# 蛋白质网络中复合体和功能模块预测算法研究

鱼 亮 高 琳 孙鹏岗
(西安电子科技大学计算机学院 西安 710071)

**摘 要** 预测蛋白质相互作用网络中的复合体和功能模块对于理解生物系统的组织和功能具有重要的意义.到目前为止,已经出现了大量的蛋白质复合体和功能模块预测算法及相关的软件,这些算法各具特色,但同时也具有一定的局限.文中对典型的聚类预测算法进行了研究,依据算法特性对它们进行了分类,并从算法思想、关键技术以及算法性能等方面进行了分析和比较.进一步介绍了基于网络比对策略检测保守模块的算法.最后,结合蛋白质网络数据集对典型的聚类算法从运行效率和预测结果的匹配率等方面进行了比较与分析,为生物网络模块的挖掘和分析提供了有益的参考.

**关键词** 蛋白质相互作用网络;复合体;功能模块;聚类;算法
**中图法分类号** TP301 **DOI 号**: 10.3724/SP.J.1016.2011.01239

# Research on Algorithms for Complexes and Functional Modules Prediction in Protein-Protein Interaction Networks

YU Liang GAO Lin SUN Peng-Gang
(*School of Computer Science and Technology, Xidian University, Xi'an* 710071)

**Abstract** Predicting protein complexes and functional modules in protein-protein interaction (PPI) networks is very important to understand the organization and function of the biological system. So far, a lot of algorithms and related software have been proposed for finding protein modules in PPI networks. However, these algorithms have their own characteristics and limitations. This paper mainly reviewes some typical clustering algorithms. Firstly, the authors make a classification to these algorithms according to their mathematical properties, and analyze them from several aspects, such as the idea of algorithm, key technology, advantages and disadvantages. Secondly, the authors briefly introduce algorithms based on PPI networks comparison for mining conserved patterns. Finally, combining with a PPI dataset, the authors make a comparison and analysis to some clustering algorithms from efficiency and matching rate of the prediction results, which provides a useful reference for mining and analysis modules of biological networks.

**Keywords** protein-protein interaction (PPI) networks; complexes; functional modules; clustering; algorithms

## 1 引 言

识别蛋白质相互作用网络中的模块结构,是理解功能单元的组织结构以及动态性的第一步.随着对蛋白质组学和系统生物学的研究,蛋白质相互作用(Protein-Protein Interaction,PPI)网络的规模变得越来越大,越来越完善.在系统层面上理解生物的

收稿日期:2009-02-03;最终修改稿收到日期:2010-04-28.本课题得到国家自然科学基金重点项目(60933009)、高等学校博士学科点专项科研基金(200807010013)、国家自然科学基金(61072103)、中央高校基本科研业务费专项资金(K50510030006)资助.**鱼 亮**,女,1979 年生,博士研究生,讲师,主要研究方向为数据挖掘、复杂网络模块分析、聚类和计算生物学. E-mail: lyu@xidian.edu.cn. **高 琳**(通信作者),女,1964 年生,教授,博士生导师,主要研究领域为计算生物信息学、生物数据挖掘、图论与组合优化算法及其应用. E-mail: lgao@mail.xidian.edu.cn. **孙鹏岗**,男,1982 年生,博士研究生,主要研究方向为数据挖掘、复杂网络社团结构分析和计算生物学.

组织结构,是后基因时代一个关键的目标.复杂的细胞处理过程是模块化的,由模块之间协同完成[1-4],与同一基本生物功能相关的模块由基因组或者蛋白质组组成.揭示生物网络中的模块结构将有助于理解功能单元如何工作[1].为了处理规模和复杂性不断增加的蛋白质相互作用数据[5],一些对蛋白质相互作用网络结构进行分析的算法应运而生.

