聚类算法（MCL）[13]通过模拟随机游走来识别图中的子图。该算法在输入图中添加环，并将其转换为马尔可夫矩阵，通过扩展和膨胀操作反复更新矩阵，直至图被分割为多个子图，每个子图即为预测的蛋白质复合体。然而，通过此方法只能得到互斥的聚类结果。也就是说，虽然 MCL 可以处理噪声节点，但无法处理重叠簇。为克服这一局限性，提出了软正则化 MCL（SR-MCL）[14]算法。BOPS[15]首先计算边的平衡权重并删除最大权重边来划分网络，直到每个连通分量的节点数不超过阈值 MAXP。然后在每个划分后的子网内枚举所有连通子图，计算凝聚度并根据凝聚度进行降序排序，筛选并得到候选簇。最后，保留重叠度超过阈值的候选簇，作为最终的蛋白质复合体。除了通过对图进行划分外，还可以基于节点的密度特征进行划分。DPFO[16]通过改进的密度峰值聚类（DPC）和模糊 C 均值（FCM）对 PPI 网络进行划分，从而识别出重叠的蛋白质复合体。

基于该思想的蛋白质复合体识别算法通过划分 PPI 网络实现蛋白质复合体的识别，忽略了蛋白质群体间真实存在的多节点协同作用，难以捕捉生物网络中多对多的交互特性。

2.2　基于启发式算法

基于启发式算法通过将种子蛋白质识别为单一簇，然后贪婪地扩展这些簇，以实现对蛋白质的聚类。其中经典算法包括 ClusterONE[17]和 DPClus[18]。ClusterONE 用于处理 PPI 网络中的重叠簇问题。从单个种子顶点开始，通过贪婪算法找到高密度的簇，然后量化每对簇之间的重叠程度后合并重叠簇，丢弃不符合特定标准的复合体候选簇。该算法将高密度区域作为蛋白质复合体。DPClus 引入了“簇外围”的概念，通过公共邻居计数分配边权重，并根据相邻边权重之和确定节点权重。该算法从加权最高的节点开始，将节点作为初始种子，逐步扩展集群，直至满足密度和簇外围的条件。DPClus 能够更精准地识别高密度区域。DMPC[19]在 PPI 网络中将所有密度为 1 的最大簇作为种子，然后递归地将当前簇相连的邻居节点加入每个簇中，得到更大的簇，并计算簇的密度。当扩展后簇的密度仍大于阈值时，进一步调用 C4.5 决策树对该簇进行判定，若判定为复合体，才继续在其基础上进行下一轮扩展，否则立即停止扩展。通过这个过程，DMPC 能够准确识别出支持重叠的蛋白质复合体。另一种使用种子扩展的启发式算法被称为核心-附属。Gavin 等[20]首先提出蛋白质复合体由核心和附属组成的观点。其核心思想是将蛋白质复合体分解为两类功能单元，即核心蛋白质和附属蛋白质。利用网络中蛋白质之间的连接模式，将稳定存在且频繁重复出现的蛋白质集合定义为“核心”，再将其他较为松散的、动态加入或脱离核心的蛋白质定义为“附属”蛋白。基于此观点，CORE[21]算法引入核心-附属（Core-Attachment）模型。首先预测复合体的核心部分，然后识别与核心部分相互作用的附属蛋白质。COACH[22]算法从每个蛋白质节点的邻域图中筛选节点度数不低于平均值的节点作为预备核心蛋白质。首先，利用核心剔除算法递归地提取出密度较高的子图，过滤得到最终的核心；接着，在核心的直接邻域筛选与核心蛋白质一半以上发生交互的节点作为附属蛋白，并将它们与核心蛋白结合，得到预测蛋白质复合体。在上述算法思想的基础上，发展出了 WCOACH[23]，WPNCA[24]，Multiobjective[25]，PCD-

BA[26]和 GCAPL[27]等算法。

基于启发式的蛋白质复合体识别算法在 PPI 网络上通过一对一的节点交互设计实现，同样忽略了生物网络中多对多的复杂交互特性。

2.3 基于图嵌入算法

图嵌入是一种将图中节点（或其他图结构，如边、子图）映射到低维向量空间的技术，它尽量保留原始拓扑图中的结构信息，并使得嵌入空间的节点之间的相似度可以近似于原始拓扑网络中的相似度[28]。例如，ELF-DPC[29]提出了一个集成学习框架，其思路是利用投票回归模型来增强蛋白质复合体的识别能力。DPCMNE[30]通过将局部与全局的拓扑信息映射到低维向量空间，得到不同粒度下节点的嵌入表示，以此来加权原始拓扑图。具体地，首先采用 Louvain 算法对原始 PPI 网络进行模块划分，将紧密连接的蛋白质节点合并形成多个模块，进而生成不同尺度的网络视图。针对不同的视图，使用 DeepWalk 算法学习蛋白质的低维特征表示。最后利用这些嵌入特征构建加权网络，并通过核心-附属策略识别蛋白质复合体。这种方式不仅考虑了全局拓扑信息，同时又融合了局部拓扑信息。AdaPPI[31]首先利用自适应图卷积网络为每个节点学习多阶拓扑信息的低维嵌入；接着在嵌入空间中根据节点间余弦相似度挖掘所有最大团，并计算每个团体的嵌入密度，将高密度团体作为核心候选；随后通过核心-附属策略识别蛋白质复合体。通过这种方式，AdaPPI 能够高效挖掘并准确识别可重叠的蛋白质复合体。除此之外，研究人员基于图嵌入思想，引入超图模型并结合各种生物信息和蛋白质动态特征识别蛋白质复合体，如 HGST[32]和 HyperGraphComplex[33]。

基于图嵌入的蛋白质复合体识别算法大多将 PPI 网络映射到低维向量空间，忽略了蛋白质节点中的高阶非成对的复杂关系。引入超图模型的图嵌入方法忽略了 PPI 网络中低阶成对的关系。

综上所述，现有计算方法主要在低阶原始拓扑图上设计算法来识别蛋白质复合体，并没有考虑到低阶原始网络的建模方式难以表征蛋白质群体间真实存在的多节点协同作用。基于超图蛋白质复合体识别方法只考虑了高阶拓扑信息，忽略了 PPI 网络中成对的相互作用信息，而结合高阶拓扑信息和低阶拓扑信息有助于蛋白质复合体的识别。为了弥补上述不足，本文通过超图特有的多元关系建模能力表征生物网络中的复杂结构，并结合高阶超图和低阶 PPI 网络拓扑信息，提出了基于超图网络嵌入的蛋白质复合体识别算法（PCIHNE）。

3 本文方法

本章详细介绍了基于超图网络嵌入的蛋白质复合体识别算法 PCIHNE。PCIHNE 包含 4 个步骤：超图构建、分层压缩、节点嵌入和节点聚类。首先将原始 PPI 网络转换为超图网络。其次在分层压缩阶段，利用超图模块度方法将超图递归地压缩为多个不同层次的较小超图。然后在节点嵌入阶段，将超图卷积应用于不同层次的多个小超图中，得到每个节点在不同尺度下的嵌入表示。将这些节点嵌入表示合并，得到最终的节点嵌入表示。最后在节点聚类阶段，利用基于核心附属策略来识别蛋白质复合体。图 1 展示了算法整体流程。

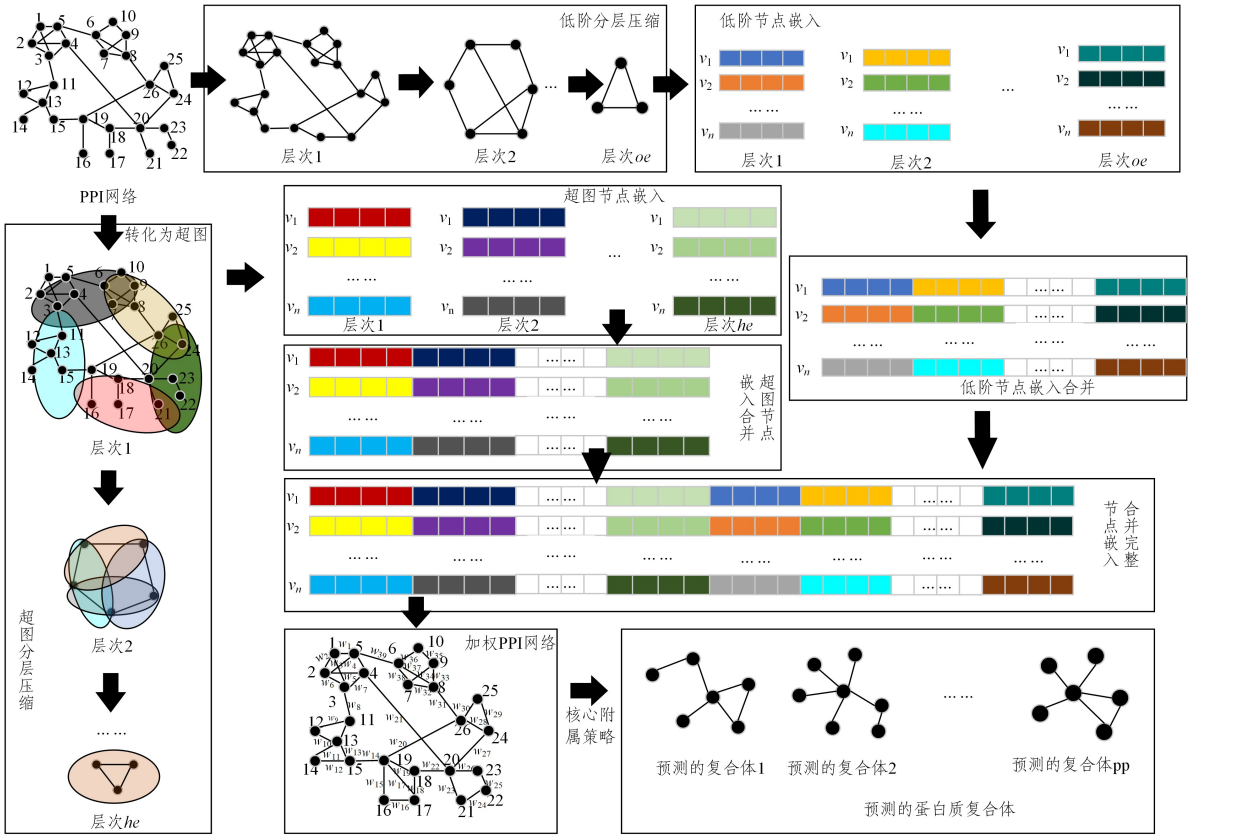


图 1 算法流程图

Fig. 1 Flowchart of algorithm

3.1　超图构建

原始 PPI 网络通常表示为无向图 $G(V_{\text{ppi}},E_{\text{ppi}})$。本文将原始 PPI 网络转化为超图 $\mathcal{G}(V,E,\boldsymbol{W},\boldsymbol{X})$，其中，$V$ 表示节点集，E 表示超边集，$\boldsymbol{W}$ 表示由超边权重组成的对角矩阵。特征矩阵 $\boldsymbol{X}\in R^{n*q}$ 由所有节点的特征向量组成，其中 n 是节点数，q 是节点特征的维度。

