蛋白质相互作用网络功能模块检测的研究综述

冀俊忠 [1] 刘志军 [1] 刘红欣 [1] 刘椿年 [1]

摘 要 蛋白质相互作用 (Protein-protein interaction, PPI) 网络是生命活动中一种极其重要的生物分子关系网络, 利用计算方法从 PPI 网络中检测功能模块是目前生物信息学中一项重要的研究课题. 本文首先总结了功能模块检测过程的基本流程, 说明了预处理和后处理的作用; 其次, 提出了一种模块检测方法的分类体系, 并对其中一些代表性的检测算法进行了阐述; 再次, 给出了模块检测常用的数据库、评价指标和相关软件工具, 并通过实验对代表性算法进行了性能对比. 最后, 通过对该领域挑战性问题的分析预测了模块检测未来的研究方向, 以期对相关研究提供一定的参考.

关键词 蛋白质相互作用网络, 功能模块检测, 检测流程, 分类体系, 性能对比

引用格式 冀俊忠, 刘志军, 刘红欣, 刘椿年. 蛋白质相互作用网络功能模块检测的研究综述. 自动化学报, 2014, **40**(4): 577−593

DOI 10.3724/SP.J.1004.2014.00577

An Overview of Research on Functional Module Detection for Protein-protein Interaction Networks

JI Jun-Zhong[1] LIU Zhi-Jun[1] LIU Hong-Xin[1] LIU Chun-Nian[1]

Abstract As a bimolecular relationship network, the protein-protein interaction (PPI) network plays an important role in biological activities. Using computational approaches, mining functional modules from a PPI network is currently a challenge in bioinformatics. This paper firstly gives a workflow of detecting functional modules from PPI data, and illustrates the effects of preprocessing and post processing. Next, a systematic category of functional module detection is proposed, and many typical detecting algorithms in each category are described. And then, the paper lists some public databases, evaluating metrics, related software tools, and experimentally compares and analyzes the performances of some representative algorithms on the same data. Finally, the existing problems and prospects in this field are presented, which offers some references for researchers engaged in PPI network analyzing.

Key words Protein-protein interaction (PPI) network, functional module detection, detection process, classification system, performance comparisons

Citation Ji Jun-Zhong, Liu Zhi-Jun, Liu Hong-Xin, Liu Chun-Nian. An overview of research on functional module detection for protein-protein interaction networks . *Acta Automatica Sinica*, 2014, **40**(4): 577−593

随着人类基因组计划的完成, 蛋白质组学的研究已经成为生命科学乃至自然科学领域中最重要的研究课题之一. 蛋白质组学是一种对蛋白质所含特性进行系统化研究的学科, 其研究目的是为生物系统在健康和疾病状况下的结构、功能和调控提供详细的描述. 由于蛋白质很少以独立个体的方式实现其被赋予的功能, 即在同一细胞的生化过程中所涉及到的蛋白质一般通过彼此之间的相互作用形成大分子复合物后才能完成其生物学功能, 例如遗传物质复制、基因表达调控、细胞信号转导、新陈代谢、细胞增殖、细胞凋亡等过程和活动都依赖于蛋白质之间的相互作用. 因此蛋白质相互作用及相互作用网络的研究和分析很自然地成为理解生命活动中细胞组织、过程和功能的基础[1]. 而且生命科学的研究表明, 对蛋白质相互作用的研究不仅能从系统角度理解各种生物学过程, 揭示疾病的发生机制, 而且能够帮助人们寻找新的药物靶标, 为新药研发起到积极的作用[2]. 因此蛋白质相互作用的研究迅速成为后基因组时代最重要的研究领域, 并受到人们越来越多的重视.

收稿日期 2013-01-05 录用日期 2013-05-07
Manuscript received January 5, 2013; accepted May 7, 2013
国家重点基础研究发展计划 (973 计划) (2014CB744601), 国家自然科学基金 (61375059, 61332016), 教育部博士点学科专项科研博导类基金 (20121103110031), 北京市教委科研计划重点项目 (北京市自然科学基金重点 B 类项目) (KZ201410005004) 资助
Supported by National Basic Research Program of China (973 Program) (2014CB744601), National Natural Science Foundation of China (61375059, 61332016), Specialized Research Fund for the Doctoral Program of Higher Education (20121103110031), and the Beijing Municipal Education Research Plan Key Project (Beijing Municipal Fund Class B) (KZ201410005004)
本文责任编委 张学工
Recommended by Associate Editor ZHANG Xue-Gong
1. 北京工业大学计算机学院多媒体与智能软件技术北京市重点实验室 北京 100124
1. Beijing Municipal Key Laboratory of Multimedia and Intelligent Software Technology, College of Computer Science and Technology, Beijing University of Technology, Beijing 100124

近年来, 随着酵母双杂交[3−4]、基于质谱的串联亲和纯化[5−6]、蛋白质芯片[7−8] 等高通量生物实验方法的发展和文本挖掘技术在蛋白质相互作用预测中的广泛应用[9−11], 蛋白质相互作用 (Protein-protein interaction, PPI) 的可用数据日益丰富, 并促成了越来越多的蛋白质相互作用网络 (PPI 网络) 的形成. PPI 网络是指一个生命有机体内的所有蛋白质之间相互作用组成的网络, 在一个 PPI 网络中, 不同时间和空间阶段通过相互作用完成某一特定分子进程的蛋白质集合称为蛋白质功能模块. 因为理解大量生物学数据所包含的生物学意义是后基因时代非常重要的研究任务[12], 所以面对大量可用的 PPI 网络数据, 如何快速、有效地识别各种具有生物学功能的功能模块 (也称蛋白质复合物) 就成为蛋白质组学研究中一项极为关键的科学问题.

通过生物实验进行 PPI 网络功能模块的检测是过去传统而经典的方法, 但该方法在检测费用、时间和质量上的局限性[13−14], 远远无法满足后基因时代人类对生命科学研究的实际需要. 而生物学的研究发现 PPI 网络中那些紧密连接的蛋白质区域通常会与蛋白质的功能模块相对应[15]. 于是以机器学习和数据挖掘为基础, 通过识别 PPI 网络中紧密连接结构来发现蛋白质功能模块的方法迅速兴起, 并成为解决这一问题的有效途径. 近十年, 新的思想层出不穷, 新的检测方法不断涌现, *Nature*[6, 13, 16]、*Science*[15, 17]、*Proceedings of National Academy of Sciences* (*PNAS*)[18−19]、*Nucleic Acids Research*[20]、*Bioinformatics*[21−23]等国际权威杂志以及许多重要的国际会议都多次报道该方面的研究工作, 在生物信息领域掀起了利用计算方法进行 PPI 网络蛋白质功能模块检测的研究热潮.

本文对 PPI 网络功能模块检测方法的研究现状进行了综述, 并结合对该领域目前所面临的主要问题的分析, 预测了未来可能的发展方向. 论文试图为 PPI 网络功能模块的检测勾画出一个较为全面和清晰的概貌, 以期对该领域的相关研究提供有益的参考.

1 功能模块的检测过程

一个 PPI 网络可以抽象为一个节点带标签的无向图 $G = (V, E)$, 其中 V 表示节点集合, E 表示边的集合. PPI 网络是一种高度动态和结构化的复杂网络[24], 它具有无尺度分布、小世界性质和功能模块化 3 个拓扑特征. 无尺度分布是指 PPI 网络中连接度为 k 的节点出现的概率满足幂律分布[25]; 小世界性质是指 PPI 网络所含的蛋白质节点具有较短的平均最短路径长度和较高的平均聚集系数[26]; 而功能模块化则指在 PPI 网络中相互连接密集的蛋白质形成一个模块, 并完成同样的功能. 这 3 个结构特征制约或影响着功能模块检测过程和方法的设计, 尤其是功能模块化的特征更是许多检测方法实现的理论依据.

PPI 网络功能模块检测的流程通常包括 5 个步骤: 数据预处理、PPI 网络形成、聚类算法、后处理、功能模块输出, 如图 1 所示.

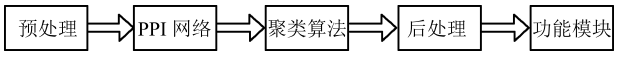


图 1 PPI 网络功能模块检测的流程

Fig. 1 Detection process of function modules from PPI networks

在以上流程中, PPI 网络形成、聚类算法和功能模块的输出是模块检测过程中三个基本且必要的步骤, 而预处理和后处理则是可选的操作步骤.

当前 PPI 数据众多、来源广泛, 双杂交技术、质谱分析、文本挖掘等方法都可以提供蛋白质间的相互作用数据. 由于不同方法检测标准不一致, 且受实验环境等诸多因素的影响, 致使 PPI 数据库中经常会包含一些无用数据. 这些无用数据主要表现为: 冗余数据、不完全数据及噪声数据. 冗余数据是指自相互作用和重复相互作用的数据. 不完全数据是指数据库中相互作用数据的缺失 (假阴性), 即仍有很多潜在的相互作用未能被有效地检测, 而噪声数据是指已被检测到的, 但在实际细胞中并不存在的蛋白质相互作用数据 (假阳性). 文献 [27] 对采用酵母双杂交技术得到的相互作用数据的研究表明只有大约 50 % 的相互作用数据是可靠的. PPI 网络本身的复杂性已经为检测方法提出了较高的要求, 而上述 PPI 数据的不完备特征更加剧了检测的难度. 于是, 就有必要对蛋白质相互作用数据进行预处理, 以减少无用数据, 保证使用数据的可靠性. 除了清除数据库中冗余数据的去重操作外, 加权操作是一种常用的通过相似度或距离的度量对相互作用数据的可靠性进行评估的方法. 文献 [21] 对基因产物的多种距离度量标准进行了性能的比较和分析, 并提出了基于系数计算的拓扑距离准则: 两个蛋白质的公共相互作用节点数越多, 它们的功能相似性就越大. 基于这种思想, Brun 等在文献 [28] 中采用 Czekanowski-Dice 距离来计算任意节点间的距离. 而文献 [18] 则提出将节点对之间最短路径的长度作为相互作用的权重度量标准. 进一步, Pei 等在文献 [29] 中将两节点间所含路径权重的乘积定义为整条路径的强度. 它将 k-length 路径强度与 k-length 最大路径强度的比率作为两节点间的距离. 此外, 近年来也涌现出一些结合生物信息知识来计算权重的

方法, 其实现机理是根据蛋白质的功能注释, 将节点间的距离表示为它们的生物学相似性. 其中最常用的方法是结合基因本体论 (Gene ontology, GO) 的注释术语来计算基因产物的相似性. 如 Resink 方法[30] 将两个基因产物的相似性定义为各自注释条目间的最大语义相似度. Jiang 和 Conrach 的方法[31] 将蛋白质节点间的距离定义为 GO 中连接两节点的最短路径上的权重之和. 而 Lin 的方法[32] 则将蛋白质的相似性定义为各自注释条目间共同最近祖先节点的信息量与两组注释条目的平均信息量的比值. 后处理步骤, 是对初始聚类结果的修正过程, 通常利用已有的生物学知识对聚类的初始结果进行过滤、合并或融合, 以形成具有生物学意义的功能模块. 例如 RNSC (Restricted neighor hoods search clustering)[33] 算法采用模块大小、模块密度和模块中节点间的功能同质性三种度量标准完成对初始结果的过滤. DECAFF[34] 算法也采用节点的功能信息来衡量初始模块中节点间相互作用的可靠性, 进而过滤掉那些可靠性较差的模块. 而 CASCADE[35] 算法则根据不同模块之间的连接关系定义了模块间的相似性度量标准, 并通过迭代地合并那些满足相似性阈值的模块来得到最终的聚类结果. 一般来说, 通过一系列后处理的操作, 模块的聚类效果都能得到明显的提高.