对蛋白质相互作用网络的研究,目前集中在检测蛋白质复合体和功能模块上[6-12],该问题已经转化为在蛋白质相互作用图中识别高度连通(或稠密)子图问题.已有文献[6-7,10]说明,模块就是它们各自内部连接相对紧密,而其相互之间连接却相对松散,这些模块一般对应的是有意义的生物单元,如蛋白质复合体和功能模块.Bader 和 Hogue[7] 提出了包含 3 个步骤的MCODE(Molecular Complex Detection)算法,用它来识别蛋白质相互作用网络中的复合体.King 等人提出了 RNSC(Restricted Neighborhood Search Clustering)算法[13],该算法基于代价函数将蛋白质相互作用网络中的蛋白质划分成不同的复合体.类似的结果还有 Pržulj 等人提出的预测网络中的稠密子图为蛋白质复合体[10].2007 年 Li 等人提出了一个新的构思,被称为 rank-HSP (Heavies Subgraph Problem)[12],它利用两个动态系统来发现蛋白质复合体和功能模块.

聚类方法已经被证明对于分析 PPI 网络来说是一个非常好的策略,许多不同类型的聚类算法已经用于寻找 PPI 网络中的蛋白质复合体和功能模块[14].MCL(Markov Clustering)算法[15-16] 是一种快速的可扩展无监督聚类算法,它的基本思想是基于图论中的随机流来提取蛋白质相互作用网络中的复合体.将聚类分析方法应用于蛋白质相互作用网络,通常还要涉及到把蛋白质网络转化成有权网络.Pereira-Leal 等人在 2004 年提出了一种近似方法[17],通过大量实验给蛋白质之间的相互作用赋权重.Pereira-Leal[17] 和 Arnau[14] 等人将两个蛋白质之间最短路径的长度,作为它们之间相互作用的权重.还有一些可选的方法,它们将网络划分成子网络,再基于子网络的拓扑结构来识别模块.除此之外,很多其它的网络聚类方法也被应用于分析蛋白质相互作用网络,其中包括基于边介数的聚类方法[18-23].Palla 等人于 2005 年提出了一种新的基于紧密连接子网过滤的聚类算法[24],将其应用于蛋白质网络中[25-27],可发现重叠的复合体.近年来,又有学者结合信息论的思想给出了模块发现的算法[28-29].除聚类算法外,还可以通过基于网络比对的方法,比较不同物种的蛋白质相互作用网络来发现保守复合体[30-31].

在本文中,我们主要分析了近几年用于预测 PPI 网络中复合体和功能模块的一些有代表性的聚类算法;依据算法特性将典型聚类算法分为以下几类:基于层次聚类的算法、基于图划分的聚类算法、基于密度的局部搜索聚类算法以及其它聚类算法;主要分析介绍了这些聚类算法的基本思想、所采用的关键技术以及算法的性能;进一步介绍了基于蛋白质网络比对策略的保守模块检测算法;最后,结合蛋白质数据集对典型的聚类算法从运行效率、预测结果的匹配率等方面进行了比较和分析,为生物网络模块的挖掘和分析提供有益的参考.

## 2 蛋白质复合体和功能模块

蛋白质复合体(Protein Complex)[32] 是指在同一时间和空间通过相互作用组成一个多分子机制的一组蛋白质,例如转录因子复合物和 RNA 拼接等.蛋白质功能模块(Protein Functional Module)[32] 是指在不同的时间和空间(细胞周期的不同条件或阶段,在不同细胞区室等)的蛋白质,通过相互绑定来参与某一特定的分子进程.例如负责细胞周期进展的细胞周期蛋白,酵母信息素响应路径等.目前,关于两者之间关系的讨论还非常有限,从一般意义上说,功能模块包含蛋白质复合物.近几年来,多篇论文通过实验方法识别蛋白质复合体[33-37].首先,通过化学实验测定,然后再对实验测定的结果应用生物信息学方法进行统计分析.通过化学实验可以较准确地测定某一环境下比较稳定的蛋白质复合体,但对于那些不稳定复合体,蛋白质之间的相互作用是瞬时的、动态变化的,以实验为基础的研究方法很难捕捉到这些蛋白质复合体,而且实验成本十分昂贵.

目前,普遍的做法是将蛋白质网络表示成一个图,其中节点表示蛋白质,边表示蛋白质之间的相互作用.这样,就可以利用各种图聚类算法来挖掘蛋白质复合体和功能模块.从蛋白质相互作用网络中挖掘复合物和功能模块,不仅有利于分析蛋白质网络的拓扑结构,进而探索蛋白质通过相互作用完成生命活动的奥秘,而且对预测未知蛋白质功能及蛋白质相互作用也具有极其重要的作用.