采用下述转换规则将原始 PPI 网络转换为超图。蛋白质节点 v_i 的邻域表示如式(1)所示：

$$N(v_i)=\{v_j\in V_{\text{ppi}}\mid(v_i,v_j)\in E_{\text{ppi}}\} \tag{1}$$

其中，$(v_i,v_j)\in E_{\text{ppi}}$ 表示节点 v_i 和 v_j 在原始 PPI 网络上互为相互作用。此时构造的超边如式(2)所示：

$$e_k=\{v_i\}\bigcup N(v_i) \tag{2}$$

对于重复的超边，例如节点 v_i 和 v_j 在原始 PPI 网络上互为相互作用，那么 $e_k=\{v_i,v_j\}$ 和 $e_k=\{v_j,v_i\}$ 为同一超边。

关联矩阵 $\boldsymbol{H}$ 是 $|V|\times|E|$ 的矩阵，用来描述节点与超边之间的关系，矩阵 $\boldsymbol{H}$ 中的元素 $h(v_i,e_k)$ 表示节点 v_i 是否属于超边 e_k，计算方法如式(3)所示：

$$h(v_i,e_k)=\begin{cases}1, & v_i\in e_k\\ 0, & v_i\notin e_k\end{cases} \tag{3}$$

其中，当蛋白质节点 v_i 属于超边 e_k 时，$h(v_i,e_k)=1$，反之则为 0。

设 $\boldsymbol{D}_e=\text{diag}(D_{e_1},D_{e_2},\cdots,D_{e_k})$ 表示超边度矩阵，其中超边度 D_{e_k} 表示超边 e_k 所包含的节点个数。设 $\boldsymbol{D}_v=\text{diag}(D_{v_1},D_{v_2},\cdots,D_{v_i})$ 表示节点度矩阵，其中节点度 D_{v_i} 表示某个节点 v_i 在超图中的加权度。超边度 D_{e_k} 和节点度 D_{v_i} 的计算方法分别如式(4)和式(5)所示：

$$D_{e_k}=\sum_{v\in V}h(v_i,e_k) \tag{4}$$

$$D_{v_i}=\sum_{e\in E}\omega(e_k)h(v_i,e_k) \tag{5}$$

其中，$\omega(e_k)=1/|e_k|$ 表示每条超边的权重，$|e_k|$ 表示超边 e_k 中节点的数量。

3.2　分层压缩

本文使用超图模块度将超图 $\mathcal{G}$ 划分为多个较小超图，实现分层压缩，以此获取超图网络的局部和全局拓扑信息[34]。

传统的超图模块度引入一个零模型(Null Model)，并基于该零模型直接定义模块度。然而，在生物网络中，由 PPI 网络转换而来的超边大小往往存在较大差异。如果直接使用传统的超图模块度，超边大小的不一致可能导致模型偏向大的超边，从而影响分层压缩结果的合理性。对此，本文使用了一种基于节点度保持的超图模块度方法[35]。该方法需要将超图转化为一个加权图，并保证转换后加权图中节点的度数与原始超图一致，从而避免大超边造成的度偏差。在此加权图上定义零模型，从而得到新的超图模块度。

将超图转换为加权图，把超图 $\mathcal{G}$ 中的每条超边 e_k 替换为由其所有蛋白质节点构成的完全子图(Clique)，把所有超边对应的完全子图叠加起来得到完整的加权图，此时对应的邻接矩阵如式(6)所示：

$$\boldsymbol{A}^{\text{clique}}=\boldsymbol{HWH}^{\text{T}} \tag{6}$$

其中，$\boldsymbol{H}$ 是原超图的关联矩阵，$\boldsymbol{W}$ 是超边权重矩阵。去掉该邻接矩阵中的自环后，蛋白质节点 v_i 在该加权图的度数等于其在邻接矩阵中对应行的元素和，定义如式(7)所示：

$$deg(v_i)=\sum_j A_{ij}^{\text{clique}}=\sum_{e_k\in E}h(v_i,e_k)\omega(e_k)(D_{e_k}-1) \tag{7}$$

其中，D_{e_k} 表示超边 e_k 所包含的顶点数。此时，$deg(v_i)$ 表示加权图中的节点度，而非原图中的节点度 D_{v_i}。

为了让加权图中每个节点的度数恰好等于其在原超图中的度，归一化每条超边的“贡献”，即在矩阵层面引入 $(\boldsymbol{D}_e-\mathbf{I})^{-1}$。故归一化后的加权图的邻接矩阵定义如下：

$$\boldsymbol{A}^{\text{hyp}}=\boldsymbol{H}\cdot\boldsymbol{W}\cdot(\boldsymbol{D}_e-\mathbf{I})^{-1}\cdot\boldsymbol{H}^{\text{T}} \tag{8}$$

其中，$\boldsymbol{A}^{\text{hyp}}$ 为归一化后的加权图邻接矩阵，$\boldsymbol{W}=\text{diag}(\omega(e_1),\cdots,\omega(e_k))$ 表示超边权重的对角线，$\mathbf{I}$ 为单位矩阵。

接下来引入一个保留节点度的零模型。在该模型中，两个蛋白质节点间的期望连接数定义如式(9)所示：

$$P_{ij}^{\text{hyp}}=\frac{nd(v_i)\times nd(v_j)}{\sum_{v\in V}nd(v)} \tag{9}$$

其中，$nd(v_i)$ 和 $nd(v_j)$ 分别表示蛋白质节点 v_i 和 v_j 在原始超图中的度，$\sum_{v\in V}nd(v)$ 是原始超图中所有节点的加权度。

由加权图和零模型的期望值之差定义超图模块度，如式(10)所示：

$$B_{ij}^{\text{hyp}}=A_{ij}^{\text{hyp}}-P_{ij}^{\text{hyp}} \tag{10}$$

其中，A_{ij}^{hyp} 是加权图中蛋白质节点 v_i 和 v_j 在邻接矩阵中对应的值；P_{ij}^{hyp} 是期望权重。此时，超图模块度的表达式为：

$$Q^{\text{hyp}}=\frac{1}{2m}\sum_{ij}B_{ij}^{\text{hyp}}\delta(v_i,v_j) \tag{11}$$

其中，$\delta(v_i,v_j)$ 是克罗内克函数，当蛋白质节点 v_i 和 v_j 在同一个模块中时取值为 1，否则为 0；m 是加权图中所有边的权重之和。

对于超图 $\mathcal{G}$，将每个蛋白质节点独立为单一社区。然后，对于每个蛋白质节点 v_i，临时从所属社区“摘除”，并计算其加入到每个候选社区的模块度 Q^{hyp}。当 Q^{hyp} 增加时，将节点 v_i 移入对应社区，否则将其放回原社区。在迭代过程中，当在某一层超图上对所有节点进行一次移动之后，若没有任何节点的模块度增益 $\Delta Q^{\text{hyp}}>0$ 时，则该层视为收敛。

分层压缩过程分为两个步骤：模块度优化和模块聚合。首先，通过超图模块度表达式 Q^{hyp}，将超图 $\mathcal{G}$ 划分为多个社区，此时将每个社区看作一个“超级节点”，并重新构建一个新的较小超图 $\mathcal{G}_1$。然后，将模块度优化重新应用于超图 $\mathcal{G}_1$，重复迭代，实现分层压缩，得到一系列多级压缩的超图网络 $\mathcal{G}_{\text{all}}=\{\mathcal{G}_1,\mathcal{G}_2,\cdots,\mathcal{G}_{he}\}$。此时可以从高阶超图的视角分析局部和全局拓扑信息，学习蛋白质节点在不同尺度下的嵌入表示。

低阶 PPI 网络上的分层压缩是将原始 PPI 网络转换为度均大于 2 的子图，并用 Louvain 算法压缩成多级较小的 PPI 网络，定义如式(12)所示：

$$Q^{\text{low}}=\frac{1}{2m}\sum_{ij}\left(w_{ij}-\frac{k_ik_j}{2m}\right)\delta(v_i,v_j) \tag{12}$$

其中，$w_{ij}=1/dr_i\cdot dr_j$ 表示在子图上蛋白质节点 v_i 和 v_j 之间边的权重，dr_i 和 dr_j 表示蛋白质节点 v_i 和 v_j 的度，$k_i=\sum_j w_{ij}$ 是在子图上与蛋白质节点 v_i 相邻边的权重和，m 是子图中边的总和。同样地，通过模块度优化和模块聚合，将子图划分为多级较小的网络，得到 $G_{\text{all}}=\{G_1,G_2,\cdots,G_{oe}\}$。