由于 PPI 网络通常具有同一个功能模块中蛋白质相互连接密集, 不同模块间蛋白质相互连接稀疏的特点, 所以聚类过程是模块检测过程中最为核心的步骤. 经过近 10 年的快速发展, 已经涌现出了许多基于不同聚类机理的 PPI 网络功能模块检测方法. 下面将主要围绕这个核心步骤进行讨论, 分析和比较现有检测算法的基本原理、特点、不足以及有待解决的问题.

2　功能模块的检测方法

如前所述, 利用计算方法进行蛋白质网络功能模块的检测已经成为后基因时代生物信息学领域中一个非常活跃的研究领域. 根据所采用的计算模型和机理, 我们可以将其划分为基于传统图理论的检测方法和基于非传统图理论的检测方法两大类. 其中基于传统图理论的检测方法又可进一步分为基于密度的聚类方法、基于层次的聚类方法和基于划分的聚类方法; 而非传统图理论的检测方法可细分为基于流模拟的聚类方法、基于谱分析的聚类方法、基于核心–附属关系的聚类方法和基于群集智能的聚类方法. 这种功能模块检测方法的分类体系如图 2 所示, 在每类方法下又可包含多个不同的具体算法.

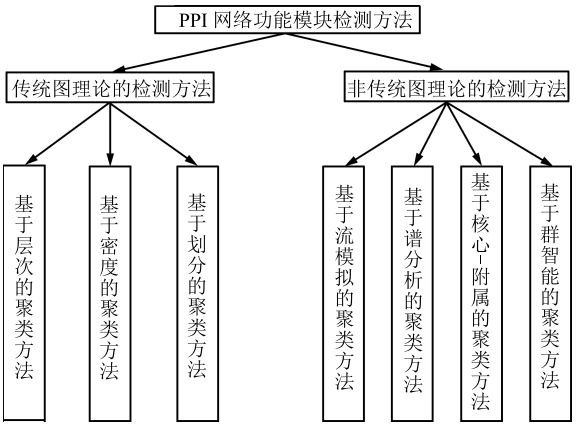


图 2　PPI 网络功能模块检测方法分类图

Fig. 2　The classification of detection methods about functional modules from PPI networks

2.1　传统图理论的检测方法

PPI 网络通常是用图来表示的, 所以基于拓扑结构信息的传统图理论分析法是功能模块检测最常用的聚类方法. 基于传统图理论分析的方法依据检测结果所体现出的特征又可细分为如下 3 类方法.

2.1.1　基于密度的聚类方法

基于密度聚类的方法通过在 PPI 网络中搜索浓密连接的子图来发现功能模块. 其中最具代表性的算法是极大团 (Maximize cliques) 方法[36], 该文献共实现了 3 种算法, 第 1 种算法通过穷尽地检测完全连接的子图 (团) 来进行检测. 由于这种团方法的使用受限于 PPI 网络数据的不完备, 所以该文献又实现了超磁聚类 (Super paramagnetic clustering, SPC) 和蒙特卡罗模拟 (Monte Carlo, MC) 两种算法. 实验结果表明 SPC 在识别与外部存在较少连接的簇时具有优势, 而 MC 在识别拥有共同节点的簇和高密度图时显示出更好的性能.

一些能够替代密度函数作为优化目标的模块检测算法也被陆续提出. 如 Bader 等提出了一种基于蛋白质的连接值来检测蛋白质复合物的计算方法 — MCODE (Molecular complex detection) 算法[37], 它首先利用节点的局部邻域密度给 PPI 网络中每个节点进行加权, 然后选择具有最高权值的节点作为初始聚类的种子节点, 并由种子节点向外扩张形成最后的簇 (蛋白质模块), 尽管该方法可以产生有重叠的簇, 但所产生簇的个数较少, 会使某些簇包含的蛋白质数量过大; 而且由于权重大的节点彼此之间的连接并不一定稠密, 故该方法不能完全保证得到的簇是稠密的.

Adamcsek 等设计了在 PPI 网络中发现功能模块的 CFinder 软件[22], 该软件使用一种团渗透的方法[16] 来检测具有 k 个节点的 k-团簇, 然后通过合并

相邻的两个 k-团簇 (若两个 k-团簇有 $k-1$ 个公共节点, 则称之为相邻) 来得到 PPI 网络的功能模块. 该方法的最大特点是能够检测有重叠的功能模块, 但是因为很难选取一个合理的参数 k, 并且由于对功能模块的 k-团簇特性的严格限制, 使得 CFinder 在聚类时往往会漏掉一些常规的功能模块.

文献 [38] 给出了一种通过簇边界的跟踪进行聚类的 DPClus 算法, 它不仅利用模块密度而且利用新定义的簇特性 CP_{nk} 完成功能模块检测. DPClus 方法首先基于两个蛋白质所包含的公共邻居加权它们所连接的边, 并通过节点的加权度来加权节点; 然后选择具有最高权值的节点或者度最大的节点 (种子节点) 作为初始簇, 并根据簇外节点与簇内节点的连接边数及边上的权重来定义簇外节点加入到簇的优先级; 最后根据密度阈值和簇特性阈值, 通过逐个添加当前簇外优先级最高的节点来迭代地扩展当前簇. 一旦一个模块确定就将其从 PPI 网络中删除, 并从剩下的网络中继续检测功能模块, 直到剩余的网络中不存在相互作用的边为止.

Pei 等提出了一种基于子图质量度量来进行模块检测的新方法[39], 该方法首先给出了一种浓密子图质量的统计度量, 然后通过双层的启发式搜索选择好的种子, 最后通过子图精炼来实现对模块重叠的控制.

基于密度聚类的方法是进行 PPI 网络功能模块检测最经典的方法, 其基本假设是网络中的稠密子图通常对应着蛋白质的功能模块. 尽管该方法能够有效地检测出 PPI 网络中浓密连接的子结构, 但是一般情况下很难对网络中大量的稀疏节点进行聚类. 而随着研究的深入, 人们发现 PPI 网络中的一些稀疏子图实际上也对应着某些真实的功能模块, 因此, 基于密度聚类的方法在进行 PPI 网络功能模块检测时也具有一定的局限性.

2.1.2 基于层次的聚类方法

由于 PPI 网络在本质上具有模块化的层次特性, 所以基于层次聚类的方法很自然地适用于蛋白质功能模块的检测. 这类方法一般采取迭代地融合节点或递归地分裂图来实现 PPI 网络的节点聚类. 比如, Ravasz 等提出利用无尺度性质刻画的模块—中心结构来进行功能模块的识别[17]. 该方法将一些中心 (Hub) 节点视为网络结构形成中的决定性节点和层次结构的骨架, 并强调低度的节点可以通过互连形成一个小的模块. 在模块中的一个核心 (Core) 节点不仅与其他小模块中的核心节点相连, 而且会与一个高度的节点连接, 于是逐渐成为由一组小模块形成的大模块中的核心节点. 利用 Hub 重复上述步骤, 最终形成一个具有层次的模块结构.

文献 [19] 通过对不同来源和类型的实验数据进行集成分析, 能够识别出在统计意义上具有相关行为的模块. 这些模块不仅能揭示许多生物学过程, 而且从整体上体现了层次化的模块结构. Arnau 等提出通过度量蛋白质间距离完成蛋白质融合的层次聚类方法[23]. 该方法分为两步, 首先根据 PPI 网络中蛋白质互连所需要的最小步数来计算任意 2 个蛋白质间的距离, 然后使用基于距离的层次聚类方法对蛋白质进行分簇. 该方法最突出的特点是能在距离计算过程中结合已知的生物学知识.

文献 [40] 中提出了一种基于层次的功能模块检测的 Jerarca 方法, 它包含三个处理步骤, 首先, 利用 UVCluster、RCluster 和 SCluster 三种方法中的任一方法来计算蛋白质节点间的加权距离; 其次, 根据加权距离, 采用经典的非加权组平均算法和邻接交叉算法将 PPI 网络的距离矩阵转换为层次结构的树; 最后, 根据簇内和簇间节点的连接分布进行最优的层次划分得到预测的功能模块.

文献 [18] 提出了另一种在 PPI 网络中进行模块结构发现的方法, 它实现的依据是: 任意 2 节点间的最短路径很可能是在功能联系上最相关的路径, 网络中任一节点相对于其他节点都有一个唯一的最短路径概要, 任一模块都是由具有相似最短路径概要的成员构成. 在网络模块的层次结构中, 中介数 (Betweenness) 是一种检测模块间桥路 (Bridge) 的通用指标, 通过迭代地去除具有最高中介数的节点或边, 文献 [41] 将图划分为 2 个或多个子网络, 最终发现指定数量的具有层次的模块划分.

基于层次的聚类结果最终可以用一个树形结构清晰地表达 PPI 网络模块化的组织结构. 由于在分裂或合并过程中不同的节点被划分到不同的子图中, 所以层次聚类的方法很难检测到节点交叠的功能模块; 而且该过程通常采用局部节点间的信息, 一旦某个节点的层次分类错误, 其下层的子树都将可能发生错误, 因此该类方法最明显的缺点是聚类的结果对网络的噪声非常敏感.

2.1.3 基于划分的聚类方法

划分聚类的目的是寻找一个最优的网络划分, 使得划分到同一簇的对象之间尽可能 “接近” 或相似, 而不同簇中的对象间尽可能 “远离” 或不同; 并且能够对所有稀疏连接的节点做出合理的解释.

King 等在 2004 年提出了一种利用代价函数识别 PPI 网络最佳划分的 RNSC 算法[33]. RNSC 从一个随机聚类的划分开始, 通过边缘节点在不同簇间的移动来改变簇的划分, 最后把具有最低代价值的划分作为最好的聚类. 为得到好的性能, 该方法需要功能模块数的先验知识, 而且聚类结果依赖于初

始划分的质量.

Dunn 等在文献 [42] 中使用 Edge-betweenness 算法[43] 将 PPI 网络划分为相互连接的子图结构. Edge-betweenness 聚类方法通过迭代地删除 PPI 网络中具有最大边介数的中心边来获得网络子图所构成的功能模块. 由于边介数的计算需要整个图的拓扑信息, 且在聚类过程中也可以使用全局信息, 所以聚类结果中能够包含稀疏连接的节点. 因此该方法的一个优点是可以使用稀疏连接节点的功能注释信息来减少 PPI 网络噪声数据对检测结果的影响.

亲缘传播 (Affinity propagation, AP)[44] 是近年来出现的一种新型的聚类方法, 它首先通过在数据点之间交换实值消息来识别数据库中代表性的标本 (簇的中心), 然后迭代地将数据点划分到最具代表性的标本所在的簇中.

划分聚类的方法是一种易于理解和实现的方法. 然而, 其最大的缺点在于需要事先知道确定的聚类数目, 而且和层次的聚类方法一样, 它也不能得到节点重叠的功能模块.

2.2 非传统图理论的检测方法

除了上面利用传统图理论的检测方法外, 近年来也涌现出许多将其他机理融合于图聚类过程的检测方法, 如基于流模拟的聚类方法、基于谱分析的聚类方法、基于核心 — 附属连接关系的聚类方法和基于群集智能的聚类方法等.

2.2.1 基于流模拟的聚类方法

流模拟的方法通过分析 PPI 网络中每个蛋白质对其他蛋白质施加的生物学或拓扑影响的程度来实现功能模块的检测. 具体方法是通过对生物或功能流在网络中流动情况的模拟, 在各类具体的外部激励下建立网络的行为模型来刻画系统内在的随机行为. Nabieva 等在文献 [45] 中提出了一种功能流的概念, 他们将每个已知功能注释的蛋白质看作是一个功能流的源, 功能流以图中相互作用的连接边为导管从源向未知功能的蛋白质传播. 迄今为止, 已有许多研究利用基于流模拟的方法在 PPI 网络中进行蛋白质的功能预测与分析.