## 3 基于层次的聚类算法

层次聚类(hierarchical clustering)算法是一类

传统聚类算法，已被广泛应用于复杂网络（包括社会网络和生物网络等）的研究. 核心思想是基于各个节点之间连接的相似性，把网络自然地划分为各个子网络. 根据从网络中移除边还是向网络中添加边，可将算法分为分裂（divisive algorithm）和凝聚（agglomerative algorithm）两类[38].

分裂算法的基本思想是从网络着手，试图找到已连接的相似性最低的节点对，然后移除它们之间的边. 重复这个过程，就逐步把网络划分成越来越小的子网络. 这个过程可以终止于任何边的移除，而此时网络的组成就认为是若干个模块. 可以利用树状图或者系统生成树来表示分裂算法的流程，如图 1，这样能够更好地描述整个网络逐步分解成若干个越来越小的子网络这一连续的过程. 图 1 中最底层的圆表示网络中的一个个节点. 如果我们把虚线从上部水平向下移动，就会得到越来越小的多个模块，移到底部时，网络中的每个节点就构成一个独立的模块. 利用虚线在树状图中的任何位置水平断开，就会得到此状态下的模块结构.

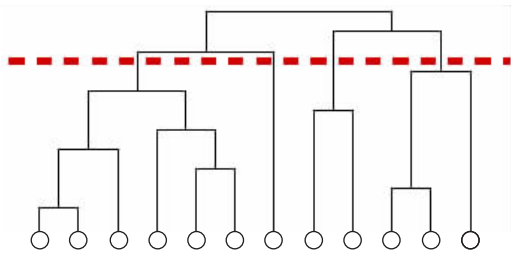

图 1　层次聚类算法的结果用树状图或者系统生成树来说明

相反地，在凝聚算法中，对于节点数为 $n$ 的网络，首先将其看成是每个节点为一个模块的具有 $n$ 个子网络的模块结构，利用某种方法计算出各节点对之间的相似性，然后从相似性最高的节点对开始，向网络中逐渐地添加边，有边相连的节点就构成了一个新的更大的模块. 这一过程可以中止于任何一点，而最终这个网络的组成就认为是若干个模块. 与分裂算法类似，从最开始的由 $n$ 个独立节点和 0 条边构成的空网络到最终网络结构的整个流程，也可以利用树状图或者系统生成树来表示. 如图 1，如果水平虚线从树的底端逐步上移，各节点也逐渐聚合成为更大的模块. 当虚线移至顶端时，即表示整个网络就总体成为一个模块. 同样，在该树状图的任何一个位置用虚线水平断开，就对应着一种模块结构.

下面主要介绍凝聚算法中具有代表性的 Newman 快速算法[19]、分裂算法中具有代表性的 Girvan-Newman（GN）算法[18] 和 Highly Connected Subgraphs（HCS）算法[11].

### 3.1　GN 算法

GN 算法是 Girvan 和 Newman 在 2002 年提出的一种新的在网络中寻找模块结构的方法[18]. 它的基本思想是，通过不断的从网络中删除介数（betweenness）最大的边，将各个模块彼此之间分离开来，从而得到大型网络的潜在模块结构. 这里提到的边的介数（betweenness）是对 Freeman[39-40] 提出的点介数的一种扩展，它定义为经过该边的任意两点之间的最短路径的条数. 它是将大的网络划分成各个模块结构的一种新的度量标准. 该算法已经被应用于蛋白质相互作用网络，并且通过实验说明能够有效地识别蛋白质复合体和功能模块[41].

GN 算法的基本步骤是：

1. 计算网络中所有边的介数；
2. 找到介数最大的边并将它从网络中删除；
3. 重新计算受删除边影响的边的介数；
4. 重复步 2，直到所有边都被删除.

与以往的算法相比，GN 算法不必依赖多余的信息来判断得到的模块结构是否具有实际意义，而是可以直接从拓扑结构进行分析；在模块结构已知的情况下，基于最短路径的算法简单、直观，而且相对于以前的方法能给出很好的结果.