3.3　节点嵌入

将超图卷积分别应用于超图 $\mathcal{G}_{\text{all}}=\{\mathcal{G}_1,\mathcal{G}_2,\cdots,\mathcal{G}_{he}\}$，学习蛋白质节点嵌入表示。原始拓扑图仅考虑一阶邻近信息，不足以捕捉节点之间的复杂关系，导致构建出一个稀疏的邻接

矩阵，不利于节点在超图中的有效嵌入学习。为了更好地整合蛋白质节点之间的潜在联系，构造了一个属性相似性图。它可以发现节点之间的隐含关系，有助于提升超图卷积的表现。属性相似图的邻接矩阵 $\boldsymbol{S}$ 由节点相似性 S_{ij} 构成，S_{ij} 和原始 PPI 网络的邻接矩阵 $\boldsymbol{A}$ 的定义如式(13)和式(14)所示：

$$S_{ij}=x_i^{\mathrm{T}}x_j \tag{13}$$

$$A_{ij}=\begin{cases}1, & v_i,v_j\in E_{\mathrm{ppi}}\\ 0, & v_i,v_j\notin E_{\mathrm{ppi}}\end{cases} \tag{14}$$

其中，$x_i=\sum_{j\in V_{\mathrm{ppi}}}A_{ij}$。

虽然属性相似图邻接矩阵 $\boldsymbol{S}\in R^{n*n}$ 可以度量节点之间潜在连接的强度，但它可能会产生一些额外的边，导致一些不相关节点的连接[36]。这里根据文章进行改进，如式(15)所示：

$$\hat{S}=\begin{cases}S_{ij}, & S_{ij}\geqslant \min\limits_{v_k\in N(v_i)}S_{ik}\\ 0, & S_{ij}<\min\limits_{v_k\in N(v_i)}S_{ik}\end{cases} \tag{15}$$

其中，$N(v_i)$表示节点 v_i 的邻居节点集合；$\min\limits_{v_k\in N(v_i)}S_{ik}$ 表示节点 v_i 与其所有邻居节点的相似度中的最小值。当节点 v_i 和 v_j 的相似度 S_{ij} 大于等于节点 v_i 与所有邻居相似度中最小值时，保留 S_{ij}，否则取为 0。

通过属性相似图，得到增强的原始 PPI 网络的邻接矩阵 $\hat{\boldsymbol{A}}$，其中矩阵中的值 $\hat{A}_{ij}$ 定义如式(16)所示：

$$\hat{A}_{ij}=A_{ij}+\mu\hat{S}_{ij} \tag{16}$$

其中，A_{ij} 是原始 PPI 网络中节点 v_i 和 v_j 对应邻接矩阵的值，μ 是一个平衡超参数。

使用图卷积得到节点特征 $\boldsymbol{\varphi}^{(1)}$，并将其作为第一层超图卷积的节点特征，定义如式(17)所示：

$$\boldsymbol{\varphi}^{(1)}=(\boldsymbol{D}^{-\frac{1}{2}}\hat{\boldsymbol{A}}\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{\varphi}^{(0)}\boldsymbol{W}_{\mathrm{ppi}}^{(0)}) \tag{17}$$

其中，$\boldsymbol{D}$ 是原始 PPI 网络上的对角度矩阵，对角元素 $D_{ii}=\sum_{j=1}^{n}\hat{A}_{ij}$；$\boldsymbol{W}_{\mathrm{ppi}}^{(0)}$ 是原始 PPI 网络上节点的权重矩阵，对角元素 $\omega(v_i)=\sum_{j=1}^{n}A_{ij}$；$\boldsymbol{\varphi}^{(0)}$ 是原始 PPI 网络中节点的邻接矩阵 $\boldsymbol{A}$。

使用 ReLU 作为非线性激活函数，构建一个两层超图卷积。通过第一层超图卷积得到节点的直接超边邻居信息，第二次超图卷积进一步传播节点信息到更远的邻居，得到更全面的节点嵌入表示。两层超图卷积的定义如式(18)和式(19)所示：

$$\boldsymbol{Z}^{(1)}=\mathrm{ReLU}(\boldsymbol{D}_v^{-\frac{1}{2}}\boldsymbol{H}\boldsymbol{W}\boldsymbol{D}_e^{-1}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{D}_v^{-\frac{1}{2}}\boldsymbol{\varphi}^{(1)}\boldsymbol{\theta}) \tag{18}$$

$$\boldsymbol{Z}^{(2)}=\mathrm{ReLU}(\boldsymbol{D}_v^{-\frac{1}{2}}\boldsymbol{H}\boldsymbol{W}\boldsymbol{D}_e^{-1}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{D}_v^{-\frac{1}{2}}\boldsymbol{Z}^{(1)}\boldsymbol{\theta}) \tag{19}$$

其中，$\boldsymbol{\varphi}^{(1)}$ 是第一层图卷积输出的节点表示；$\boldsymbol{D}_v$ 是节点度矩阵；$\boldsymbol{\theta}$ 是超图卷积的权重参数矩阵，通过正态分布随机生成。

将第二次超图卷积 $\boldsymbol{Z}^{(2)}$ 作为对应节点嵌入表示，并将两次超图卷积应用于每个超图 $\mathcal{G}_{\mathrm{all}}=\{\mathcal{G}_1,\mathcal{G}_2,\cdots,\mathcal{G}_{he}\}$，得到每个超图中蛋白质节点嵌入表示矩阵 $\mathcal{GZ}_{\mathrm{all}}=\{\mathcal{GZ}^1,\mathcal{GZ}^2,\cdots,\mathcal{GZ}^{he}\}$。

若第 he 个超图中的节点 v_i 为“超级节点”，则需要将“超级节点”u 的嵌入表示分配给对应模块中的节点。假设第 he 个超图中，超级节点 u 所包含的节点集合为 $C(u)$，此时将“超级节点”u 的嵌入表示 $\mathcal{GZ}_u^{he}$ 按照原始节点权重，分配至其包含的每个节点，得到节点 v_i 在该超图中的嵌入表示：

$$\boldsymbol{DG}_{c_i^{he}}^{he}(v_i)=\omega(v_i)\cdot\mathcal{GZ}_u^{he} \tag{20}$$

其中，$\omega(v_i)$表示在原始 PPI 网络上节点 $v_i\in C(u)$的权重。将每个节点在不同超图中得到的嵌入表示进行拼接，得到超图上节点嵌入表示，如式(21)所示：

$$\boldsymbol{HD}(v_i)=[\boldsymbol{DG}_{c_i^1}^{1}(v_i),\boldsymbol{DG}_{c_i^2}^{2}(v_i),\cdots,\boldsymbol{DG}_{c_i^{he}}^{he}(v_i)] \tag{21}$$

其中，$\boldsymbol{HD}(v_i)$表示节点 v_i 在超图上的嵌入表示，$\boldsymbol{DG}_{c_i^{he}}^{he}(v_i)$是节点 v_i 第 he 个嵌入表示，C_i^{he} 是节点 v_i 所属的第 he 层。

在原始图中的节点嵌入阶段，将 DeepWalk 随机游走和 Word2Vec 模型应用于 $G_{\mathrm{all}}=\{G_1,G_2,\cdots,G_{oe}\}$。对每个层次使用 DeepWalk 随机游走生成节点序列，用于 Word2Vec 模型，生成相应的蛋白质嵌入表示 $\boldsymbol{GZ}_{\mathrm{all}}=\{GZ^1,GZ^2,\cdots,GZ^{oe}\}$。同样地，如果该节点为“超级节点”，则对应模块中节点的嵌入表示计算与超图一致。若子图中节点 v_i 在原 PPI 网络中与其他节点 v_j 有连边，则节点 v_j 的嵌入表示与节点 v_i 一致。此时得到节点在 PPI 网络中的嵌入表示 $\boldsymbol{OD}(v_i)=[DG_{c_i^1}^{1}(v_i),DG_{c_i^2}^{2},\cdots,DG_{c_i^{oe}}^{oe}(v_i)]$，并将这些嵌入表示连接起来得到完整的嵌入表示。蛋白质节点 v_i 的最终嵌入表示 $\boldsymbol{FD}(v_i)$的定义如式(22)所示：

$$\boldsymbol{FD}(v_i)=[\boldsymbol{DG}_{c_i^1}^{1}(v_i),\boldsymbol{DG}_{c_i^2}^{2}(v_i),\cdots,\boldsymbol{DG}_{c_i^{he}}^{he}(v_i),DG_{c_i^1}^{1}(v_i),DG_{c_i^2}^{2},\cdots,DG_{c_i^{oe}}^{oe}(v_i)] \tag{22}$$

3.4 节点聚类

在聚类生成阶段，首先基于嵌入表示计算蛋白质间的余弦相似度来构建一个加权 PPI 网络 G_w。给定蛋白质 v_i 和蛋白质 v_j 的嵌入向量 $\boldsymbol{FD}(v_i)=[v_{i_1},\cdots,v_{i_{fd}}]$ 和 $\boldsymbol{FD}(v_j)=[v_{j_1},\cdots,v_{j_{fd}}]$，它们之间的余弦相似度定义如式(23)所示：

$$\begin{aligned}cosine(v_i,v_j)&=\frac{\boldsymbol{FD}(v_i)\cdot\boldsymbol{FD}(v_j)}{\|\boldsymbol{FD}(v_i)\|\times\|\boldsymbol{FD}(v_j)\|}\\&=\frac{\sum_{c=1}^{fd}(v_{i_c}*v_{j_c})}{\sqrt{\sum_{c=1}^{fd}(v_{i_c})^2}*\sqrt{\sum_{c=1}^{fd}(v_{j_c})^2}}\end{aligned} \tag{23}$$