MCL (Markov clustering)[46] 就是通过模拟在 PPI 网络中流的自由行走来检测蛋白质复合物的经典算法, 它定义了指派节点概率的 Expansion 操作和改变节点游走概率的 Inflation 操作来模拟随机游走的扩展和收缩行为. 初始时每个漫游者都会选择不同的路径游走, 一旦进入一个稠密区域便不会轻易离开, 通过反复地使用 Expansion 操作和 Inflation 操作把 PPI 网络划分为许多稠密子图. 作为一种具有随机性的聚类算法, MCL 尽管容易受到给定条件的影响, 但仍然具有很强的鲁棒性, 能够适应网络结构的变化并在一定程度上能克服数据噪声对检测结果的影响.

在文献 [47] 中, Hwang 等通过建立动态的信号转换模型 (Signal transduction model, STM) 来表达网络中每一蛋白质引起的扰动信号的转换行为. STM 算法首先为每个簇选择有代表性的蛋白质, 然后综合信号转换和图拓扑信息不断地精炼簇, 并形成最终的功能模块. 该方法能够识别较大、低密度、任意形状且生物功能较密集的模块. 文献 [35] 对 STM 算法做了进一步的扩展, 重点对传播路径进行了调整. 它将相互作用对间的最短路径调整为准全路径度量, 提出了一种类全路径的 CASCADE 算法, 取得了更好的结果.

2007 年, Cho 等提出了另一种基于流的 PPI 网络功能模块检测算法[48−49], 该算法首先利用基因本体 (GO) 注释信息定义了语义相似、语义相关 2 种新度量来评估蛋白质相互作用的可靠性, 然后基于新度量将 PPI 网络转换为加权图, 最后采用基于流的识别方法得到有蛋白质重叠的功能模块.

Feng 等将 PPI 数据和基因微阵列数据相结合, 提出了一种新的检测蛋白质复合物的图碎片算法[50], 该算法首先利用最大网络流的改进算法发现 PPI 网络中较大的密集子图, 然后利用微阵列数据信息迭代地将密集子图分裂成足够小的碎片, 这种方法充分利用了 PPI 网络的密度信息和微阵列的基因信息, 能在识别精度、效率和新蛋白质复合物发现的能力方面获得很好的性能.

2.2.2 基于谱分析的聚类方法

谱聚类方法以谱图为理论基础, 采用矩阵分析技术将待求解问题转化为带约束的二次型优化问题, 近年来被广泛应用于图分割和复杂网络聚类等领域[51]. 基于矩阵特征向量的计算, 谱方法首先通过对图邻接矩阵的研究, 利用矩阵的谱来确定图的相关性质, 然后根据顶点之间的权值对图进行划分.

Kamp 等在果蝇的 PPI 网络上研究了网络的谱属性与其结构特征的关系, 发现离散的谱密度能够对应 PPI 网络的拓扑结构, 尤其擅长刻画环状结构[52]. 文献 [20] 利用谱方法对芽殖酵母的 PPI 网络进行分析, 提出了一种利用超几何分布的 p-值函数来实现网络划分的谱方法, 有效地揭示了复杂 PPI 网络隐含的拓扑结构. Sen 等利用酵母 PPI 网络连通矩阵上的特征模态分析进行蛋白质功能的聚类, 他们发现在同一特征向量中的蛋白质不仅具有相似的度, 而且更有可能发生相互作用[53]. 这些研究表明网络的谱和拓扑结构之间有着紧密的联系, 通过对网络谱的分析可以达到对其聚类的目的.

2010 年, Qin 等也使用谱聚类方法对 PPI 网络

的模块识别问题进行了研究, 提出了一种基于 PPI 网络属性确定模块数的方法, 并在 PPI 网络的 4 种相似图上进行了有效性检验[54]. 该方法首先分别基于邻接相似度、公共邻居相似度、传输相似度以及交换相似度构建 PPI 网络的拉普拉斯矩阵; 然后根据 PPI 网络中孤立子图及 Hub 节点的个数得到预测模块个数 k, 最后采用 k-均值聚类方法来处理特征向量矩阵, 进而得到最终的模块.

文献 [55] 提出了一种基于扩散模型的谱聚类 ADMSC (Adjustable diffusion matrix-based spectral clustering) 方法, 该方法将 PPI 网络模块结构的发现看作是粒子扩散过程中的随机行走问题, 通过引入能调节扩散矩阵的功率因子来处理 PPI 网络蛋白质的异类问题. 实验表明, ADMSC 是一种简单、快速且鲁棒的方法, 尤其在 PPI 网络所含簇的规模近似相等时, ADMSC 方法更为有效.

谱聚类方法通过计算特征向量, 构建特征向量空间, 不仅能降低数据的维数, 简化算法的计算量, 而且从理论上讲也能在任意形状的样本空间中收敛于全局的最优解. 但是特征向量的选择将直接关系到最终的聚类效果, 所以如何准确地选择特征向量的个数仍是一个亟待解决的关键问题.

2.2.3 基于核心 – 附属关系的聚类方法

根据对实验所确定的蛋白质复合物的生物信息学分析, 蛋白质复合物内在组织具有核心附属的结构, 一个功能模块通常由核心蛋白和附属蛋白组成[56]. 位于 “核心” 内的蛋白质具有高度的共表达特性及高度的功能相似性, 附属蛋白质位于功能模块的 “核心” 外围, 辅助 “核心” 实现相关的生物功能. 近年来, 基于这种 “核心 – 附属” 的概念, 研究者设计了一些利用核心依附关系进行功能模块检测的算法.

Core 算法[57] 是新近提出的一种通过先分别识别蛋白质复合物的核心蛋白和附属蛋白再进行模块检测的算法. 该算法分 3 个步骤: 1) 预测核心蛋白质; 2) 识别每个核心蛋白质的附属蛋白质并去除不重要的核心蛋白质; 3) 利用预测复合体内部及外部相互作用数的关系计算评分值, 并根据评分对所预测蛋白质复合物的重要性排序. 该算法识别蛋白质复合物的能力较强, 但不能处理复合物间蛋白质的重叠问题.

基于类似思想, Wu 等提出了 COACH (Core-attachment based method) 算法[58], 该算法分为两个阶段, 第 1 阶段从邻接图中定义核心顶点, 然后从中检测蛋白质复合物的核心蛋白质; 第 2 阶段将附属蛋白质逐个连接到核心蛋白质所代表的复合物中, 从而形成具有生物学意义的功能结构. 与 Core 相比, COACH 算法具有的最大优势是能够处理复合物间蛋白质的重叠问题.

2012 年, Ma 等也提出了一种新的基于 “核心 – 附属” 结构的检测方法[59], 它首先利用 PPI 网络的特征值和特征向量构建一个虚拟网络. 因为虚拟网络中的每一个最大基团对应着一个功能模块的 “核心”, 所以模块 “核心” 的检测问题就能够被转换为求解网络的最大基团问题. 然后利用 “核心 – 附属” 关系将附属节点融入到各自的核心形成具有生物学意义的复合物. 尽管这种方法具有好的预测精度和强的鲁棒性, 但是由于具有比较高的复杂度, 所以难以在较大的 PPI 网络中得到应用.

2.2.4 基于群集智能的聚类方法

群集智能算法是一类基于简单个体相互作用时涌现的整体智能行为而提出的元启发搜索方法. 其特点是在求解各种组合优化问题时, 通过模拟社会型生物 (如蜜蜂、蚂蚁、鸟、鱼等) 群体间的协作行为来有效获取问题的近似最优解. 群集智能算法发展迅速, 不仅已在机器学习、数据挖掘领域得到了广泛的应用, 而且也在 PPI 网络的功能模块检测中开始了一些新的探索.

2008 年, Sallim 等提出利用 ACO 算法求解旅行商问题 (Traveling salesman problem, TSP) 的机制来完成蛋白质网络聚类的 ACOPIN (Ant colony optimization in protein interaction networks) 算法[60]. 它首先通过转换 PPI 网络的邻接矩阵获得蛋白质间的距离矩阵; 然后应用 ACO 获取遍历所有蛋白质的最优路径; 最后基于彼此间距离较近的蛋白质通常在最优路径序列上也相邻的现象, 断开最优路径上距离较长的邻近节点便可形成对应的功能模块簇. 该方法是首次将 ACO 引入到 PPI 网络功能模块检测中的群集智能算法, 但该方法的有效性还需要更系统而深入的研究.

2012 年, Ji 等也提出了基于 ACO 的 PPI 网络功能模块检测方法 NACO-FDM[61], 该方法结合 PPI 网络的拓扑结构和蛋白质的功能信息, 设计了更有效的启发函数来指导蚁群最优解的构建, 同时通过一系列新的划分、过滤和合并策略得到了较好的模块检测结果. 但是该方法仍沿袭了 ACO 算法在求解大规模组合优化问题时易陷入局部最优的固有缺陷. 为了克服这个不足, 他们随后又将 ACO 和多智能体进化 (Multi-agent evolution, MAE) 相结合, 提出了一种新的 ACO-MAE 算法[62]. 新算法能发挥 ACO 优化机制与 MAE 进化机制的各自优势, 有效地避免求解过程陷入局部最优, 从而提高了检测的质量.

同年, 雷秀娟等提出了一种基于连接强度的蚁

群优化聚类算法[63], 该算法引入了连接强度的概念对蚁群聚类算法中的拾起/放下规则加以改进. 核心思想是: 以连接强度作为拾起规则, 对节点进行聚类, 并根据放下规则放弃部分不良数据, 产生最终聚类结果. 但是算法的性能有待于在规模较大、噪声较高的 PPI 网络上进行更为系统的验证.

3 功能模块检测的实验评测

3.1 公共数据集合

随着生物信息学的广泛研究和深入发展, 已经有多种不同生物组织的综合 PPI 数据集合能够用于蛋白质功能模块检测的测试. 由于这些开放的数据库分属于世界上不同的研究机构, 各自都有自己的数据结构、格式和描述模式, 而且数据处理技术也不尽相同, 所以还没有形成统一的标准. 目前最常使用的公共数据集合包括:

慕尼黑蛋白质序列信息中心 (Munich information center for protein sequences, MIPS)[64] 是一个高质量的关于哺乳动物 (包含人类) 的 PPI 数据集合. 其内容源于多种已发表的科学文献中所提供的实验证据, 并通过领域专家的收集、整理获得的, 它是一个可用于多种研究目的的开放数据库.

相互作用蛋白质数据库 (Database of interacting proteins, DIP)[65] 是最早从研究文献中进行蛋白质相互作用二进制数据的提取、存储和组织的数据库之一. 由于该数据库采用开放机制, 不仅其内容可以随着分子相互作用信息的增加而更新, 而且实验数据的可靠性会根据数据质量评价方法的发展而在基因组级别上及时完成测试, 所以 DIP 能够为高通量 PPI 数据可靠性的评价、预测方法的开发、PPI 网络属性的研究提供参考.

生物分子相互作用网络数据库 (Biomolecular interaction network database, BIND)[66] 是 BOND (Biomolecular object network databank) 的一部分, 它包含生物分子相互作用、反应、复合物、路径等信息. BIND 数据库被组织成一种综合的可用计算机处理的信息模式, 能够为研究人员发现相互作用和分子机制提供有力支持. 此外, BIND 拥有的基因和蛋白质潜在注释的可视化手段也为分子相互作用网络的研究提供了便利.

通用交互数据集 (General repository for interaction datasets, BioGRID)[67] 是一个将物理和遗传相互作用相结合, 且数据持续更新的数据库. 它一共包含 22 种生物体的 130 000 多条非冗余的相互作用数据, 其中酵母相互作用数据超过 55 000 条, 成为当前酵母相互作用的最大数据库.