但是，从算法的流程可以看到原始的 GN 算法并没有明确给出算法的结束条件. 因此，在事先不知道存在多少个蛋白质复合体和功能模块的情况下，将 GN 算法应用于蛋白质相互作用网络发现蛋白质模块时，GN 算法不知分解进行到哪一步停止. 这样，还必须加入算法的结束标准，比如由 Newman 等人提出的衡量网络划分质量的一个标准-模块性（Modularity）[42]，利用计算所得模块的 $Q$ 值来确定较好的划分位置，即算法何时结束. 但是，$Q$ 是一个全局优化函数，对其进行优化是一个 NP-完全问题[43]. 目前，如何对 $Q$ 进行优化已成为一个关键的热点问题[44-48].

对于处理小规模的网络来说，GN 算法可以快速、准确地寻找到网络中的模块结构. 但是，从时间复杂度方面来讲，GN 算法耗时比较大，对于稀疏网络时间复杂度为 $O(n^3)$，其中 $n$ 是网络节点数. 如果将 GN 直接应用于处理较大规模的生物网络，如 PPI 网络，时间复杂度是不能忍受的. 于是在 2007 年，Yang 和 Lonardi[41] 提出了一种 GN 算法的并行执行方法，线性加速几乎达到了 32 个处理器. 他们将该方法应用于分析大规模的 PPI 网络，分析结果

表明并行执行的 GN 算法能够高效地分析 PPI 网络,并识别出高可靠性的蛋白质复合体和功能模块.除此之外,我们还可以将 GN 算法与其它聚类算法相结合来处理 PPI 网络[49].

### 3.2 Newman 快速算法

Newman 在 2003 年提出了基于模块性(modularity)[42]的一种新算法:快速算法[19].快速算法总的时间复杂度是 $O((m+n)n)$,对于稀疏网络为 $O(n^2)$,如 PPI 网络,其中 $n$ 和 $m$ 分别是网络中的节点数和边数.它比 GN 算法的时间复杂度大大减少,因此能够高效地分析大规模的蛋白质相互作用网络.

快速算法实际上是基于贪婪算法思想的一种凝聚算法,算法基本步骤如下:

1. 将网络初始化为 $n$ 个模块,$n$ 是网络中的节点数,即每个节点作为一个独立模块.最初始的 $e_{ij}$ 和 $a_j$ 满足以下条件:

$$e_{ij}=\begin{cases}1/2m, & \text{如果节点 } i \text{ 和节点 } j \text{ 之间有边相连}\\ 0, & \text{其它}\end{cases},$$

$$a_i=k_i/2m,$$

其中 $k_i$ 是节点 $i$ 的度,$m$ 为网络中边的总数.在后面的步骤中,如果模块中的节点数大于 1,则 $e_{ij}$ 是模块 $i$ 与模块 $j$ 之间的边在 $2m$ 条边中所占的比例,$k_i$ 是模块 $i$ 的度,即模块 $i$ 与其它模块之间的边的条数,$a_i$ 是 $k_i$ 在 $2m$ 条边中所占比例.这里讨论的网络是无向的,因此分母是 $2m$.

2. 依次合并有边相连的模块对,并计算合并后的 $Q$ 值增量 $\Delta Q$:

$$\Delta Q=e_{ij}+e_{ji}-2a_ia_j=2(e_{ij}-a_ia_j).$$

根据贪婪算法的思想,步 2 的时间复杂度是 $O(m+n)$.

3. 重复步 2,不断合并模块,直到整个网络被合并成一个大的模块.最多要执行 $(n-1)$ 次合并.

利用快速算法分析蛋白质相互作用网络时,当整个算法结束后,可以得到一个蛋白质模块的分层树结构.通过选择在不同的层次断开网络,可以得到不同的蛋白质模块划分结果.在这些蛋白质模块划分中,根据 Newman 提出的模块性衡量标准,选择一个对应的 $Q$ 值是局部最大的,就得到较好的蛋白质模块划分结果.

相对于 GN 算法,快速算法在时间复杂度方面有了很大的改善,而且可以应用于分析具有几百万个节点的网络,因此应用 Newman 快速算法发现具有大数据量的蛋白质相互作用网络中的复合体和功能模块也非常可行.但是,通过实验分析发现,在检测蛋白质复合体和功能模块的准确性方面,Newman 快速算法不如 GN 算法.