采用基于核心附属策略从加权 PPI 网络 G_w 中识别蛋白质复合体。核心附属策略包括两个步骤：寻找核心蛋白质和添加附属蛋白质。通过深度优先搜索算法生成所有的极大团(每个极大团包含超过 2 个蛋白质节点，且与其他极大团不重叠)作为候选核心集($CCSet$)。候选核心集 $CCSet$ 中的所有团根据其密度值降序排列，密度值的定义如式(24)所示：

$$density(CC_k)=\sum_{v_i,v_j\in CC_k}cosine(v_i,v_j) \tag{24}$$

其中，$CC_k\in CCSet$ 表示候选核心集 $CCSet=\{CC_1,CC_2,\cdots,CC_p\}$中的第 p 个候选核心。

从候选核心集 $CCSet$ 中选择合适的种子核心 $SCSet$，将 CC_1 移动到核心集 $SCSet$ 中。对于 $CCSet$ 中的任一其他团 CC_i，如果 CC_i 与 CC_1 有相同的节点，并且 $|CC_1-CC_i|\geqslant 3$，则 CC_i 通过 CC_i-CC_1 进行更新；否则，CC_i 从候选核心集 $CCSet$ 中删除。迭代，直到 $CCSet$ 为空，得到 $SCSet$ 为种子核心。

选择合适的附属蛋白质添加到核心蛋白质中。附属蛋白质的选择标准是通过计算蛋白质节点 v_i 与种子核心 $SC_j\in SCSet$ 之间的连接度，连接度定义如式(25)所示：

$$connectivity(v_i,SC_j)=\frac{\sum_{k\in SC_j}cosine_{ik}}{|SC_j|},v_i\notin SC_j \tag{25}$$

当 $connectivity(v_i,SC_j)>\lambda$ 时，将蛋白质节点 v_i 作为种子核心 SC_j 的附属蛋白质。当网络中所有节点都划分到所属类中，则算法结束，从而得到预测蛋白质复合体 PPC。算法

PCIHNE 的详细流程如算法 1 所示。

算法 1　PCIHNE 算法

输入：PPI 网络 G，扩展阈值 λ，嵌入维度 d

输出：预测蛋白质复合体集合 PPC

1. 初始化 PPC=∅，CCSet=∅ 和 SCSet=∅；
2. 根据式(1)和式(2)将一个 PPI 网络 G 转换为超图$\mathcal{G}$；
3. 根据式(11)，对超图$\mathcal{G}$进行模块度优化和模块聚合得到$\mathcal{G}_{all}=\{\mathcal{G}_1,\mathcal{G}_2,\cdots,\mathcal{G}_{he}\}$；
4. 根据式(12)，对 PPI 网络对应子图进行模块度优化和模块聚合得到$G_{all}=\{G_1,G_2,\cdots,G_{oe}\}$；
5. 每个超图$\mathcal{G}_{all}=\{\mathcal{G}_1,\mathcal{G}_2,\cdots,\mathcal{G}_{he}\}$使用式(18)和式(19)进行两次超图卷积得到对应超图中的节点嵌入表示$\mathcal{GZ}_{all}=\{\mathcal{GZ}^1,\mathcal{GZ}^2,\cdots,\mathcal{G}Z^{he}\}$；
6. 每个低阶图$G_{all}=\{G_1,G_2,\cdots,G_{oe}\}$使用 DeepWalk 和 Word2Vec 模型，得到$\mathcal{GZ}_{all}=\{GZ^1,GZ^2,\cdots,GZ^{oe}\}$；
7. 根据式(20)将$\mathcal{GZ}_{all}$和$\mathbf{GZ}_{all}$的节点嵌入表示进行转化，得到每个节点对应的嵌入表示 $HD(v_i)=[D\mathcal{G}_{c_i^1}^1(v_i),\cdots,D\mathcal{G}_{c_i^{he}}^{he}(v_i)]$和 $OD(v_i)=[DG_{c_i^1}^1(v_i),\cdots,DG_{c_i^{oe}}^{oe}(v_i)]$；
8. 合并 $HD(v_i)$和 $OD(v_i)$，得到原始 PPI 网络中每个节点最终的嵌入表示 $FD(v_i)=[D\mathcal{G}_{c_i^1}^1(v_i),D\mathcal{G}_{c_i^2}^2(v_i),\cdots,D\mathcal{G}_{c_i^{he}}^{he}(v_i),DG_{c_i^1}^1(v_i),DG_{c_i^2}^2(v_i),\cdots,DG_{c_i^{oe}}^{oe}(v_i)]$；
9. 根据式(23)构建加权 PPI 网络 G_w；
10. 使用深度优先搜索算法生成极大团作为候选核心集 CCSet；
11. 将候选核心集 CCSet 通过式(24)进行降序排序，得到 $CCSet=\{CC_1,CC_2,\cdots,CC_p\}$；
12. 从候选核心集 CCSet 选择合适的种子核心并将其移动到 $SCSet=\{SC_1,SC_2,\cdots,SC_{pc}\}$；
13. 选择合适的附属蛋白添加到核心蛋白质中；
14. 通过式(25)计算蛋白质节点 v_i与种子核心 SC_j之间的连接度 connectivity；
15. 当 $connectivity(v_i,SC_j)>\lambda$时，将蛋白质节点 v_i作为种子核心 SC_j的附属蛋白质，否则执行步骤 18。
    $SC_j=SC_j\cup\{v_i\}$；
17. 重复步骤 14 和步骤 15，直到所有蛋白质节点均划分到种子核心 SC_j；
18. $PPC=PPC\cup\{SC_j\}$；
19. 重复步骤 14－步骤 18，直到核心集 CCSet=∅；
20. 返回 PPC。

3.5　算法时间复杂度分析

本文算法整体时间复杂度主要由超图构建、分层压缩、节点嵌入及节点聚类 4 部分组成。设 n，E，L 和 d 分别是 PPI 网络的节点、边集、分成压缩的层次和嵌入维度。具体而言，算法首先将原始网络映射为超图，时间复杂度为 $O(n+|E|)$。其次，在分层压缩阶段，算法基于超图模块度的优化和聚合，对每一层超图进行社区划分，并将划分出的每个社区聚合为一个“超级节点”。若第 i 层边数为 $|E_i|$，则该层聚合及重构的时间复杂度为 $\sum_{i=0}^{L-1}O(|E_i|)$。若不采用分层压缩，则每层操作都在完整的图结构上执行，时间复杂度为 $O(|E|(1+d))$。通过分层压缩策略，边数 $|E_i|$ 随层数快速衰减，设平均每层压缩因子为 $c>1$，则有 $\sum_{i=0}^{L-1}|E_i|\approx|E|(1+c^{-1}+\cdots)=O(|E|)$。所以，分层压缩后整个算法的边级操作由原来的 $O(|E|(1+d))$变为 $O(L\cdot|E_{avg}|(1+d))$，$|E_{avg}|\ll|E|$。再次，在节点嵌入阶段，对每层压缩后的超图进行超图卷积，每次的时间复杂度为 $O(|E_i|\cdot d)$，故 L 层的时间复杂度为 $\sum_{i=0}^{L-1}O(|E_i\cdot d|)$。最后，节点聚类阶段，在最终的 PPI 加权图上使用核心-附属策略，其时间复杂度为 $O(n+|E|)$。

4　实验结果与分析

4.1　数据集

实验使用了酿酒酵母蛋白质相互作用网络 Gavin1[37]，Gavin2[38]，K_extend[39]，DIP[40]，BioGRID[41]，以及人类蛋白质互作用网络数据 STRING[42]。以 CYC2008[43] 和 MIPS[44] 作为酿酒酵母蛋白质复合体的金标准，CORUN[45] 作为人类蛋白质复合体的金标准，进行参数分析和聚类结果评估。表 1 列出了 PPI 网络的具体信息，这些数据集均去除了所有自相交和重复相交。

表 1　PPI 数据集信息

Table 1　PPI dataset informations

数据集	节点	连边
Gavin1	1352	3210
Gavin2	1430	6531
K_extend	3672	14317
DIP	4930	17201
BioGRID	4187	20454
STRING	9477	77295

4.2　评价指标

为了评估本文方法的性能，使用了 F-measure 和 Accuracy 作为评价指标[46]。F-measure 基于测试的精确率和召回率进行计算，Accuracy 则是灵敏度和阳性预测值的几何平均值。根据已有研究[16,20,47]，F-measure 和 Accuracy 能够反映算法在不同维度上的表现。F-measure 体现算法对预测结果精度与召回率的综合平衡，Accuracy 则能够直观反映算法预测结果与真实复合体之间的整体吻合程度。将这两个指标组合起来，有助于更加客观、全面地评估算法性能。本文将 F-measure＋Accuracy 作为综合评价指标，记为 F1＋ACC。

4.3　参数分析

本文对扩展阈值 $\lambda\in[0.1,1]$和嵌入维度 $d\in(2^4,2^{10})$这两个主要参数对算法聚类结果的影响进行分析。BioGRID 是一个低通量数据集，并且具有高可靠性和高准确性。使用 BioGRID 数据集，以最大化 F1＋ACC 值作为目标，使用网格搜索方法进行参数分析。

如图 2 所示，对于扩展阈值 λ，当 $\lambda=0.4$ 时，F1＋ACC 值最大。对于嵌入维度 d，当 $d=512$ 时，F1＋ACC 值达到最大。为了评估参数对预测结果的影响，对 λ 和 d 进行了敏感性分析。比较不同参数组合下的 F1＋ACC 值，当 $\lambda=0.4$，$d=512$ 时，该算法在 BioGRID 数据集上取得了最好的性能。此外，在 $\lambda\in[0.3,0.5]$和 $d\in[256,640]$区间内，F1＋ACC 值与最优值的差异很小，说明 PCIHNE 算法在该参数范围内取值均能获得较好的性能。所选参数虽然在其他部分数据集上未达到最优性能，但整体上仍能在所有数据集上实现优于对比算法的性能，实验结果如图 3 和图 4 所示。因此，该参数组合被选为统一的参数设置。