蛋白质相互作用数据库 (Protein interaction database, IntAct)[68] 是一个用于建模、存储和分析分子相互作用数据的开放数据库和软件工具. 它的可用数据完全源于已发表的文献, 并经过生物学家对实验方法、条件和作用子域 (Domain) 的手工标注而成. 除 PPI 数据外, 它还包括 DNA、RNA 以及小分子相互作用的大量信息.

人类蛋白质参考数据库 (Human protein reference database, HPRD)[69] 是一个丰富的与人类相关的蛋白质信息资源数据库, 它包括子域结构、蛋白质功能、蛋白质相互作用、PTMs (Post-translational modifications)、酶作用物关系、亚细胞定位、组织表达和基因的疫病关联等数据. 该数据主要源于包括酵母双杂交在内的各种生物实验方法.

相互作用基因/蛋白质检索搜索工具 (Search tool for the retrieval of interacting genes/proteins, STRING)[70] 是一个可用来探索和分析基因组关联关系的可计算全球资源库. 由于基因组关联关系分别是从功能 (同一功能的需求)、位置 (所处位置的临近度) 和事件 (所涉及的基因融合事件) 三种不同途径获得的, 所以 STRING 也包含了评价和比较各种预测方法重要性的评分框架. 该数据库仍在不断地更新和发展, 目前包含了来自 1 133 种生物组织的 5 214 234 个蛋白质信息.

在线预测人类相互作用数据库 (Online predicted human interaction database, OPHID)[71] 是一个基于 Web 的在线预测人类蛋白质相互作用的数据库. 它将来自于 BIND、HPRD 和 MINT 等数据源中的人类 PPI 信息与从酵母菌、秀丽隐杆线虫、果蝇和小家鼠等生物预测得到的 PPI 信息结合到一起, 构成了哺乳动物及真核生物的蛋白质相互作用资源库. 其中所含的 23 889 个相互作用可用蛋白质结构域、基因共表达和基因本体术语来评价. OPHID 不仅可进行单项条件或多项条件查询, 而且能使用定制的图形可视化程序对结果进行相应的可视化表示.

分子相互作用数据库 (Molecular interaction database, MINT)[72] 也是一个对各种期刊文献中实验部分所涉及的分子相互作用数据进行提取, 并以结构化格式进行存储的数据库. 它主要侧重于对蛋白质间实际相互作用数据的整理, 目前还包含了一个从直系同源体推导出的人类蛋白质相互作用数据库 HomoMINT.

人类蛋白质相互作用数据库 (Human protein interaction database, HPID)[73] 不仅能够提供从已有结构和实验数据中计算得到的人类蛋白质相互作用信息, 而且为用户发现的人类蛋白质相互作用数据提供了一个存储的场所. 此外, 该数据库还提供了一些非常实用的基于 Web 的应用程序, 能够实现根

据用户所提交的查询来预测潜在的蛋白质相互作用, 并对蛋白质相互作用网络进行可视化等功能.

3.2 评价方法

许多文献都对模块检测的结果评测进行了研究[20, 74−77], 虽然其中也包含利用不同的数据表示方式对聚类结果进行的评测[74], 但大部分研究都使用如下三种流行的度量方法来评价预测模块的质量和计算预测方法的精度[20, 75−77].

3.2.1 精度、召回率和 F 度量

精度 (Precision)、召回率 (Recall) 和 F 度量 (F-measure) 是信息提取和机器学习中常用的评价指标. 为了评价所预测的功能模块与实际功能模块的匹配度, 许多研究都利用如下邻域亲和评分公式来进行计算:

$$NA(p,b)=\frac{|V_p\bigcap V_b|^2}{|V_p|\times|V_b|} \tag{1}$$

其中, 预测模块用 $p=(V_p,E_p)$ 表示, 实际的基准模块表示为 $b=(V_b,E_b)$. 若 $NA(p,b)\geq\omega$, 则认为 p 和 b 模块相匹配 (一般取 $\omega=0.2$ 或 0.25). 如果令 P 为通过计算方法预测得到的功能模块集合, B 为在基准网络中实际的功能模块集合, 则 P 中至少与一个实际的功能模块相匹配的模块数量为 $N_{cp}=|\{p|p\in P,\exists b\in B,NA(p,b)\geq\omega\}|$, 从另一角度看, B 中至少与一个预测的功能模块相匹配的模块数量可表示为 $N_{cb}=|\{b|b\in B,\exists p\in P,NA(p,b)\geq\omega\}|$, 于是, 模块检测方法的精度和召回率可定义如下[75]:

$$\text{Precision}=\frac{N_{cp}}{|P|} \tag{2}$$

$$\text{Recall}=\frac{N_{cb}}{|B|} \tag{3}$$

F 度量是精度和召回率的调和平均值, 其大小为

$$\text{F-measure}=\frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \tag{4}$$

3.2.2 灵敏度、正的预测率和准确度

灵敏度 (Sensitivity, Sn)、正的预测率 (Positive predictive value, PPV) 和准确度 (Accuracy, Acc) 是新近提出的一组用来评价模块检测方法性能的度量指标. 设 $m=|V_p|$, $n=|V_b|$, T_{ij} 表示在第 i 个基准模块和第 j 个预测模块中共有的蛋白质的数量, 则灵敏度和正的预测率定义如下[76−77]:

$$\text{Sn}=\frac{\sum\limits_{i=1}^{n}\max\limits_{j}\{T_{ij}\}}{\sum\limits_{i=1}^{n}N_i} \tag{5}$$

$$\text{PPV}=\frac{\sum\limits_{j=1}^{m}\max\limits_{i}\{T_{ij}\}}{\sum\limits_{j=1}^{m}T_{\cdot j}} \tag{6}$$

其中, N_i 是第 i 个基准模块中所包含的蛋白质数量. 而 $T_{\cdot j}=\sum_{i=1}^{n}T_{ij}$. 一般来说, Sn 越高说明算法对实际功能模块中蛋白质的预测具有越好的覆盖率, 而 PPV 值高则说明预测到的功能模块更有可能是实际的模块. 作为一种综合度量, 预测的准确率 Acc 可用 Sn 和 PPV 的一种几何平均来求得:

$$\text{Acc}=\sqrt{\text{Sn}\times\text{PPV}} \tag{7}$$

3.2.3 *p* 值度量

随着蛋白质组学研究的深入, 人类获得的蛋白质的生物学知识越来越多, 这使将一个蛋白质与其功能注释相对应成为可能. 于是, 一个蛋白质簇 (预测到的蛋白质模块) 的发生对于一个给定功能注释在统计学上的意义就可以通过一个超几何分布的等式来进行计算[20]:

$$p\text{-value}=1-\sum_{i=0}^{k-1}\frac{\begin{pmatrix}|F|\\ i\end{pmatrix}\begin{pmatrix}|V|-|F|\\ |C|-i\end{pmatrix}}{\begin{pmatrix}|V|\\ |C|\end{pmatrix}} \tag{8}$$

其中, $|V|$ 为 PPI 网络中包含的蛋白质总数, C 为一个预测得到的模块, F 为一个功能组, k 为 C 中包含 F 中的蛋白质数目, p-值度量又称为功能同质性指标. 基于该式的含义, 一个预测得到的功能模块所对应的功能应是在所有候选功能组中具有最小 p-值的那个功能组. 也就是说, 一个具有低 p-值的预测模块表示它被来自于同一功能组中的蛋白质所充实, 所以它更有可能是真正的蛋白质功能模块. 因此, 通过设置一个具有统计意义的公共阈值, 由检测方法得到的那些低于公共阈值的功能模块的数量就能够体现每种方法各自的综合性能.

3.3 几种典型检测算法的性能对比

基于上述评估方法, 我们从分类体系中的每一类中选择一种具有代表性的蛋白质功能模块的检测算法进行了测试、实验对比和分析. 这 7 种算法包括 MCODE、Jerarca、RNSC、MCL、ADMSC、Core 和 ACO-MAE. 实验数据来自最新版本的 DIP 数

据库 (20120818full), 原始数据包含 5 084 个蛋白质节点和 22 466 对相互作用, 是一个含有较大规模和较高噪声的代表性数据集. 通过清除数据库中的冗余数据 (自相互作用和重复的相互作用) 后, 可以得到包含 5 032 个蛋白质节点和 22 091 对相互作用的数据. 在如下实验中只把包含蛋白质数量大于 3 的聚类集合作为检测到的有效功能模块来对检测结果进行性能的评价与分析. 选取的基准功能模块集包含 428 个模块, 由 MIPS[78]、Aloy[79] 和 SGD 数据库[80] 集成得到.

图 3 列出了 7 种不同算法在 Precision、Recall、F-measure、Sn、PPV 和 Acc 6 项评价指标上的实验对比. 不难发现, MCODE、Core 和 ACO-MAE 算法在 Precision 上取得了较好的结果, 其中 MCODE 获得了最高值 0.473, 相比其他算法有较明显的优势. ACO-MAE、RNSC、Core、Jerarca 和 MCL 在 Recall 上取得较好的值, 分别为 0.549、0.442、0.428、0.418 和 0.416. 而在 F-measure 上取得最高值的 3 个算法分别是 Core (0.360)、ACO-MAE (0.335) 和 RNSC (0.292). 尽管 MCODE 在 Precision 上取得了最大值, 但是其 Recall 值是最小的, 所以在它们的综合评价指标 F-measure 上, MCODE 未能取得好的表现. 另外, ADMSC 算法在这 3 项指标上均没有突出的表现. 尽管 Core 算法在 Precision 和 Recall 上都不是最佳的结果, 但是均达到了较好的性能, 所以在 F-measure 上 Core 算法取得了最好值. 在 Sn 指标上最好的 4 个算法是 ADMSC、ACO-MAE、Core 和 MCL, 它们的取值分别是 0.400、0.370、0.350 和 0.321. 而算法 RNSC、ACO-MAE、Jerarca 和 MCL 在 PPV 上取得了较好的表现, 对应的评价值分别为 0.361、0.342、0.317 和 0.311. 在 Acc 指标上, 获得最好表现的 4 种算法分别是 ACO-MAE (0.356)、ADMSC (0.333)、MCL (0.316) 和 RNSC (0.310). 在后 3 项指标的评价中, MCODE 也均没能取得突出的表现, 而 ACO-MAE 获得了较好的结果. 总之, 不同算法在 6 个不同的评价指标上各有优劣, 没有一种算法能够在各项指标都能得到最好的结果, 这也说明不同的聚类机理对模块的检测都具有一定的偏爱. 相对而言, 算法 ACO-MAE 在 6 项评价指标上表现得更稳定一些.

表 1 标注了 7 种典型算法所属的类型、检测结果的节点覆盖率 (能够识别的蛋白质的比率), 并列出了所检测的模块在三种不同生物学意义 (生物过程、细胞组成和模块功能) 上的最小的 p-值. 基于密度的检测方法, 由于直接选取 PPI 网络中的稠密子图作为聚类结果, 因此该方法的检测结果中容易丢弃大量的节点, 检测到的模块所包含的节点仅占网络节点数的 16.1 %; 层次方法 Jerarca、流方法 MCL 以及谱方法 ADMSC 聚类结果的节点覆盖率较高, 尤其是 ADMSC 方法, 它将网络中的绝大多数节点都聚类到了不同的模块中, 因此覆盖率达到最高的 98.9 %; 基于划分的 RNSC 方法, 由于加入 3 种严格的过滤策略, 因此检测模块具有较低的覆盖率; Core 是基于核心 – 附属关系的检测方法, 它要选取近似稠密子图的 “核心” 区域, 因此该方法在聚类结果中也丢弃了大量的节点. 基于群集智能的 ACO-MAE 方法, 将聚类问题转换为最优路径问题, 避免了仅检测稠密子图的缺陷, 所以丢弃较少的节点. 后 3 列的 p-值分别列举了在生物过程、细胞组成和模块功能三种不同生物意义上所得到的最小的 p-值. 从表 1 可见, 尽管这些算法的检测机制不尽相同, 但它们都能有效检测出具有较高生物学意义的功能模块.