### 3.3 HCS 算法

HCS(Highly Connected Subgraphs)算法[11]是一种基于图的连通性的层次聚类算法.它根据相似数据得到一个"相似图",在应用于蛋白质相互作用网络时,节点代表蛋白质,边代表蛋白质之间的正确的相互作用概率大于某一阈值.在这个相似图中,蛋白质模块结构就是高度连通的子图,即子图中边的连通性大于 1/2 倍的子图节点数.HCS 算法的目标是将蛋白质相互作用图中的所有稠密子图挖掘出来.它的基本思想可以简单描述如下:首先对当前图求最小割,判定该图是否满足稠密图的定义,若不满足,则将最小割从当前图中移除产生两个诱导子图,采用递归方法对两个诱导子图再求最小割;如此反复,直到诱导子图是最小割或者成为孤立点.由于生物系统的模块结构都是分层的组织结构,而且完成特定功能的模块结构是相对稠密的子图,HCS 算法对稠密图的定义所蕴含的意义又十分贴切现实数据,因此利用 HCS 算法发现蛋白质相互作用网络中的模块结构能取得很好的结果.

经严格的数学证明[11] HCS 算法具有以下几个特性:

特性 1. 每一个高度连通子图的直径至多是 2.

特性 2. 如果 $S$ 是一个最小割,它把图划分成两个导出子图,$\overline{H}$ 是较小的子图,它包含 $k$ 个节点 $(k>1)$,那么 $|S|\leq k$,当且仅当 $\overline{H}$ 是一个紧密连接子网(完全连通图)时等号成立.

特性 3. 假定图 $G$ 不是高度连通的,但是它的直径是 2.从图 $G$ 中删除最小割 $S$ 得到两个导出子图 $H$ 和 $\overline{H}$,其中 $|V(\overline{H})|\leq|V(H)|$,$|V(H)|$ 表示子图 $H$ 中的节点数.那么,① $\overline{H}$ 中的每一个节点都和 $S$ 相关;② $\overline{H}$ 是一个紧密连接子网;③ 如果 $|V(\overline{H})|>1$,那么 $\overline{H}$ 中的每一个节点都与 $S$ 中的一条边相关.

特性 4. ① 高度连接子图中边的数量与其节点数成二次性关系;② 在 HCS 算法中,每一次迭代删除的边数与其节点数至多成线性关系.

HCS 算法的时间复杂度是多项式的,经过一些启发式的改进[11],该算法可以在合理的时间内处理具有几千个节点的网络,如 PPI 网络.同时它给出了高度连通子图的定义:子图的连通性大于子图中节点数的 1/2 倍,这也提供了一个结束模块检测的标准.将 HCS 应用于 PPI 网络时,可以通过这个结束标准来确定何时算法结束,得到所研究 PPI 网络中的全部模块.除了已检测到的蛋白质模块具有前面提到的 4 个可证明的好特性外,HCS 算法使得在进行 PPI 网络模块划分之前,不需要预先知道或者

猜测最终检测到的模块数量. 更重要的一点是，特性 1 指出 HCS 算法检测到的模块直径至多是 2，这是同源性的一个主要特点. 因为如果两个节点同源，那么它们要么邻接，要么具有一个或者多个相同的邻居节点. 这样通过 HCS 检测到的蛋白质复合体和功能模块与真实的蛋白质复合体和功能模块之间重合率就越高，也就越能反映真实的生物特性，为实际的生物实验提供有效的帮助.

显然，基于层次聚类的方法能够以树状结构呈现整个蛋白质网络的层次化模块组成方式，但是却很难识别交叠的蛋白质复合物和功能模块，而且对噪声很敏感.

### 3.4　MOHCS 算法

通过分析著名的频繁子图挖掘算法、稠密子图挖掘算法和频繁稠密子图挖掘算法的特点，并深入研究稠密图的性质，我们基于证明所得的稠密图的一些性质，给出稠密子图的新定义[50]，并据此新定义提出了一种在线性时间内发现网络中重叠的稠密子图算法——MOHCS(Mining Overlapping Highly Connected Subgraphs)算法[50]. 通过将 MOHCS 算法用于分析酵母蛋白质相互作用网络数据，发现 MOHCS 算法挖掘出来的稠密子图的规模整体上不断变小，并且在初期的下降速度非常快. 实验结果表明[50]，无论是在性能方面还是在挖掘结果方面，MOHCS 算法都能令人满意.