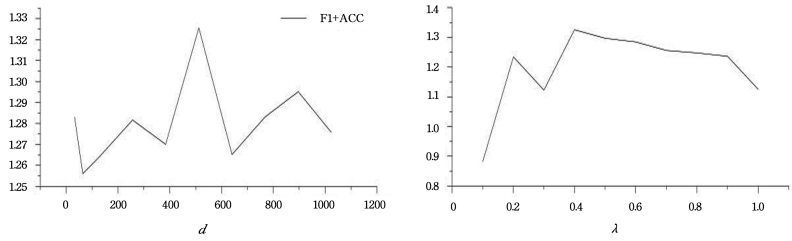


图 2 参数 λ 和 d 在 BioGRID 数据集上的聚类结果影响

Fig. 2 Impact of clustering results parameters λ and d on the BioGRID dataset

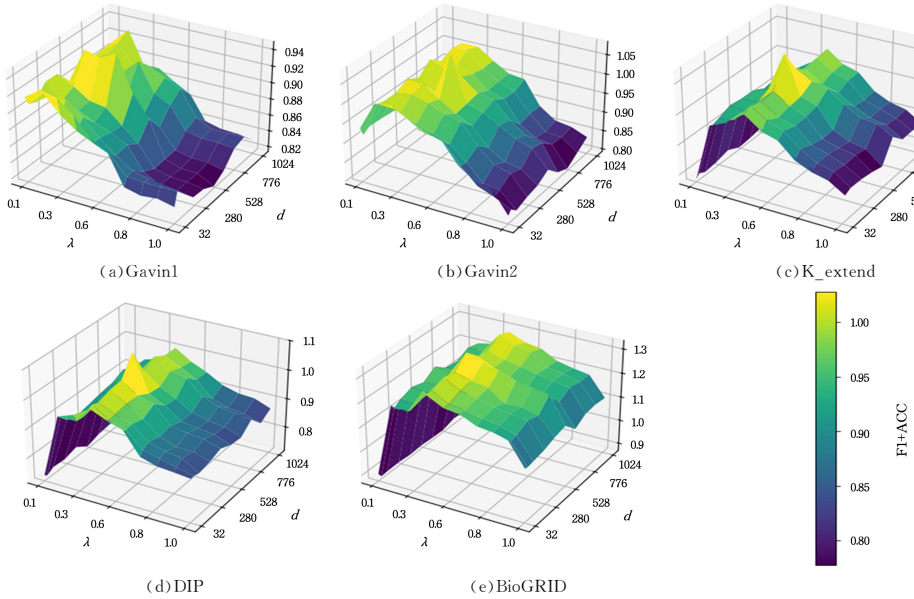


图 3 参数 λ 和 d 在不同数据集上的聚类结果影响(以 CYC2008 作为金标准)

Fig. 3 Impact of clustering results parameters λ and d on different datasets(CYC2008 as benchmarks)

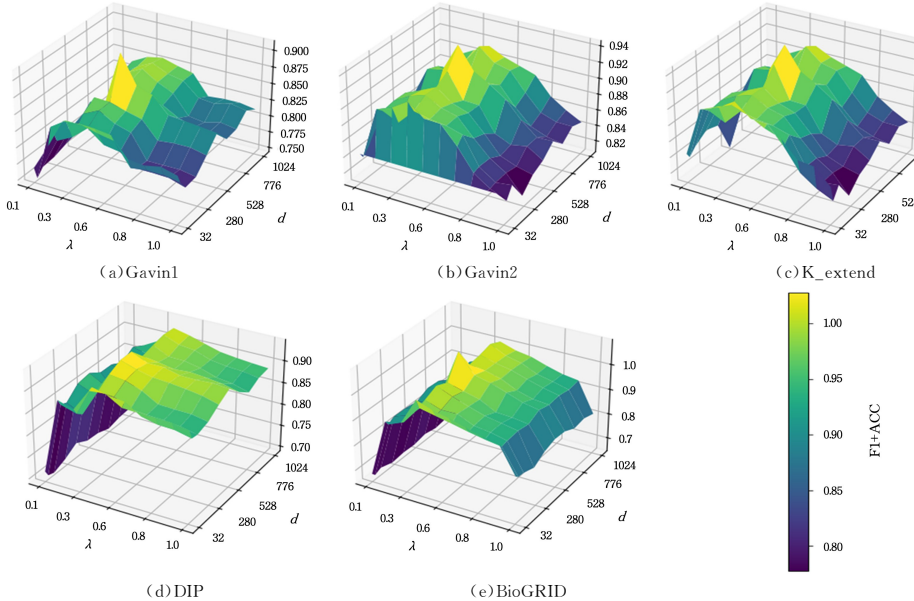


图 4 参数 λ 和 d 在不同数据集上的聚类结果影响(以 MIPS 作为金标准)

Fig. 4 Impact of clustering results parameters λ and d on different datasets(MIPS as benchmarks)

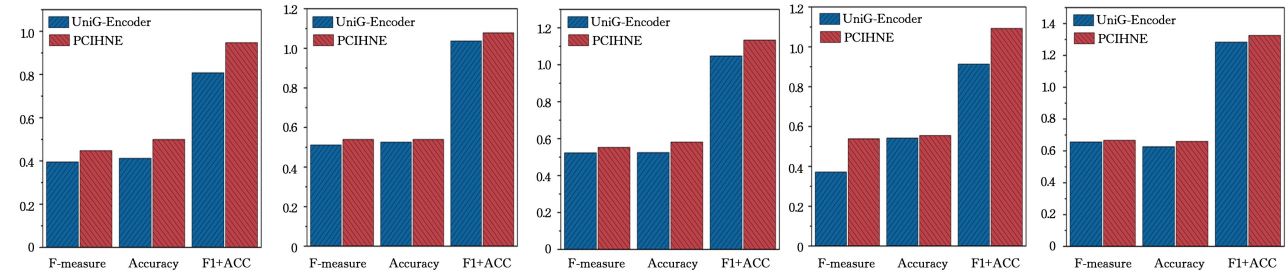
4.4 超图网络嵌入实验分析

本文 PCIHNE 算法中使用了超图网络嵌入表示方法,为了评估该表示方法的有效性,将其与超图神经网络算法 UniG-Encoder[48] 和 HyperGT[49] 进行对比分析。

UniG-Encoder 是基于全局超图自编码器框架的方法,对整个超图进行编码与解码,重构超边结构依次捕捉全局关系。GyperGT 是基于 Transformer 的超图注意力网络,通过对超边内节点进行自注意力聚合以捕捉超边局部关系。

图 5 展示了 CYC2008 作为蛋白质复合体金标准,PCIHNE 算法和 UniG-Encoder 算法在 Gavin1, Gavin2,

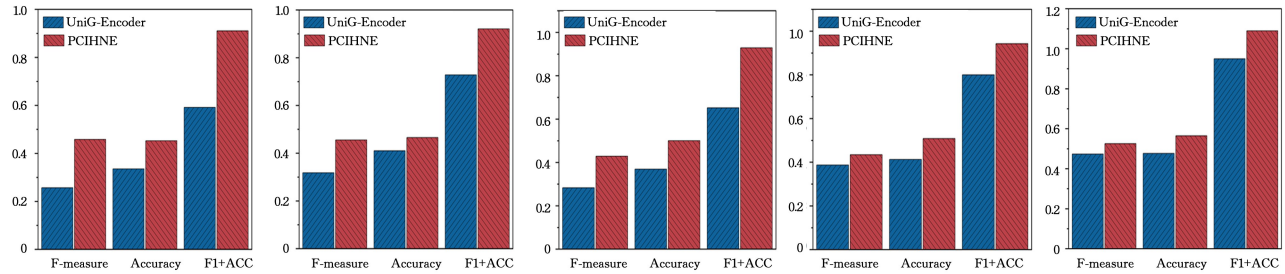
K_extend，DIP 和 BioGRID 数据集上的实验结果。实验结果表明，PCIHNE 算法的 F-measure 和 Accuracy 在所有数据集上均优于 UniG-Encoder 算法，F1＋ACC 值平均提高了 9.77%。图 6 展示了 MIPS 作为蛋白质复合体金标准，PCIHNE 算法和 UniG-Encoder 算法在 Gavin1，Gavin2，K_extend，DIP 和 BioGRID 数据集上的实验结果。同样地，PCIHNE 算法的 F-measure 和 Accuracy 在所有数据集上均优于 UniG-Encoder 算法，F1＋ACC 值平均提高了 21.44%。



(a)Gavin1　(b)Gavin2　(c)K_extend　(d)DIP　(e)BioGRID

图 5　PCIHNE 算法和 UniG-Encoder 算法的评估结果(以 CYC2008 作为金标准)

Fig. 5　Evaluation results of PCIHNE algorithm and UniG-Encoder algorithm(CYC2008 as benchmarks)