表 2 列出了这 7 种典型算法的实现语言、是否开源、运行平台、使用难易, 以及在获得最后结果时在时空资源上的占用情况, 其中, Jerarca 和 ACO-

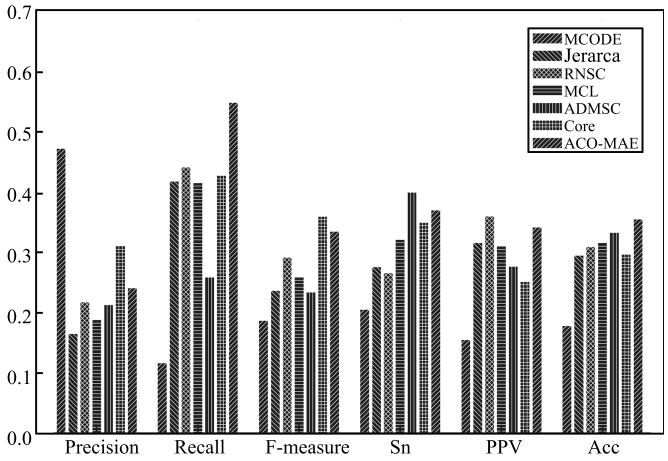


图 3　几种典型算法的评价指标比较

Fig. 3　Comparative results of several typical algorithms in terms of various evaluation metrics

表 1 不同算法的最小 p-值列举

Table 1 The minimum p-value of different algorithms

方法	类型	覆盖率	p-value		
			Biological process	Cellular component	Molecular function
MCODE	密度	0.1611	1.95E−27	2.55E−31	4.78E−10
Jerarca	层次	0.7696	1.90E−23	3.29E−23	3.92E−24
RNSC	划分	0.4253	6.91E−28	9.92E−39	4.71E−11
MCL	流模拟	0.8225	2.89E−34	2.64E−49	2.85E−33
ADMSC	谱方法	0.9893	6.01E−36	1.06E−51	8.05E−35
Core	核心−附属	0.3432	4.97E−38	3.54E−59	3.37E−20
ACO-MAE	群智能	0.6222	1.60E−57	1.98E−63	4.18E−16

表 2 不同算法的运行情况

Table 2 The running status of several different algorithms

方法	程序/语言	是否开源	运行平台	是否易用	占用内存 (MB)	运行时间 (s)
MCODE	安装包	否	Windows	是	260	10
Jerarca	二进制	否	Windows/Linux	是	1 092	1 895
RNSC	C++	是	Linux	多步运行	2.2	5
MCL	C	是	Linux	是	17	4
ADMSC	Matlab	是	Windows	需参数设置	1 050	523
Core	Matlab	是	Windows	是	264	78 024
ACO-MAE	C++	是	Windows	需参数设置	1 176	2 767

MAE 为迭代算法, 所列时间分别为迭代 10 次、67 次的结果. 从表 2 可见, 这些算法在检测过程中所需要的时空资源相差很大, RNSC 是其中占用内存资源最少的算法, MCL 是运行最快的算法, Core 所需的运行时间最长, 而 ACO-MAE 则是占用内存资源最多的算法.

为了能够更具体、形象地显示检测到的结果, 在图 4 中我们给出了各算法检测到的一个功能模块的实例对比. 图 4 (a) 是由 20 个蛋白质组成的标准功能模块 (mRNA cleavage factor module). 图中圆形节点表示组成该模块的蛋白质, 节点标注即为蛋白质的名字, 节点间的连线为该 DIP 数据库中蛋白质之间的相互作用. 图 4 (b) 是 MCODE 方法的检测结果, 它检测出的模块包含 13 个节点, 它们全部和标准的模块匹配; 图 4 (c) 是算法 Jerarca 检测出的结果, 仅包含 12 个节点, 是所有方法中包含节点最少的, 而且其中包含 3 个噪声节点 (方形节点); 图 4 (d) 是 RNSC 方法的检测结果, 它包含 17 个蛋白质, 其中 16 个与标准功能模块所对应的蛋白质匹配; 图 4 (e) 是流方法 MCL 的检测结果, 该方法检测的模块所包含的节点数是最多的, 由 26 个节点组成, 但其中仅有 18 个能和标准模块匹配; 图 4 (f) 是 ADMSC 算法的检测结果, 和标准模块一样由 20 个节点组成, 但检测得到的模块中有 4 个噪声节点; 图 4 (g) 是 Core 算法的检测结果, 和 RNSC 算法检测的结果类似, 均有 17 个节点组成, 且都包含同一个噪声节点 yjl033w, 但它们的匹配节点不完全相同; 图 4 (h) 是 ACO-MAE 算法的检测结果, 它由 19 个节点组成, 其中有 18 个节点和标准模块所含的蛋白质相匹配.

3.4 相关的软件和工具

3.4.1 功能富集分析软件

功能富集分析的统计原理是用超几何分布模型来检验一组基因 (或基因产物) 中某个功能类的显著性. 通常采用 p-值度量来分析蛋白质功能模块在某项功能上的富集程度.

1) DAVID (The database for annotation, visualization and integrated discovery, http://david.abcc.ncifcrf.gov/home.jsp)[81] 是一个生物信息数据平台, 它集成了一些生物学数据和相关的分析工具, 为大规模的基因和蛋白质列表提供综合性的

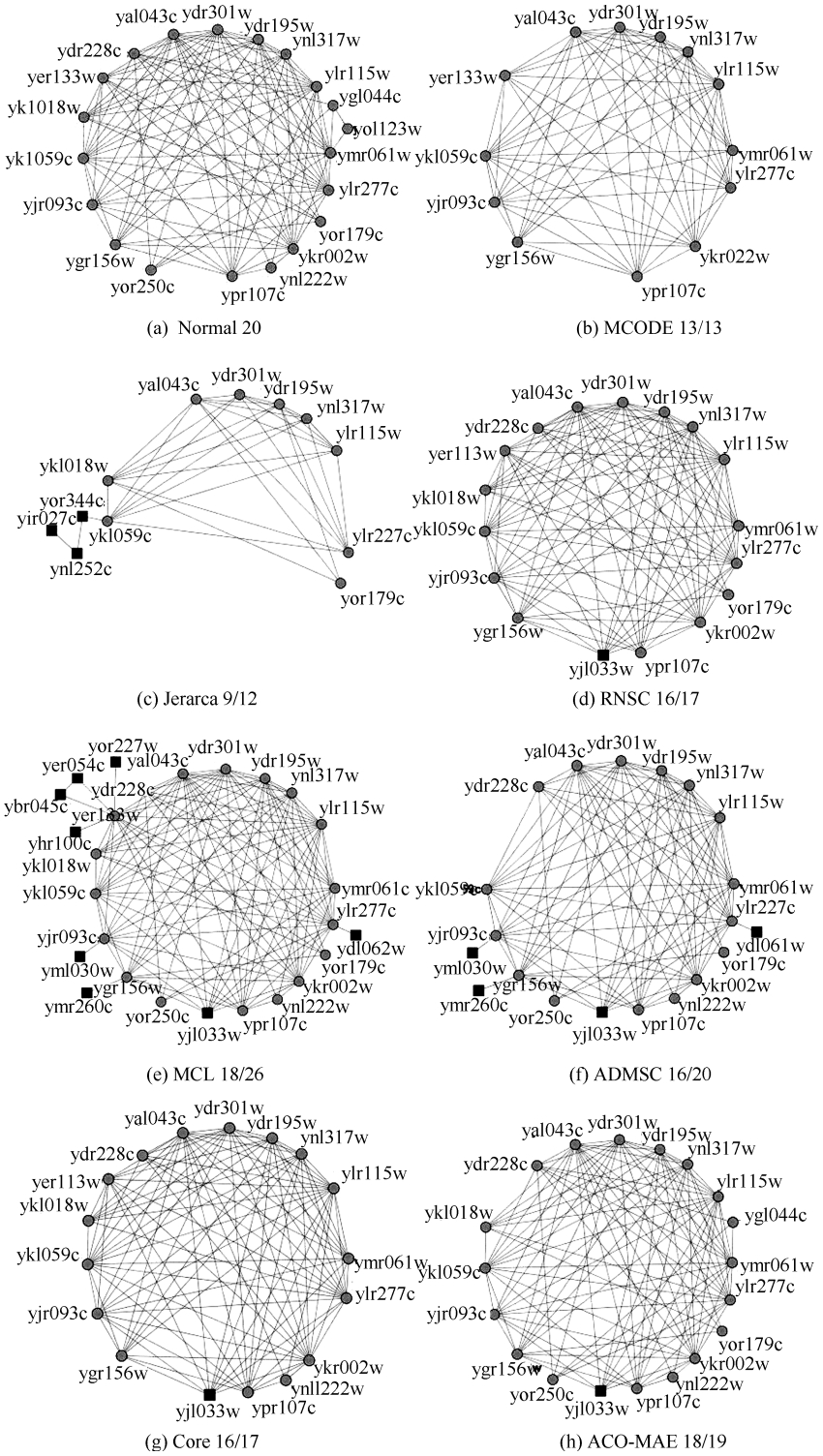


图 4　不同算法在模块 Anaphase promoting 上的实例对比

Fig. 4　One instance comparison on Anaphase promoting module for different algorithms

生物功能注释信息. DAVID 从 2003 年第 1 版发布以来, 目前已发展到 6.7 版, 它由一个知识库和 5 个基于网络功能注释的工具套件组成. 这些工具包括基因功能分类工具、功能注释工具、基因 ID 转换工具、基因名称查看器和病原体基因组浏览器. 用户上传自己的基因列表后, 该平台不仅能提供典型的基因 – 术语富集分析, 而且能提供相似基因查询、基因列表浓缩以及基因和蛋白质标识的转换等服务.

2) AmiGO (http://amigo.geneontology.org/cgi-bin/amigo/term_enrichment)[82] 是由 GO 国际协会负责开发的一个网络应用程序, 该程序不仅允许用户查询、浏览 GO 协会提供的基因本体和注释数据, 而且能够可视化基因本体及其相关注释信息. 只要用户提供自己检测的基因产物 (功能模块) 及基因产物来源的背景数据集, 便可通过该程序获得在生物过程、细胞组成和模块功能三种不同生物意义上计算得到的生物富集 p-值. 另外, AmiGO 也是一个开源软件, 可以被下载和安装到本地进行浏览和注释.

3) 基因本体富集分析软件工具箱 (Gene ontology enrichment analysis software toolkit, GOEAST (http://omicslab.genetics.ac.cn/GOEAST/))[83] 是由中国科学院遗传与生物研究所开发的基因本体富集分析的在线软件工具包. GOEAST 具有 3 个主要特征: a) 可以根据术语在三个 GO 分类 (生物过程、分子功能和细胞组成) 层次图中的关系来图形化地显示富集的 GO 术语, 从而对富集 GO 术语间的相关性提供更好的理解; b) 能够支持多种来源和多个物种 (60 多种原核生物和真核生物物种) 的数据分析; c) 允许通过多个实验的 GO 富集状态的交叉比较来识别它们之间的功能相关性.

4) 京都基因与基因组百科全书 (Kyoto encyclopedia of genes and genomes, KEGG (http://www.genome.ad.jp/kegg/))[84] 是一个能系统分析基因功能、带高阶函数信息关联的基因组信息的知识库. 它包含 3 个数据库: a) PATHWAY 库用于根据相互作用分子网络来完成高阶函数的表示; b) GENES 库用于存储所有全序基因组和一些偏序基因组的基因类别; c) LIGAND 用于存储在细胞、酶分子和酶活动中的化学组合物. KEGG 既提供了 JAVA 图形工具来浏览、比较和操作基因组图, 也提供了计算工具来完成序列、图形对比和路径计算. KEGG 数据库信息每天都在更新, 而且资源的使用一直是免费的和开放的.