## 4　基于图划分的聚类算法

在基于图划分的聚类算法中，将网络中的主体抽象成图中的节点，而将主体之间的关系抽象成图中节点之间的连边. 基于图划分的聚类方法有时也称为基于目标函数的聚类算法. 它的指导思想是使得被划分到同一模块的对象之间相似度最大，而不同模块之间的相似度最小. 其中，比较有代表性的算法就是 RNSC 算法.

### 4.1　RNSC 算法

RNSC(Restricted Neighborhood Search Clustering)算法[13]是一种基于代价函数的聚类算法. 它的核心思想是给每一个可能的模块定义一个相应的代价值，这个值代表了模块聚类的好坏，然后在所有可能的模块中寻找代价值较小的模块.

RNSC 算法定义了两个独立的代价函数，一是整数值代价函数，一是实数值代价函数. 这两个函数在算法的执行过程中用在不同的阶段. 实数值代价函数表达更准确，但比较复杂，运行效率较低，因而只用作算法开始搜索前的预处理操作. 整数值代价函数比较简单，贯穿整个算法过程. RNSC 算法为了得到代价值较小的模块，初始时先随机地对节点进行模块划分，接下来再随机地将一个节点从一个模块移动到另一个模块来取得较小的模块代价值，一次移动就是以一个接近最佳的值来降低模块代价值. 由此不难看出，RNSC 是一种局部搜索算法.

RNSC 算法中加入了变换步骤，来改善局部搜索容易陷入局部最小值的缺陷. 另外，为了避免重复划分，RNSC 算法还包含了一系列禁止操作[13].

利用 RNSC 算法对 PPI 网络进行分析，为了得到更接近真实生物复合体的模块，这个过程共分为两步：首先，利用 RNSC 算法对蛋白质网络进行模块化；接着，利用过滤条件对所得的模块进行过滤. RNSC 算法定义了 3 个过滤条件：模块大小、模块密度和功能同源性.

(1) 模块大小

RNSC 算法依据以下两个原则，将小规模模块丢弃. ① 在同样重叠比例的情况下，最有可能的是已检测到的大复合体与已知复合体之间进行重叠. ② 小的已知复合体在现有的 PPI 网络中，通常密度都会很小，因此利用模块算法很难检测到它们. 对于模块大小的下界，会根据实际的 PPI 网络来定义，所有小于这个下界的模块都将被丢弃，只保留大于或等于下界的模块.

(2) 模块密度

蛋白质复合体有一个典型特性就是复合体内部的蛋白质之间交互频率很高. 因此，可以认为具有低密度的模块很难反映已知蛋白质复合体，而且通过丢弃密度小于某个域值的模块，可以提高该算法的预测率.

(3) 功能同源性

在同一已知复合体中的蛋白质通常都具有高的功能同源性，即在同一个已知复合体中的蛋白质，属于同一个功能块的可能性很大. 功能的同源性用 $P$ 值来表示，它由超几何分布给出，表达式如下：

$$P=1-\sum_{i=0}^{k-1}\frac{\binom{|F|}{i}\binom{|V|-|F|}{|C|-i}}{\binom{|V|}{|C|}},$$

其中，$|C|$ 表示检测到的模块 $C$ 中的蛋白质个数，$|F|$ 表示已知复合体或功能模块 $F$ 中的蛋白质个数，$k$ 表示在 $C$ 中具有 $k$ 个蛋白质包含在 $F$ 中，$|V|$

表示整个蛋白质相互作用网络中的蛋白质总数.这个式子表示在模块 $C$ 中至少有 $k$ 个蛋白质在已知复合体或功能模块 $F$ 中的概率,用于衡量检测到的蛋白质模块对于某个已知蛋白质功能的富集程度.对于检测到的模块 $C$,计算其与所有已知复合体或功能模块 $F$ 对应的 $P$ 值,将 $P$ 值最小的作为模块 $C$ 的 $P$ 值.

根据实验定义一个 $P$ 值的门限值,将 $P$ 值大于门限值的模块丢弃掉.一般门限值的范围在 $10^{-2}\sim 10^{-8}$ 之间,具体的值要根据实际的网络数据来确定.