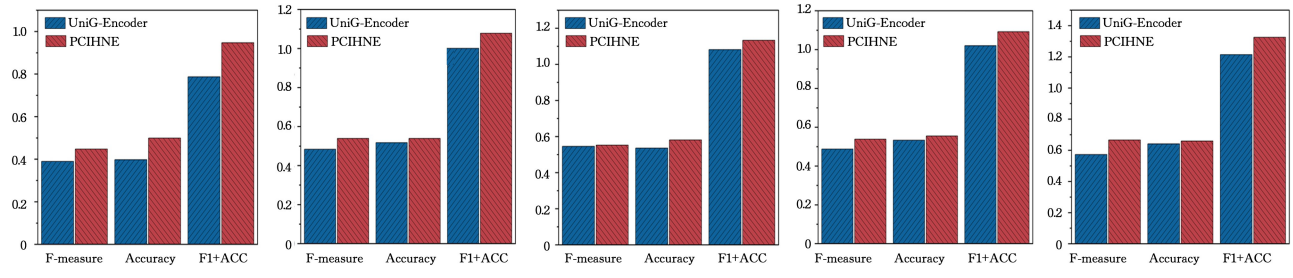


(a)Gavin1　(b)Gavin2　(c)K_extend　(d)DIP　(e)BioGRID

图 6　PCIHNE 算法和 UniG-Encoder 算法的评估结果(以 MIPS 作为金标准)

Fig. 6　Evaluation results of PCIHNE algorithm and UniG-Encoder algorithm(MIPS as benchmarks)

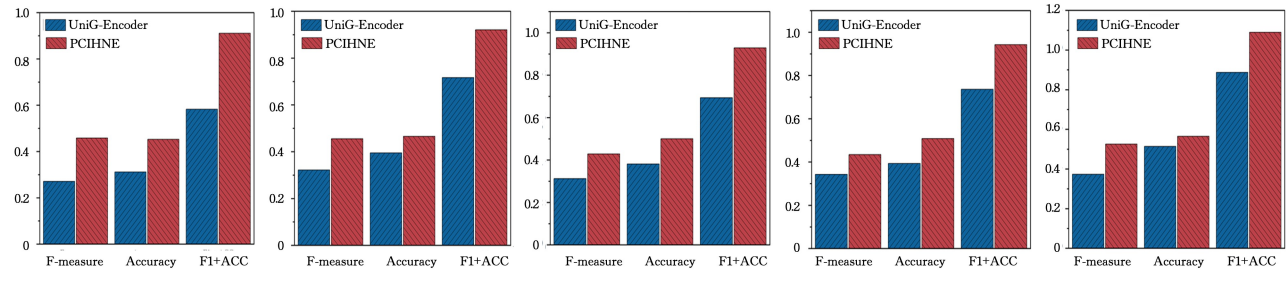
图 7 展示了 CYC2008 作为蛋白质复合体金标准，PCIHNE 算法和 HyperGT 算法在 Gavin1，Gavin2，K_extend，DIP 和 BioGRID 数据集上的实验结果。结果表明，PCIHNE 算法的 Accuracy 和 F-measure 在所有数据集上均优于 HyperGT 算法，F1＋ACC 值平均提高了 9.43%。图 8 展示了 MIPS 作为蛋白质复合体金标准，PCIHNE 算法和 HyperGT 算法在 Gavin1，Gavin2，K_extend，DIP 和 BioGRID 数据集上的实验结果。同样地，PCIHNE 算法的 Accuracy 和 F-measure 在所有数据集均上优于 HyperGT 算法，F1＋ACC 值平均提高了 23.53%。



(a)Gavin1　(b)Gavin2　(c)K_extend　(d)DIP　(e)BioGRID

图 7　PCIHNE 算法和 HyperGT 算法的评估结果(以 CYC2008 作为金标准)

Fig. 7　Evaluation results of PCIHNE algorithm and HyperGT algorithm(CYC2008 as benchmarks)



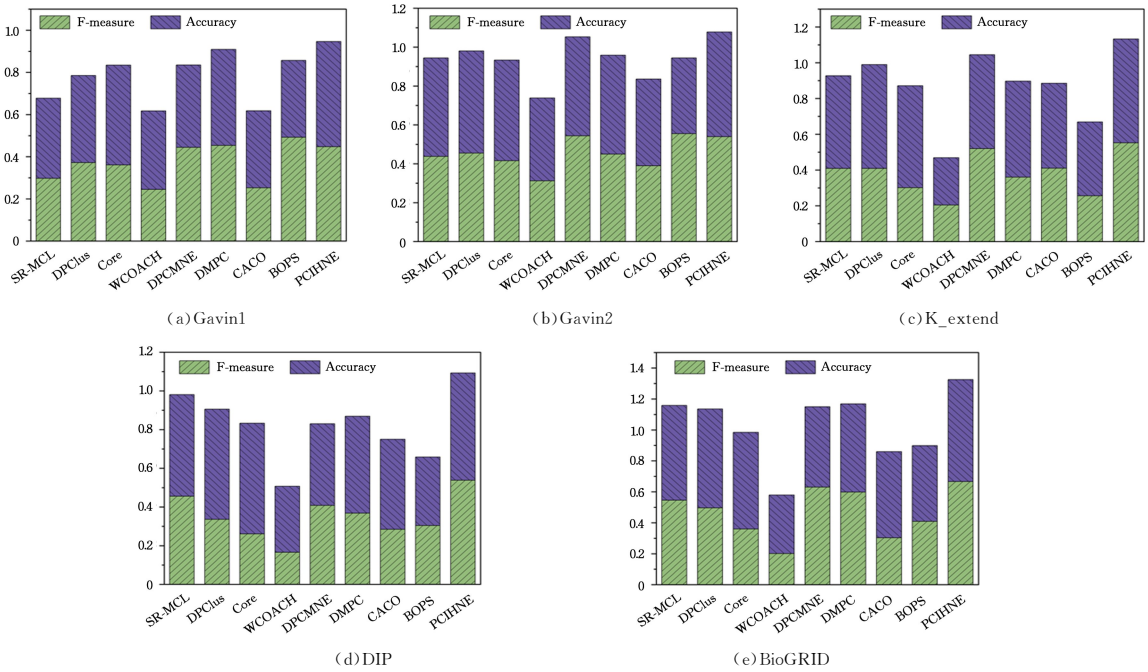
(a)Gavin1　(b)Gavin2　(c)K_extend　(d)DIP　(e)BioGRID

图 8　PCIHNE 算法和 HyperGT 算法的评估结果(以 MIPS 作为金标准)

Fig. 8　Evaluation results of PCIHNE algorithm and HyperGT algorithm(MIPS as benchmarks)

4.5 对比实验

为了全面评估 PCIHNE 算法的性能，将其与蛋白质复合体识别算法 SR-MCL，DPClus，Core，WCOACH，DPCMNE，DMPC，CACO 和 BOPS 进行比较。图 9 展示了 CYC2008 作为蛋白质复合体金标准，在 Gavin1，Gavin2，K_extend，DIP 和 BioGRID 数据集上的实验结果。实验结果表明，PCIHNE 算法的 Accuracy 在所有数据集上均优于其他算法，F-measure 在 Gavin1，K_extend，DIP 和 BioGRID 数据集上优于其他算法，在 Gavin2 数据集上也接近最大值。与其他算法相比，PCIHNE 算法在 Gavin1，Gavin2，K_extend，DIP 和 BioGRID 数据集上的 F1＋ACC 值分别平均提高了 23.49%，16.72%，34.25%，37.94%和 33.64%，取得了最好的性能。



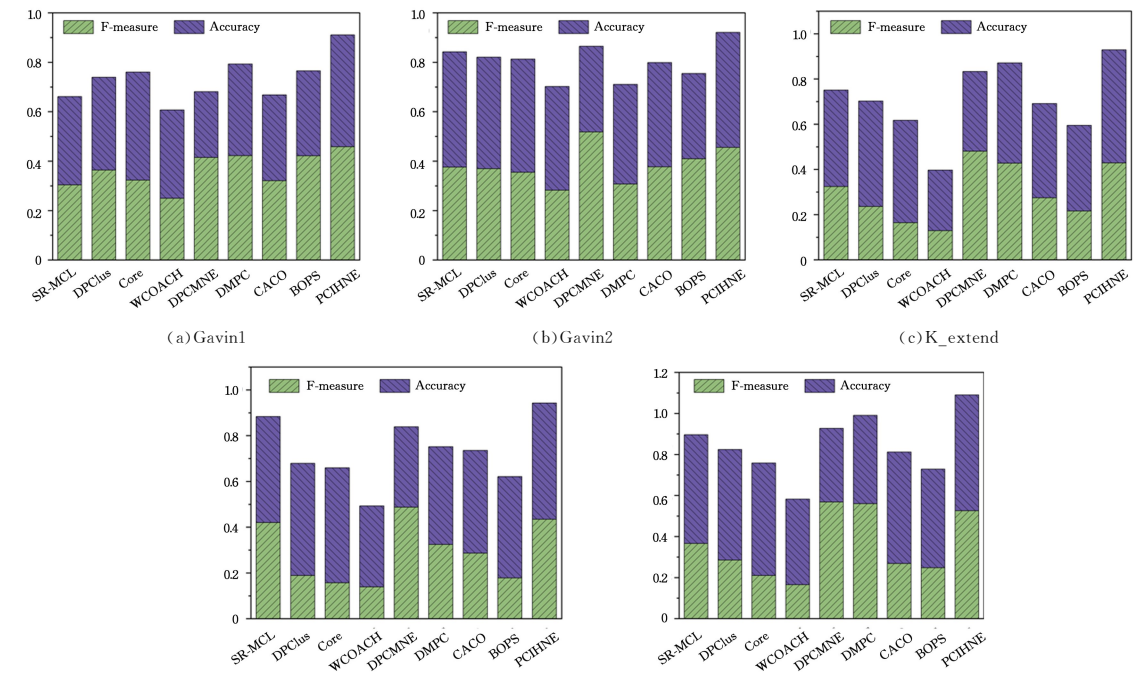
(a)Gavin1 (b)Gavin2 (c)K_extend

(d)DIP (e)BioGRID

图 9 不同算法在数据集的评估结果(以 CYC2008 作为金标准)

Fig. 9 Evaluation results by different algorithms(CYC2008 as benchmarks)