5) GO: TermFinder[85] 是一个开源软件, 可用于访问基因本体信息和查找具有显著性富集的指定基因组的本体术语. 它由一套面向对象的 Perl 模块组成, 这些模块能访问基因本体信息, 并按 GO 术语来评估和可视化一组基因的共同注释. 它能够通过计算每个基因注释的统计显著性, 从微阵列和其他生物学数据中获得有用的结论. 该软件能够在 Perl 可运行的任何系统中使用, 使用方式既可作为一个命令行应用程序以单或批量模式来运行, 也可作为一个基于 Web 的 CGI 脚本来运行.

3.4.2 常用可视化软件

蛋白质相互作用及其网络的可视化一直以来是蛋白质组学研究中的一种辅助工具, 前面提到的部分数据库和软件平台都包含了一些可视化的功能. 除此之外, 还有如下一些常用的可视化软件.

1) Cytoscape[86] 是由 Leroy Hood 实验室 (美国系统生物学研究所)、Trey Ideker 实验室 (加州大学圣地亚哥分校) 等单位联合开发的一个开源的网络可视化、数据集成分析的软件包, 其开发和使用的主要目的是为了满足系统生物学的建模需求. 由于提供了专用于数据分析、数据集成和可视化的插件框架, 所以 Cytoscape 能被灵活地应用于系统生物学的许多研究问题. 比如, 它可用于大规模蛋白质相互作用、蛋白质–DNA 和遗传交互作用的分析, 并可通过与功能注释数据库的连接, 形象地进行网络整合以及谱、表型及分子状态的表达. 由于 Cytoscape 可以通过简易插件对其功能进行扩展, 所以软件包的功能日益丰富和完善.

2) Osprey[87] 是由多伦多大学的生物信息学研究组开发的一个关于复杂相互作用网络可视化的系统, 它是一个能够促进蛋白质相互作用网络和蛋白质功能模块研究的开源软件平台. Osprey 能生成数据丰富的图形化表示, 并用不同颜色来表示节点的基因功能及节点间的相互作用. 当鼠标停留在对象上时, 它以分支建模的形式快速定制和组织网络图. 尽管 Osprey 本身主要依赖于 BIND 和 GRID 两个蛋白质相互作用数据库, 但用户定义的大规模数据集也能够快速地与 Osprey 结合来完成不同方法的比较.

3) CFinder[22] 是由匈牙利布达佩斯大学生物物理系与匈牙利科学院生物物理研究所共同开发的一种基于团 (Clique) 渗透方法发现和可视化复杂网络中具有交叠特征浓密子图的软件. 该软件可完成在网络中寻找团, 把具有交叠的团用不同的颜色表示, 并加不同注释等功能. CFinder 可以运行在 Windows、Linux 和 Macintosh 等环境, 为基因网络、社会网络和能用图表示的微阵列数据提供了一种快速、高效的聚类方法.

4 功能模块检测的应用、存在的问题与研究展望

在 PPI 网络中准确地进行功能模块检测, 对于了解生命过程中某些未知蛋白质的功能, 理解生命活动的本质, 进行疾病的诊断以及病理的研究具有重要的实际意义. 例如, 在蛋白质功能预测方面, 如果所发现的模块中大部分蛋白质都具有某项功能, 那么可以预测同模块中未注释的蛋白质也应该具有

该功能; 在理解网络的结构和功能方面, 由于细胞的功能和生化反应都是由位于相同功能模块中彼此相互作用的蛋白质共同协调完成的, 所以在 PPI 网络中进行功能模块检测是正确理解这些基础细胞网络结构和功能的重要手段; 在疾病的诊断和药物研发方面, 如果某些特殊蛋白质相互作用发生异常时会导致疾病, 而通过功能模块检测能够获悉该特殊蛋白质与哪些蛋白质发生作用, 那么就可以设计药物去防止和阻断该相互作用的发生, 进而达到疾病治疗的目的.

正是由于功能模块检测在医药研究中具有潜在的应用价值, 所以通过生物实验来识别功能模块的方法一直是过去系统生物学中的一项主要内容. 但是, 伴随蛋白质相互作用数据的与日剧增, 人们通过实验方法确定 PPI 网络中蛋白质功能模块已经难以满足后基因时代人类对生物体行为理解、蛋白质功能预测和药物设计的需要. 因此利用复杂网络、图论和信息论等领域的一些计算方法及时、高效地进行功能模块的检测无疑是解决该问题的有效途径. 尽管近年来人们已经将机器学习、数据挖掘中的一些理论和方法运用于 PPI 网络功能模块的检测, 也取得了许多丰硕的成果. 但是随着研究的深入, 一些极具挑战性的问题被凸显出来:

1) 相互作用数据不可靠问题: 从生物实验获得的 PPI 数据完整性不足, 存在着许多噪声数据. 这是由于一方面高通量蛋白质相互作用的数据比小规模的数据有着更高的假阳性率; 另一方面, 受实验方法的局限性, 数据库中有可能丢失一些实际存在的相互作用 (假阴性). 因此, 为获得好的挖掘结果, 必须提高检测算法对假阳性、假阴性的识别能力, 也就是说, 需要加强算法的鲁棒性. 尽管目前许多研究都从多种信息融合的角度给出了一些有效的方法, 但克服噪声数据对检测准确率的影响仍然是功能模块检测将面临的重要问题.

2) 检测算法效率问题: 一个 PPI 网络通常具有成千上万的蛋白质及更多的相互作用关系, 是一个大规模的复杂网络. 于是在利用计算方法进行 PPI 网络功能模块检测时, 必须要解决的一个现实问题是如何提高所采用算法的时间性能以更好地满足实际需求. 目前所提出的大多数方法的时间复杂度都较高, 这将制约基于计算方法进行 PPI 网络功能模块检测的发展和实际应用. 所以对该问题的研究应引起人们更多的重视.

3) 功能模块的重叠问题: PPI 网络中的一个蛋白质有可能参与多个不同的生物学功能, 即一个蛋白质有可能属于多个不同的功能模块, 可见, 簇的重叠是功能模块检测方法必须要解决的重要问题. 近年来, 已有一些学者对该问题进行了研究和探索, 提出了一些有针对性的方法. 但结合生物机理对其进行的系统性研究还较缺乏. 因此, 如何在 PPI 网络中进行功能模块的重叠检测仍将是 PPI 网络功能模块检测中一个极具挑战性的研究问题.

此外, PPI 网络本身的拓扑特性也给蛋白质功能模块的检测带来了很大困难, 如 PPI 网络的无尺度分布会导致簇的大小严重不均匀、不同功能组之间的蛋白质存在频繁连接引起的模块界限的模糊等都会使蛋白质功能模块的检测成为一项极为困难的研究任务. 面对这些挑战, 结合后基因时代生物信息学的实际需求, 如下几方面的问题将是该领域今后重要的发展方向.

1) 相互作用可靠性的计算仍将是研究热点: 目前, 学者们已经意识到使用计算方法进行 PPI 网络功能模块检测的效果在很大程度上依赖于相互作用数据的质量, 并开始在预处理阶段融合多种生物信息进行相互作用的可靠性评测和判断, 但是冗余、不完全及噪声数据对功能模块检测质量的影响仍然非常严重. 所以, 研究如何利用鲁棒的计算方法去验证已存在相互作用的可靠性或利用已有数据预测真实的相互作用将是克服数据不良影响, 发挥功能模块检测计算方法作用的关键.

2) 新型融合方法和检测技术将发挥重要作用: 为了提高功能模块检测的准确性, 人们已经通过本体、贝叶斯网、神经网络、核方法等技术将拓扑结构、功能注释以及生物进化等信息融合到蛋白质功能模块检测过程中, 取得了一定的效果. 随着系统生物学研究的进一步深入, 人们获得的蛋白质相互作用的相关证据将与日俱增. 利用先进计算机理论和技术进行多源知识融合, 设计新颖、高效且准确检测重叠模块的搜索机制将是该领域未来追求的研究目标.

3) 功能模块的检测将与生物进化的研究紧密结合: PPI 网络的形成一方面为预测未知的蛋白质功能, 了解特定生物功能的分子机制提供了基础, 另一方面也为生物进化研究提供了新的视角. 由于 PPI 网络所具有的许多性质都是通过在适应环境中形成的, 所以功能模块的检测与生物进化的研究有必然内在的联系. 目前已有一些研究通过网络比对方法在不同生物种类的 PPI 网络上发现保守的功能模块[88−91], 但这些方法的研究和应用还没形成更大规模. 但可以相信, 随着 PPI 网络数据的发展和学术研究的需要, PPI 网络功能模块的检测与生物进化的广泛结合将成为必然趋势.

4) 与动态 PPI 网络相关的功能模块检测研究将成为新的研究热点: 目前使用的 PPI 网络是一个高度平均和理想化的结构[92], 它包含了细胞中不同条件、不同时间、甚至不同空间上的各种相互作用,

这种结构不仅会引起相互作用数据的高噪声, 而且其反映出的一些细胞网络的特点可能并非是真实的. 而动态 PPI 网络更接近于细胞的真实情况, 所以对动态 PPI 网络进行功能模块检测的研究对于揭示潜在的疾病标志物和药物靶标将更具实际意义.

致谢

感谢美国纽约州立大学布法罗分校计算机系主任 Aidong Zhang 教授, 与她的交流和讨论使我们对蛋白质功能模块检测方面的研究有了更完整的理解. 此外, 感谢给予本文指导和建议的同行专家.

References

1 Zhang A D. *Protein Interaction Networks: Computational Analysis*. Cambridge: Cambridge University Press, 2009

2 Li Man-Sheng, Liu Qi-Jun, Li Dong, Liu Pei-Lei, Zhu Yun-Ping. Progress of literature mining for protein-protein interaction information. *Science in China (Series C)*, 2010, **40**(9): 805−819
(李满生, 刘齐军, 李栋, 刘培磊, 朱云平. 蛋白质相互作用信息的文本挖掘研究进展. 中国科学: 生命科学, 2010, **40**(9): 805−819)

3 Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National academy of Sciences of the United States of America*, 2001, **98**(8): 4569−4574

4 Uetz, P, Giot L, Cagney G, Mansfield T A, Judson R S, Knight J R, Lockshon D, Narayan V, Srinivasan M, Pochart P, Qureshi-Emili A, Li Y, Godwin B, Conover D, Kalbfleisch T, Vijayadamodar G, Yang M J, Johnston M, Fields S, Rothberg J M. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature*, 2000, **403**(6770): 623−627

5 Aebersold R, Mann M. Mass spectrometry-based proteomics. *Nature*, 2003, **422**(6928): 198−207

6 Ho Y, Gruhler A, Heilbut A, Bader G D, Moore L, Adams S L, Millar A, Taylor P, Bennett K, Boutilier K, Yang L Y, Wolting C, Donaldson I, Schandorff S, Shewnarane J, Vo M, Taggart J, Goudreault M, Muskat B, Alfarano C, Dewar D, Lin Z, Michalickova K, Willems A R, Sassi H, Nielsen P A, Rasmussen K J, Andersen J R, Johansen L E, Hansen L H, Jespersen H, Podtelejnikov A, Nielsen E, Crawford J, Poulsen V, Sorensen B D, Matthiesen J, Hendrickson R C, Gleeson F, Pawson T, Moran M F, Durocher D, Mann M, Hogue C W V, Figeys D, Tyers M. Systematic identification of protein complexes in saccharomyces cerevisiae by mass spectrometry. *Nature*, 2002, **415**(6868): 180−183