利用带有过滤条件的 RNSC 算法可以预测到高可信的蛋白质复合体.这些预测结果可以为真正的生物实验提供有价值的参考,使生物实验更有针对性,不仅提高了效率还节省了大量的开支.RNSC 的分析结果不仅能够保证预测未知的蛋白质复合体,而且在一些情况下能够对已有的结果进行验证.今后,还可以考虑从以下两方面对现有的算法进行改进:一是在计算功能同源性方面,另外在数据方面.

总之,基于图划分的聚类算法简单,易于理解,但是聚类所得模块之间没有重叠,即蛋白质只能属于一个模块.

## 5 基于密度的局部搜索聚类算法

基于密度的算法与其它算法的一个根本区别是:它不是基于各种各样的距离的,而是基于密度的.这样就能克服基于距离的算法只能发现"类圆形"的聚类的缺点.这类算法的指导思想就是,只要一个区域中的点的密度大过某个阈值,就把它加到与之相近的聚类中去.这里我们主要介绍两种比较有代表性的算法:紧密连接子网过滤算法(Clique Percolation Method,CPM)和 MCODE 算法.

### 5.1 紧密连接子网过滤算法

紧密连接子网过滤算法(Clique Percolation Method,CPM)[24]是由 Palla 等人在 2005 年提出的一种新的模块划分算法,利用它可以分析相互重叠的模块结构,应用软件 CFinder(http://www.cfinder.org/)就是以该算法的思想为基础开发的[51].

利用 CPM 算法寻找网络中的 $k$-紧密连接子网模块,共分为两大步骤[24,38]:

1. 寻找网络中的紧密连接子网.

2. 利用紧密连接子网寻找 $k$-紧密连接子网模块.

算法的复杂度由 Palla 等人根据实际网络的计算分析得到,$t=\alpha n^{\beta\ln(n)}$,其中 $\alpha,\beta$ 是常数,$n$ 为网络中节点的数目.

Zhang 等人[26]将 CPM 算法应用到酵母的 PPI 网络中检测蛋白质复合体,经其设定的过滤条件过滤后获得 125 个大小范围在 4～46 个蛋白质复合体,并将所得结果与 MIPS(Munich Information Center for Protein Sequences: ftp://ftpmips.gsf.de/yeast/)[52]中的功能目录(FunCat)[53]数据库进行比对,匹配率达到 88%.其中,有些蛋白质复合体对应多种生物功能.如文献[26]中提到的两个例子:模块 74 参与 mRNA 处理功能和 RNA 绑定功能;模块 98 同时参与 mRNA 合成、DNA 构象变异、乙酰化和脱乙酰化作用的变异和染色体结构的组织等功能.在分析 PPI 网络方面,CPM 算法有以下两大优势.首先,CPM 是确定性方法,而其它大部分网络聚类算法都是随机性的,如 SPC(Super-Paramagnetic Clustering)[54] 算法、RNSC[13] 算法以及 MCL[17]算法.随机性聚类算法运行结果受给定条件的影响比较大,即使是相同的数据输入也可能产生不同的模块划分结果.最重要的一点是通过 CPM 方法可以找到重叠的蛋白质复合体,而像参考文献[17]中提到的,很少有其它的方法可以实现这一点.但是,CPM 算法本身存在一个缺点,它对 3-紧密连接子网基本特性的要求太过严格.这就可能导致 CPM 不能检测出在 PPI 网络中常见的一些模块[55].可以从这点进行考虑,对 CPM 进行改进,得到一个条件要求稍微宽松的新方法.

通过对 CPM 算法的分析,可以看到 CPM 算法为寻找 PPI 中感兴趣的功能区域,提供了一种快速的途径.根据近来对 PPI 网络的分析[4,6]可以看出,对 CPM 检测结果的分析[26]强有力地支持了 PPI 网络的分子结构.尽管 CPM 算法还存在很多局限性,但是它在利用计算方法系统分析 PPI 网络方面是一个有益的补充.

### 5.2 MCODE 算法

MCODE(Molecular Complex Detection)算法[7]是针对检测蛋白质网络中的蛋白质复合体而提出的一种聚类算法.它主要包括以下 3 个步骤:节点赋权重、模块预测以及可选的后期处理操作.