图 10 展示了使用 MIPS 作为蛋白质复合体金标准的实验结果。实验结果表明，PCIHNE 算法的 Accuracy 在所有数据集上均优于其他算法，F-measure 在大多数数据集上表现出更好或相当的性能。PCIHNE 算法在 Gavin1，Gavin2，K_extend，DIP 和 BioGRID 数据集上的 F1＋ACC 值最大，分别平均提高了 28.39%，16.85%，36.18%，33.17%和 33.73%。



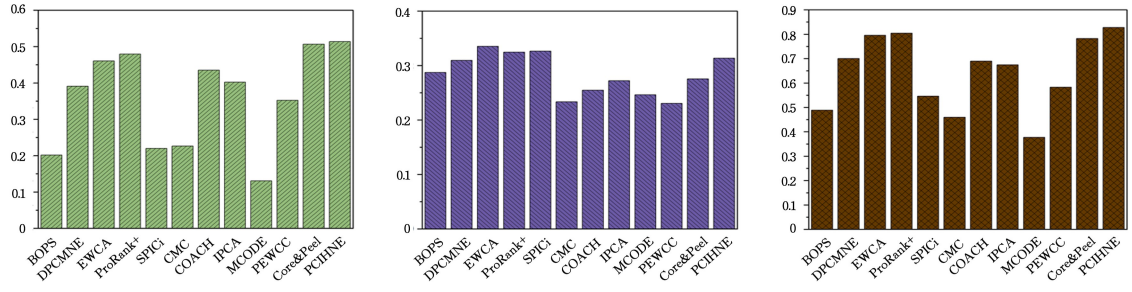
(a)Gavin1 (b)Gavin2 (c)K_extend

(d)DIP (e)BioGRID

图 10 不同算法在数据集的评估结果(以 MIPS 作为金标准)

Fig. 10 Evaluation results by different algorithms(MIPS as benchmarks)

以 CORUM 作为蛋白质复合体金标准，在人类 PPI 数据集 STRING 上，将 PCIHNE 算法与 BOPS，DPCMNE，EWCA[50]，ProRank＋[51]，SPICi[52]，CMC[53]，COACH，IPCA[54]，MCODE[55]，PEWCC[56] 和 Core&Peel[57] 算法进行对比测试，实验结果如图 11 所示。可以看出，PCIHNE 算法在人类 PPI 数据集上同样表现优异。PCIHNE 算法的 F-measure 优于所有对比算法，Accuracy 优于 BOPS，DPCMNE，CMC，COACH，IPCA，MCODE，PEWCC 和 Core&Peel 算法，与 EWCA，ProRank＋和 SPICi 算法取得了相当的水平，F1＋ACC 值在所有数据集上始终处于最高水平。与其他算法相比，PCIHNE 算法在 F-measure，Accuracy 和 F1＋ACC 上分别平均提升 77.36%，13.28% 和 39.57%，在人类数据集上取得了很好的蛋白质复合体识别性能。



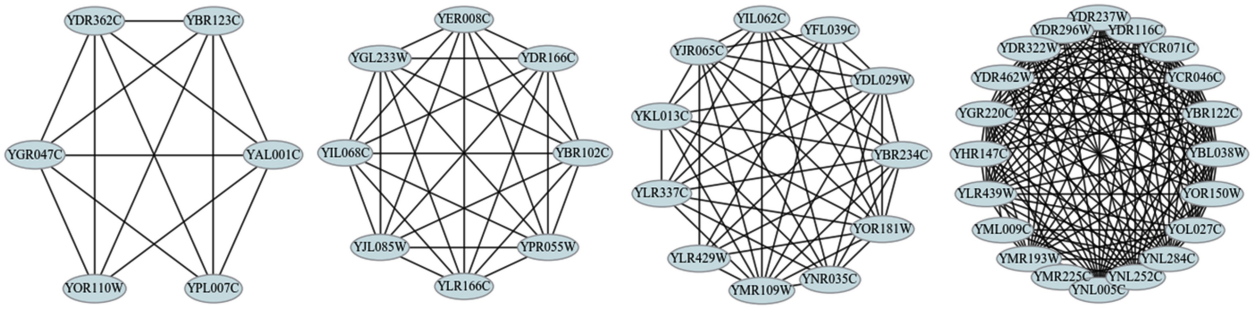
(a) F-measure　(b) Accuracy　(c) F1＋ACC

图 11　不同的算法在 STRING 数据集的评估结果（以 CORUM 作为金标准）

Fig. 11　Evaluation results by different algorithms on STRING dataset (CORUM as benchmarks)

4.6　预测复合体功能意义分析

为了验证 PCIHNE 算法的有效性，本文对预测到的蛋白质复合体进行了 GO 功能富集，并通过以下 4 个例子进行分析，如图 12 所示。图 12(a)和图 12(b)展示了 2 个预测的蛋白质复合体，它们与金标准中的蛋白质复合体完全匹配。图 12(c)和图 12(d)展示了 2 个未能与金标准数据匹配的预测复合体。表 2 列出了图 12(c)和图 12(d)所示示例的基因本体注释和显著性 p 值。通过 GO 功能富集分析，预测的蛋白质复合体(c)和(d)在生物过程(BP)、分子功能(MF)和细胞组分(CC)下均展现出显著的功能富集，具有显著的生物学意义，可能是潜在的蛋白质复合体。这表明 PCIHNE 算法具有优异的蛋白质复合体识别能力。



(a)　(b)　(c)　(d)

图 12　预测到的 4 个蛋白质复合体例子

Fig. 12　Four examples of predicted protein complexes

表 2　预测到 4 个蛋白质复合体的 GO 功能注释

Table 2　Predicted gene ontology functional annotations for four protein complexes

ID	预测的蛋白质复合体	生物过程(BP)		分子功能(MF)		细胞组分(CC)	
		GO Term	*P*-Value	GO Term	*P*-Value	GO Term	*P*-Value
a	YBR123C，YAL001C，YGR047C，YDR362C，YOR110W，YPL007C	5S class rRNA transcription by RNA polymerase III GO:0042791	2.33×10^{-14}	DNA binding GO:0003677	1.38×10^{-3}	transcription factor TFIIIC complex GO:0000127	3.76×10^{-15}
b	YJL085W，YBR102C，YLR166C，YGL233W，YER008C，YDR166C，YIL068C，YPR055W	vesicle docking involved in exocytosis GO:0006904	3.05×10^{-17}	small GTPase binding GO:0031267	5.14×10^{-3}	exocyst GO:0000145	1.15×10^{-19}
c	YBR234C，YDL029W，YFL039C，YIL062C，YJR065C，YKL013C，YLR337C，YLR429W，YMR109W，YNR035C，YOR181W	actin nucleation GO:0045010	3.57×10^{-14}	actin binding GO:0003779	2.12×10^{-16}	actin cortical patch GO:0030479	1.48×10^{-19}

（续表）

ID	预测的蛋白质复合体	生物过程(BP)		分子功能(MF)		细胞组分(CC)	
		GO Term	*P*-Value	GO Term	*P*-Value	GO Term	*P*-Value
d	YBL038W, YBR122C, YCR046C, YCR071C, YDR116C, YDR237W, YDR296W, YDR322W, YDR462W, YGR220C, YHR147C, YLR439W, YML009C, YMR193W, YMR225C, YNL005C, YNL252C, YNL284C, YOL027C, YOR150W	mitochondrial translation GO:0032543	1.47×10^{-29}	structural constituent of ribosome GO:0003735	1.34×10^{-22}	mitochondrial large ribosomal subunit GO:0005762	8.64×10^{-36}

结束语 本文提出了一种基于超图网络嵌入的蛋白质复合体识别方法(PCIHNE)。该算法通过超图网络对多元关系的直接建模能力，引入超图和分层压缩策略来建模 PPI 网络。不仅考虑了高阶超图在生物网络上表示的多节点交互关系，也考虑了 PPI 网络的低阶拓扑特征，既融合了局部拓扑信息，也利用了全局拓扑信息。在多个 PPI 网络数据集上进行测试，PCIHNE 算法的 F1＋ACC 在金标准 CYC2008，MIPS 和 CORUM 上平均提高了 29.21%，29.66%和 39.57%。该算法结合超图网络特性，可以显著提高蛋白质复合体识别方法的性能，能够很好地进行蛋白质复合体识别，促进了数据挖掘和生物信息领域的发展。

参 考 文 献

[1] ZHANG Y, JIAK B, ZHANGA D. Consistent protein functional module detection from multi-view of biological data[J]. Acta Electronica Sinica, 2014, 42(12): 2337-2344.

[2] WU Z, WANG Y, CHEN L. Network-based drug repositioning [J]. Molecular BioSystems, 2013, 9(6): 1268-1281.

[3] GÖBL C, MADL T, SIMON B, et al. NMR approaches for structural analysis of multidomain proteins and complexes in solution [J]. Progress in Nuclear Magnetic Resonance Spectroscopy, 2014, 80: 26-63.

[4] WALZTHOENI T, LEITNER A, STENGEL F, et al. Mass spectrometry supported determination of protein complex structure[J]. Current Opinion in Structural Biology, 2013, 23(2): 252-260.

[5] ALBERTS B. The cell as a collection of protein machines: preparing the next generation of molecular biologists[J]. Cell, 1998, 92(3): 291-294.

[6] DUNHAM B, GANAPATHIRAJU M K. Benchmark evaluation of protein-protein interaction prediction algorithms[J]. Molecules, 2021, 27(1): 41.

[7] HUA Y, LI J X, FENG Z H, et al. Protein-drug interaction prediction based on attention feature fusion[J]. Journal of Computer Research and Development, 2022, 59(9): 2051-2065.

[8] CAO H T, CHEN J. Prediction of multitype protein interactions combining Doc2vec and GCN[J]. CAAI Transactions on Intelligent Systems, 2023, 18(6): 1165-1172.