7 Macbeath G, Schreiber S L. Printing proteins as microarrays for high-throughput function determination. *Science*, 2000, **289**(5485): 1760−1763

8 Zhu H, Bilgin M, Bangham R, Hall D, Casamayor A, Bertone P, Lan N, Jansen R, Bidlingmaier S, Houfek T, Mitchell T, Miller P, Dean R A, Gerstein M, Snyder M. Global analysis of protein activities using proteome chips. *Science*, 2001, **293**(5537): 2101−2105

9 Peri S, Navarro J D, Amanchy R, Kristiansen T Z, Jonnalagadda C K, Surendranath V, Niranjan V, Muthusamy B, Gandhi T K B, Gronborg M, Ibarrola N, Deshpande N, Shanker K, Shivashankar H N, Rashmi B P, Ramya M A, Zhao Z X, Chandrika K N, Padma N, Harsha H C, Yatish A J, Kavitha M P, Menezes M, Choudhury D R, Suresh S, Ghosh N, Saravana R, Chandran S, Krishna S, Joy M, Anand S K, Madavan V, Joseph A, Wong G W, Schiemann W P, Constantinescu S N, Huang L, Khosravi-Far R, Steen H, Tewari M, Ghaffari S, Blobe G C, Dang C V, Garcia J G N, Pevsner J, Jensen O N, Roepstorff P, Deshpande K S, Chinnaiyan A M, Hamosh A, Chakravarti A, Pandey A. Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Research*, 2003, **13**(10): 2363−2371

10 Donaldson I, Martin J, de Bruijn B, Wolting C, Lay V, Tuekam B, Zhang S D, Baskin B, Bader G D, Michalickova K, Pawson T, Hogue C W V. PreBIND and textomy-mining the biomedical literature for protein-protein interactions using a support vector machine. *BMC Bioinformatics*, 2003, **4**: 11

11 Corney D P A, Buxton B F, Langdon W B, Jones D T. BioRAT: extracting biological information from full-length papers. *Bioinformatics*, 2004, **20**(17): 3206−3213

12 Zhang Chun-Ting. The current status and the prospect of bioinformatics. *World Sci-Tech R and D*, 2000, **22**(6): 17−20
(张春霆. 生物信息学的现状与展望. 世界科技研究与发展, 2000, **22**(6): 17−20)

13 Gavin A C, Böesche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick J M, Michon A M, Cruciat C M, Remor M, Höfert C, Schelder M, Brajenovic M, Ruffner H, Merino A, Klein K, Hudak M, Dickson D, Rudi T, Gnau V, Bauch A, Bastuck S, Huhse B, Leutwein C, Heurtier M A, Copley R R, Edelmann A, Querfurth E, Rybin V, Drewes G, Raida M, Bouwmeester T, Bork P, Seraphin B, Kuster B, Neubauer G, Superti-Furga G. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 2002, **415**(6868): 141−147

14 Tarassov K, Messier V, Landry C R, Radinovic S, Molina M M, Shames I, Malitskaya Y, Vogel J, Bussey H, Michnick S W. An in vivo map of the yeast protein interactome. *Science*, 2008, **320**(5882): 1465−1470

15 Tong A H Y, Drees B, Nardelli G, Bader G D, Brannetti B, Castagnoli L, Evangelista M, Ferracuti S, Nelson B, Paoluzi S, Quondam M, Zucconi A, Hogue C W V, Fields S, Boone C, Cesareni G. A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. *Science*, 2002, **295**(5553): 321−324

16 Palla G, Derényi I, Farkas I, Vicsek T. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 2005, **435**(7043): 814−818

17 Ravasz E, Somera A L, Mongru D A, Oltvai Z N, Barabasi A L. Hierarchical organization of modularity in metabolic networks. *Science*, 2002, **297**(5586): 1551−1555

18 Rives A W, Galitski T. Modular organization of cellular networks. *Proceedings of the National Academy of Sciences of the United States of America*, 2003, **100**(3): 1128−1133

19 Tanay A, Sharan R, Kupiec M, Shamir R. Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. *Proceedings of the National Academy of Sciences of the United States of America*, 2004, **101**(9): 2981−2986

20 Bu D B, Zhao Y, Cai L, Xue H, Zhu X P, Lu H C, Zhang J F, Sun S W, Ling L J, Zhang N, Li G J, Chen R S. Topological structure analysis of the protein-protein interaction network in budding yeast. *Nucleic Acids Research*, 2003, **31**(9): 2443−2450

21 Glazko G, Gordon A, Mushegian A. The choice of optimal distance measure in genome-wide datasets. *Bioinformatics*, 2005, **21**(S3): iii3−iii11

22 Adamcsek B, Palla G, Farkas I J, Derényi I, Vicsek T. CFinder: locating cliques and overlapping modules in biological networks. *Bioinformatics*, 2006, **22**(8): 1021−1023

23 Arnau V, Mars S, Marin I. Iterative cluster analysis of protein interaction data. *Bioinformatics*, 2005, **21**(3): 364−378

24 Li D, Li J, Ouyang S G, Wang J, Wu S F, Wan P, Zhu Y P, Xu X J, He F C. Protein interaction networks of Saccharomyces cerevisiae, Caenorhabditis elegans and Drosophila melanogaster: large-scale organization and robustness. *Proteomics*, 2006, **6**(2): 456−461

25 Stumpf M P H, Kelly W P, Thorne T, Wiuf C. Evolution at the system level: the natural history of protein interaction networks. *Trends in Ecology and Evolution*, 2007, **22**(7): 366−373

26 Strogatz S H. Exploring complex networks. *Nature*, 2001, **410**(6854): 268−276

27 Sprinzak E, Sattath S, Margalit H. How reliable are experimental protein-protein interaction data? *Journal of Molecular Biology*, 2003, **327**(5): 919−923

28 Brun C, Chevenet F, Martin D, Wojcik J, Guenoche A, Jacq B. Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network. *Genome Biology*, 2004, **5**(1): R6

29 Pei P J, Zhang A D. A topological measurement for weighted protein interaction network. In: Proceedings of the 16th IEEE Computational Systems Bioinformatics Conference. Washington DC, USA: IEEE, 2005. 268−278

30 Resink P. Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, 1999, **11**: 95−130

31 Jiang J J, Conrath D W. Semantic similarity based on corpus statistics and lexical taxonomy. In: Proceedings of the 1997 ROCLING X, International Conference on Research in Computational Linguistics. Taiwan, China, 1997. 19−33

32 Lin D K. An information-theoretic definition of similarity. In: Proceedings of the 15th International Conference Machine Learning. San Francisco, CA: Morgan Kaufmann, 1998. 296−304

33 King A D, Przulj N, Jurisica I. Protein complex prediction via cost-based clustering. *Bioinformatics*, 2004, **20**(17): 3013−3020

34 Li X L, Foo C S, Ng S K. Discovering protein complexes in dense reliable neighborhoods of protein interaction networks. *Computer Society Bioinformatics Conference — CSB*, 2007, **6**: 157−168

35 Hwang W, Cho Y R, Zhang A D, Ramanathan M. CASCADE: a novel quasi all paths-based network analysis algorithm for clustering biological interactions. *BMC Bioinformatics*, 2008, **9**: 64

36 Spirin V, Mirny L A. Protein complexes and functional modules in molecular networks. *Proceedings of the National Academy of Sciences of the United States of America*, 2003, **100**(21): 12123−12128

37 Bader G D, Hogue C W V. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, 2003, **4**: 2

38 Altaf-UI-Adim M, Shinbo Y, Mihara K, Kurokawa K, Kanaya S. Development and implementation of an algorithm for detection of protein complexes in large interaction networks. *BMC Bioinformatics*, 2006, **7**: 207

39 Pei P J, Zhang A D. A "seed-refine" algorithm for detecting protein complexes from protein interaction data. *IEEE Transactions on NanoBioscience*, 2007, **6**(1): 43−50

40 Aldecoa R, Marin I. Jerarca: efficient analysis of complex networks using hierarchical clustering. *PLoS ONE*, 2010, **5**(7): e11585

41 Holme P, Huss M, Jeong H. Subnetwork hierarchies of biochemical pathways. *Bioinformatics*, 2003, **19**(4): 532−538

42 Dunn R, Dudbridge F, Sanderson C M. The use of edge-betweenness clustering to investigate biological function in protein interaction networks. *BMC Bioinformatics*, 2005, **6**: 39

43 Girvan M, Newman M E J. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, 2002, **99**(12): 7821−7826

44 Frey B J, Dueck D. Clustering by passing messages between data points. *Science*, 2007, **315**(5814): 972−976

45 Nabieva E, Jim K, Agarwal A, Chazelle B, Singh M. Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. *Bioinformatics*, 2005, **21**(S1): i302−i310

46 Van Dongen S. A Cluster Algorithm for Graphs. Technical Report INS-R0010, National Research Institute for Mathematics and Computer Science in the Netherlands, 2000

47 Hwang W, Cho Y R, Zhang A D, Ramanathan M. A novel functional module detection algorithm for protein-protein interaction networks. *Algorithms for Molecular Biology*, 2006, **1**: 24

48 Cho Y R, Hwang W, Zhang A D. Optimizing flow-based modularization by iterative Centroid search in protein interaction networks. In: Proceedings of the 7th IEEE International Conference on Bioinformatics and Bioengineering (BIBM). Boston, MA: IEEE, 2007. 342−349

49 Cho Y R, Hwang W, Ramanathan M, Zhang A D. Semantic integration to identify overlapping functional modules in protein interaction networks. *BMC Bioinformatics*, 2007, **8**: 265

50 Feng J X, Jiang R, Jiang T. A max-flow based approach to the identification of protein complexes using protein interaction and microarray data. *Computer Society Bioinformatics Conference-CSB*, 2008, **7**: 51−62

51 Yang Bo, Liu Da-You, Liu Ji-Ming, Jin Di, Ma Hai-Bin. Complex network clustering algorithms. *Journal of Software*, 2009, **20**(1): 54−66
(杨博, 刘大有, Liu Ji-Ming, 金弟, 马海宾. 复杂网络聚类方法. . 软件学报, 2009, **20**(1): 54−66)

52 Kamp C, Christensen K. Spectral analysis of protein-protein interactions in Drosophila melanogaster. *Physical Review*, 2005, **71**(4): 041911

53 Sen T Z, Kloczkowski A, Jernigan R L. Functional clustering of yeast proteins from the protein-protein interaction network. *BMC Bioinformatics*, 2006, **7**: 355

54 Qin G M, Gao L. Spectral clustering for detecting protein complexes in protein-protein interaction (PPI) networks. *Mathematical and Computer Modelling*, 2010, **52**(11−12): 2066−2074

55 Inoue K, Li W J, Kurata H. Diffusion model based spectral clustering for protein-protein interaction networks. *PLoS ONE*, 2010, **5**(9): e12623

56 Gavin A C, Aloy P, Gradi P, Krause R, Boesche M, Marzioch M, Rau C, Jensen L J, Bastuck S, Dümpelfeld B, Edelmann A, Heurtier M A, Hoffman V, Hoefert C, Klein K, Hudak M, Michon A M, Schelder M, Schirle M, Remor M, Rudi T, Hooper S, Bauer A, Bouwmeester T, Casari G, Drewes G, Neubauer G, Rick J M, Kuster B, Bork P, Russell R B, Superti-Furga G. Proteome survey reveals modularity of the yeast cell machinery. *Nature*, 2006, **440**(7084): 631−636

57 Leung H C, Yiu S M, Xiang Q, Chin F Y. Predicting protein complexes from PPI data: a core-attachment approach. *Journal of Computational Biology*, 2009, **16**(2): 133−144

58 Wu M, Li X L, Kwoh C K, Ng S K. A core-attachment based method to detect protein complexes in PPI networks. *BMC Bioinformatics*, 2009, **10**: 169

59 Ma X K, Gao L. Predicting protein complexes in protein interaction networks using a core-attachment algorithm based on graph communicability. *Information Sciences*, 2012, **189**: 233−254

60 Sallim J, Abdullah R, Khader A T. ACOPIN: an ACO algorithm with TSP approach for clustering proteins from protein interaction network. In: Proceedings of the 2nd UKSIM European Symposium on Computer Modeling and Simulation. Liverpool: IEEE, 2008. 203−208

61 Ji J Z, Liu Z J, Zhang A D, Jiao L, Liu C N. Improved ant colony optimization for detecting functional modules in protein-protein interaction networks. In: Proceedings of the 3rd International Conference on Information Computing and Applications. Berlin, Heidelberg: Springer, 2012. 404−413

62 Ji J Z, Liu Z J, Zhang A D, Jiao L, Liu C N. Ant colony optimization with multi-agent evolution for detecting functional modules in protein-protein interaction networks. In: Proceedings of the 3rd International Conference on Information Computing and Applications. Berlin, Heidelberg: Springer, 2012. 445−453

63 Lei Xiu-Juan, Huang Xu, Wu Shuang, Guo Ling. Joint strength based ant colony optimization clustering algorithm for PPI networks. *Acta Electronica Sinica*, 2012, **40**(4): 695−702
(雷秀娟, 黄旭, 吴爽, 郭玲. 基于连接强度的 PPI 网络蚁群优化聚类算法. 电子学报, 2012, **40**(4): 695−702)

64 Pagel P, Kovac S, Oesterheld M, Brauner B, Dunger-Kaltenbach I, Frishman G, Montrone C, Mark P, Stumpflen V, Mewes H W, Ruepp A, Frishman D. The MIPS mammalian protein-protein interaction database. *Bioinformatics*, 2005, **21**(6): 832−834

65 Salwinski L, Miller C S, Smith A J, Pettit F K, Bowie J U, Eisenberg D. The database of interacting proteins: 2004 update. *Nucleic Acids Research*, 2004, 32: D449−D451

66 Alfarano C, Andrade C E, Anthony K, Bahroos N, Bajec M, Bantoft K, Betel D, Bobechko B, Boutilier K, Burgess E, Buzadzija K, Cavero R, D′Abreo C, Donaldson I, Dorairajoo D, Dumontier M J, Dumontier M R, Earles V, Farrall R, Feldman H, Garderman E, Gong Y, Gonzaga R, Grytsan V, Gryz E, Gu V, Haldorsen E, Halupa A, Haw R, Hrvojic A, Hurrell L, Isserlin R, Jack F, Juma F, Khan A, Kon T, Konopinsky S, Le V, Lee E, Ling S, Magidin M, Moniakis J, Montojo J, Moore S, Muskat B, Ng I, Paraiso J P, Parker B, Pintilie G, Pirone R, Salama J J, Sgro S, Shan T, Shu Y, Siew J, Skinner D, Snyder K, Stasiuk R, Strumpf D, Tuekam B, Tao S, Wang Z, White M, Willis R, Wolting C, Wong S, Wrong A, Xin C, Yao R, Yates B, Zhang S, Zheng K, Pawson T, Ouellette B F, Hogue C W. The biomolecular interaction network database and related tools 2005 update. *Nucleic Acids Research*, 2005, **33**: D418−D424

67 Chatr-Aryamontri A, Breitkreutz B J, Heinicke S, Boucher L, Winter A, Stark C, Nixon J, Ramage L, Kolas N, O′Donnell L, Reguly T, Breitkreutz A, Sellam A, Chen D, Chang C, Rust J, Livstone M, Oughtred R, Dolinski K, Tyers M . The BioGRID interaction database: 2013 update. *Nucleic Acids Research*, 2013, **41**: D816−D823

68 Kerrien S, Aranda B, Breuza L, Bridge A, Broackes-Carter F, Chen C, Duesbury M, Dumousseau M, Feuermann M, Hinz U, Jandrasits C, Jimenez RC, Khadake J, Mahadevan U, Masson P, Pedruzzi I, Pfeiffenberger E, Porras P, Raghunath A, Roechert B, Orchard S, Hermjakob H. The IntAct molecular interaction database in 2012. *Nucleic Acids Research*, 2012, **40**: D841−D846

69 Keshava Prasad T S, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A, Balakrishnan L, Marimuthu A, Banerjee S, Somanathan DS, Sebastian A, Rani S, Ray S, Harrys Kishore C J, Kanth S, Ahmed M, Kashyap M K, Mohmood R, Ramachandra Y L, Krishna V, Rahiman B A, Mohan S, Ranganathan P, Ramabadran S, Chaerkady R, Pandey A. Human protein reference database-2009 update. *Nucleic Acids Research*, 2009, **37**: D767−D772

70 von Mering C, Huynen M, Jaeggi D, Schmidt S, Bork P, Snel B. STRING: a database of predicted functional associations between proteins. *Nucleic Acids Research*, 2003, **31**(1): 258−261

71 Brown K R, Jurisica I. Online predicted human interaction database. *Bioinformatics*, 2005, **21**(9): 2076−2082

72 Licata L, Briganti L, Peluso D, Perfetto L, Iannuccelli M, Galeota E, Sacco F, Palma A, Nardozza AP, Santonico E, Castagnoli L, Cesareni G. MINT, the molecular interaction database: 2012 update. *Nucleic Acids Research*, 2012, **40**: D857−D861

73 Han K, Park B, Kim H, Hong J S, Park J. HPID: the human protein interaction database. *Bioinformatics*, 2004, **20**(15): 2466−2470

74 Sardiu M E, Florens L, Washburn M P. Evaluation of clustering algorithms for protein complex and protein interaction network assembly. *Journal of Proteome Research*, 2009, **8**(6): 2944−2952

75 Chua H N, Ning K, Sung W K, Leong H W, Wong L. Using indirect protein-protein interactions for protein complex predication. *Computer Society Bioinformatics Conference-CSB*, 2007, **6**: 97−109

76 Friedel C C, Krumsiek J, Zimmer R. Bootstrapping the interactome: unsupervised identification of protein complexes in yeast. In: Proceedings of the 12th Annual International Conference on Research in Computational Molecular Biology. Berlin, Heidelberg: Springer-Verlag, 2008. 3−16

77 Brohée S, van Helden J. Evaluation of clustering algorithms for protein-protein interaction networks. *BMC Bioinformatics*, 2006, **7**: 488

78 Mewes H W, Amid C, Arnold R, Frishman D, Güldener U, Mannhaupt G, Münsterkotter M, Pagel P, Strack N, Stümpflen V, Warfsmann J, Ruepp A. MIPS: analysis and annotation of proteins from whole genomes. *Nucleic Acids Research*, 2004, **32**(Database issue): D41−D44

79 Aloy P, Böttcher B, Ceulemans H, Leutwin C, Mellwig C, Fischer S, Gavin A C, Bork P, Superti-Furga G, Serrano L, Russell R B. Structure-based assembly of protein complexes in yeast. *Science*, 2004, **303**(5666): 2026−2029

80 Dwight S S, Harris M A, Dolinski K, Ball C A, Binkley G, Christie K R, Fisk D G, Issel-Tarver L, Schroeder M, Sherlock G, Sethuraman A, Weng S, Botstein D, Cherry J M. Saccharomyces Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). *Nucleic Acids Research*, 2002, **30**(1): 69−72

81 Dennis G Jr, Sherman B T, Hosack D A, Yang J, Gao W, Lane H C, Lempicki R A. DAVID: Database for annotation, visualization, and integrated discovery. *Genome Biology*, 2003, **4**(5): P3

82 Carbon S, Ireland A, Mungall C J, Shu S Q, Marshall B, Lewis S. AmiGO: online access to ontology and annotation data. *Bioinformatics*, 2009, **25**(2): 288−289

83 Zheng Q, Wang X J. GOEAST: a web-based software toolkit for Gene Ontology enrichment analysis. *Nucleic Acids Research*, 2008, **36**(Suppl 2): w358−w363

84 Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 2000, **28**(1): 27−30

85 Boyle E I, Weng S, Gollub J, Jin H, Botstein D, Cherry J M, Sherlock G. GO: TermFinder-open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics*, 2004, **20**(18): 3710−3715

86 Killcoyne S, Carter G W, Smith J, Boyle J. Cytoscape: a community-based framework for network modeling. *Methods in Molecular Biology*, 2009, **563**: 219−239

87 Breitkreutz B J, Stark C, Tyers M. Osprey: a network visualization system. *Genome Biology*, 2003, **4**(3): R22

88 Guo Xing-Li, Gao Lin, Chen Xin. Models and algorithms for alignment of biological networks. *Journal of Software*, 2010, **21**(9): 2089−2106
(郭杏莉, 高琳, 陈新. 生物网络比对的模型与算法. 软件学报, 2010, **21**(9): 2089−2106)

89 Hirsh E, Sharan R. Identification of conserved protein complexes based on a model of protein network evolution. *Bioinformatics*, 2006, **23**(2): e170−e176

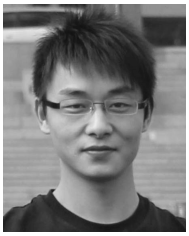
90 Dost B, Shlomi T, Gupta N, Ruppin E, Bafna V, Sharan R. QNet: a tool for querying protein interaction networks. In: Proceedings of the 11th Annual International Conference on Research in Computational Molecular Biology. Berlin, Heidelberg: Springer, 2007. 1−15

91 Dutkowski J, Tiuryn J. Identification of functional modules from conserved ancestral protein-protein interactions. *Bioinformatics*, 2007, **23**(13): i149−i158

92 Liu Zhong-Yang, Li Dong, Zhu Yun-Ping, He Fu-Chu. Progress in the evolutionary analysis of protein interaction networks. *Progress in Biochemistry and Biophysics*, 2009, **36**(1): 13−24
(刘中扬, 李栋, 朱云平, 贺福初. 蛋白质相互作用网络进化分析研究进展. 生物化学与生物物理进展, 2009, **36**(1): 13−24)

冀俊忠 北京工业大学教授. 2004 年获北京工业大学计算机应用技术专业博士学位, 2005 年和 2010 年分别在挪威科技大学、纽约州立大学布法罗分校做访问学者. 主要研究方向为机器学习, 数据挖掘, 群智能算法, 生物信息学. 本文通信作者. E-mail: jjz01@bjut.edu.cn

(**JI Jun-Zhong** Professor at Beijing University of Technology. He received his Ph. D. degree in computer science and application technology from Beijing University of Technology in 2004. He was a visiting scholar at Norwegian University of Science and Technology in 2005 and State University of New York at Buffalo in 2010, respectively. His research interest covers machine learning, data mining, swarm intelligence, and bioinformatics. Corresponding author of this paper.)



刘志军 北京工业大学硕士研究生. 主要研究方向为机器学习, 生物数据挖掘. E-mail: lzhj@emails.bjut.edu.cn

(**LIU Zhi-Jun** Master student at Beijing University of Technology. His research interest covers machine learning and biological data mining.)

刘红欣 北京工业大学硕士研究生. 主要研究方向为计算智能, 生物数据挖掘. E-mail: hong4781798@emails.bjut.edu.cn

(**LIU Hong-Xin** Master student at Beijing University of Technology. Her research interest covers computation intelligence and biological data mining.)

刘椿年 北京工业大学教授, 日本山口大学客座教授. 主要研究方向为数据挖掘, 约束逻辑程序设计, 机器学习和软计算. E-mail: ai@bjut.edu.cn

(**LIU Chun-Nian** Professor at Beijing University of Technology. He is a visiting professor at Yamaguchi University, Japan. His research interest covers data mining, artificial intelligence, inductive logic programming, constraint programming, and soft computing.)