[9] LI Z J, CHEN Y M, LIU J W, et al. A survey of computational method in protein-protein interaction research[J]. Journal of Computer Research and Development, 2008, 45(12): 2129-2137.

[10] PAN Y L, GUAN J H, YAO H, et al. Computational methods for protein complex prediction: A survey[J]. Journal of Frontiers of Computer Science and Technology, 2022, 16(1): 1-20.

[11] GAO Y, FENG Y, JI S, et al. HGNN+: General hypergraph neural networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(3): 3181-3199.

[12] SHANG J L, ZHANG Z Y, QU W W, et al. Survey of graph partitioning techniques for distributed graph computing[J]. Journal of Computer Research and Development, 2025, 62(1): 90-103.

[13] VLASBLOM J, WODAK S J. Markov clustering versus affinity propagation for the partitioning of protein interaction graphs [J]. BMC Bioinformatics, 2009, 10(1): 1-14.

[14] SHIH Y K, PARTHASARATHY S. Identifying functional modules in interaction networks through overlapping Markov clustering[J]. Bioinformatics, 2012, 28(18): 473-479.

[15] LYU J, YAO Z, LIANG B, et al. Small protein complex prediction algorithm based on protein-protein interaction network segmentation[J]. BMC Bioinformatics, 2022, 23(1): 1-20.

[16] WANG C, WANG R, JIANG K. Amethod for detecting overlapping protein complexes based on an adaptive improved FCM clustering algorithm[J]. Mathematics, 2025, 13(2): 196.

[17] NEPUSZ T, YU H, PACCANARO A. Detecting overlapping protein complexes in protein-proteininteraction networks[J]. Nature Methods, 2012, 9(5): 471-472.

[18] ALTAF-UL-AMIN M, SHINBO Y, MIHARA K, et al. Development and implementation of an algorithm for detection of protein complexes in large interaction networks[J]. BMC Bioinformatics, 2006, 7(1): 1-13.

[19] SAHOO T R, PATRA S, VIPSITA S. Decision tree classifier based on topological characteristics of subgraph for the mining of protein complexes from large scale PPI networks[J]. Computational Biology and Chemistry, 2023, 106: 107935.

[20] GAVIN A C, ALOY P, GRANDI P, et al. Proteome survey reveals modularity of the yeast cell machinery[J]. Nature, 2006, 440(7084): 631-636.

[21] LEUNG H C M, XIANG Q, YIU S M, et al. Predicting protein complexes from PPI data: a core-attachment approach[J]. Journal of Computational Biology, 2009, 16(2): 133-144.

[22] WU M, LI X, KWOH C K, et al. A core-attachment based method to detect protein complexes in PPI networks[J]. BMC Bioinformatics, 2009, 10(1): 1-16.

[23] KOUHSAR M, ZARE-MIRAKABAD F, JAMALI Y. WCOACH: protein complex prediction in weighted PPI networks[J]. Genes & Genetic Systems, 2015, 90(5): 317-324.

[24] PENG W, WANG J, ZHAO B, et al. Identification of protein complexes using weighted pagerank-nibble algorithm and core-attachment structure[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2014, 12(1): 179-192.

[25] MUKHOPADHYAY A, RAY S, MAULIK U, et al. Multiobjective approach to protein complex detection[M]// Multiobjective-Optimization Algorithms for Bioinformatics. Singapore: Springer, 2024: 171-193.

[26] WANG J, LIANG J Y, ZHAO X W, et al. Overlapping protein complexes detection algorithm based on assortativity in PPI networks[J]. Computer Science, 2019, 46(2): 294-300.

[27] WANG J, JIA Y, SANGAIAH A K, et al. A network clustering algorithm for protein complex detection fused with power-Law distribution characteristic[J]. Electronics, 2023, 12(14): 3007.

[28] XU M. Understanding graph embedding methods and their applications[J]. Siam Review, 2021, 63(4): 825-853.

[29] WANG R, MA H, WANG C. An ensemble learning framework for detecting protein complexes from PPI networks[J]. Frontiers in genetics, 2022, 13: 839949.

[30] MENG X, XIANG J, ZHENG R, et al. DPCMNE: Detecting protein complexes from protein-protein interaction networks via multi-level network embedding[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2021, 19(3): 1592-1602.

[31] CHEN H, CAI Y, JI C, et al. AdaPPI: Identification of novel protein functional modules via adaptive graph convolution networks in a protein-protein interaction network[J]. Briefings in Bioinformatics, 2023, 24(1): 523.

[32] WANG S, CUI H, QU Y, et al. Multi-source biological knowledge-guided hypergraph spatiotemporal subnetwork embedding for protein complex identification[J]. Briefings in Bioinformatics, 2025, 26(1): 718.

[33] XIA S, LI D, DENG X, et al. Integration of protein sequence and protein-protein interaction data by hypergraph learning to identify novel protein complexes[J]. Briefings in Bioinformatics, 2024, 25(4): 274.

[34] FU G, HOU C, YAO X. Learning topological representation for networks via hierarchical sampling[C]// 2019 International Joint Conference on Neural Networks (IJCNN). IEEE, 2019: 1-8.

[35] KUMAR T, VAIDYANATHAN S, ANANTHAPADMANABHAN H, et al. Hypergraph clustering by iteratively reweighted modularity maximization[J]. Applied Network Science, 2020, 5(1): 52.

[36] XIANG N, YOU M, WANG Q, et al. Hypergraph network embedding for community detection[J]. The Journal of Supercomputing, 2024, 80(10): 14180-14202.

[37] GAVIN A C, BÖSCHE M, KRAUSE R, et al. Functional organization of the yeast proteome by systematic analysis of protein complexes[J]. Nature, 2002, 415(6868): 141-147.

[38] GAVIN A C, ALOY P, GRANDI P, et al. Proteome survey reveals modularity of the yeast cell machinery[J]. Nature, 2006, 440(7084): 631-636.

[39] KROGAN N J, CAGNEY G, YU H, et al. Global landscape of protein complexes in the yeast Saccharomyces cerevisiae[J]. Nature, 2006, 440(7084): 637-643.

[40] XENARIOS I, SALWINSKI L, DUAN X J, et al. DIP, thedatabase of interacting proteins: a research tool for studying cellular networks of protein interactions[J]. Nucleic Acids Research, 2002, 30(1): 303-305.

[41] STARK C, BREITKREUTZ B J, REGULY T, et al. BioGRID: A general repository for interaction datasets[J]. Nucleic Acids Research, 2006, 34(1): 535-539.

[42] SZKLARCZYK D, GABLE A L, LYON D, et al. STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets[J]. Nucleic Acids Research, 2019, 47(1): 607-613.

[43] PU S, WONG J, TURNER B, et al. Up-to-date catalogues of yeast protein complexes[J]. Nucleic Acids Research, 2009, 37(3): 825-831.

[44] BROHEE S, VAN HELDEN J. Evaluation of clustering algorithms for protein-protein interaction networks[J]. BMC Bioinformatics, 2006, 7(1): 1-19.

[45] GIURGIU M, REINHARD J, BRAUNER B, et al. CORUM: The comprehensive resource of mammalian protein complexes—2019[J]. Nucleic Acids Research, 2019, 47(1): 559-563.

[46] IVAZEH A, ZAHIRI J, RAHGOZAR M, et al. Performance evaluation measures for protein complex prediction[J]. Genomics, 2019, 111(6): 1483-1492.

[47] OMRANIAN S, ANGELESKA A, NIKOLOSKI Z. PC2P: Parameter-free network-based prediction of protein complexes[J]. Bioinformatics, 2021, 37(1): 73-81.

[48] ZOU M, GAN Z, WANG Y, et al. Unig-encoder: A universal feature encoder for graph and hypergraph node classification[J]. Pattern Recognition, 2024, 147: 110115.

[49] LIU Z, TANG B, YE Z, et al. Hypergraph transformer for semi-supervised classification[C]// ICASSP 2024-2024 IEEEInternational Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2024: 7515-7519.

[50] WANG R, LIU G, WANG C. Identifying protein complexes based on an edge weight algorithm and core-attachment structure[J]. BMC Bioinformatics, 2019, 20(1): 1-20.

[51] HANNA E M, ZAKI N. Detecting protein complexes in protein interaction networks using a ranking algorithm with a refined merging procedure[J]. BMC Bioinformatics, 2014, 15(1): 1-11.

[52] JIANG P, SINGH M. SPICi: a fast clustering algorithm for large biological networks[J]. Bioinformatics, 2010, 26(8): 1105-1111.

[53] LIU G, WONG L, CHUA H N. Complex discovery from weighted PPI networks[J]. Bioinformatics, 2009, 25(15): 1891-1897.

[54] LI M, CHEN J, WANG J, et al. Modifying the DPClus algorithm for identifying protein complexes based on new topological structures[J]. BMC Bioinformatics, 2008, 9(1): 1-16.

[55] BADER G D, HOGUE C W V. An automated method for finding molecular complexes in large protein interaction networks [J]. BMC Bioinformatics, 2003, 4(1): 1-27.

[56] ZAKI N, EFIMOV D, BERENGUERES J. Protein complex detection using interaction reliability assessment and weighted clustering coefficient[J]. BMC Bioinformatics, 2013, 14(1): 1-9.

[57] PELLEGRINI M, BAGLIONI M, GERACI F. Protein complex prediction for large protein protein interaction networks with the Core&Peel method[J]. BMC Bioinformatics, 2016, 17(12): 37-58.

WANG Jie, born in 1988, Ph. D, associate professor, is a member of CCF (No. N2805M). His main research interests include data mining and bioinformatics, etc.

YANG Xiancan, born in 2000, master, is a member of CCF (No. T8922G). His main research interests include data mining and machine learning.

(责任编辑:何杨)