MCODE 算法基于聚类的聚集系数[56],采用节点赋权重体系.以下是 MCODE 算法包含的 3 个主要步骤:

1. 节点赋权重.利用节点邻居区域内,最大 $k$-core 的局部网络密度来给节点赋权重.

2. 模块预测. 将有权图作为输入，首先以权重最大的节点作为中心节点，由它迭代地向外扩展，将那些权重大于给定阈值的节点依次包含进来.

3. 后期处理. 该步骤为可选操作. 首先，将不是 2-core 的模块过滤掉. 接着，可能执行两个可选项：fluff 选项和 haircut 选项. fluff 选项是根据给定范围在 0.0～1.0 之间的 fluff 参数，可以增大模块的规模. haircut 选项，完成的是移除模块内所有度为 1 的节点，那么得到的模块都是 2-core 的. 如果这两个选项都执行，那么首先执行 fluff 操作，再执行 haircut 操作.

整个算法的时间复杂度是多项式 $O(nmh^3)$，其中 $n$ 是顶点的个数，$m$ 是边的个数，$h$ 是所输入的图中节点邻居区域中所包含节点的平均个数，该时间复杂度主要由节点赋权重步骤决定.

MCODE 算法中，节点赋权重只需要做一次，而且它构成了大部分的时间复杂度，一旦节点权重已得到，可以在 $O(n)$ 的时间复杂度上测试该算法的许多参数，这在估算不同参数时非常有用. 还有就是 MCODE 算法是基于局部密度的，在实现上也相对比较容易. 但是，MCODE 算法存在一个缺陷，它不能保证检测到的子图都是稠密子图. 这也就是说，将 MCODE 算法应用于蛋白质相互作用网络中发现蛋白质复合体时，不能保证检测到所有模块都有意义，这也就降低了算法的有效性. 但是就 MCODE 多项式的算法复杂度，却使它在处理大规模的蛋白质网络时具有很大的优势，可以考虑将它与其它聚类算法巧妙结合，如 GN 算法[49]，使它们在共同处理蛋白质网络数据时，互为补充，充分发挥各自的优势，得到更可信、更有意义的蛋白质复合体.

MCODE 算法的实现软件可以在网站 http://baderlab.org/Software/MCODE[7] 获得.

总之，基于密度的局部搜索聚类算法具有可以识别交叠的蛋白质复合体和功能模块的优点.

### 5.3　ICPM 算法

目前，由于高通量筛选等方法的发展，从而获得了大量的蛋白质相互作用数据，鉴于蛋白质作用数据的飞速增长，给算法的时间性能提出了新的挑战. 因此，目前迫切需要高效的算法来处理大量的蛋白质相互作用数据. CPM 算法是一种经典的紧密连接子网过滤算法，目前已经普遍应用到各种网络中，包括蛋白质相互作用网络. 但是 CPM 在查找 $k$-紧密连接子网时，没有考虑到 $k$-紧密连接子网对节点度的要求，因为组成 $k$-紧密连接子网的节点的度至少为 $k-1$，因此，我们提出了[57-58] 迭代式的紧密连接子网过滤算法 ICPM（Iterative-Clique Percolation Method）. 该算法不仅考虑到 $k$-紧密连接子网对节点度的要求，同时在查找 $k$-紧密连接子网时，先把 $k$-紧密连接子网转化为查找 $k-1$ 紧密连接子网，再把 $k-1$ 紧密连接子网转化为查找 $k-2$ 紧密连接子网直到递归到最小紧密连接子网（这里假设为 3-紧密连接子网），然后通过给 3-紧密连接子网增加一个节点来获得 4-紧密连接子网，再通过给 4-紧密连接子网增加一个节点来获得 5-紧密连接子网直到获得 $k$-紧密连接子网. 由于 ICPM 在查找过程中舍弃了一些不能构成 $k$-紧密连接子网的节点，即节点的度小于 $k-1$ 的节点，从而节省了算法运行的时间.

目前蛋白质作用网络存在大量的噪声问题，因此已有很多方法对蛋白质间的作用大小进行度量，从而形成了一个带权重的网络，权值在 0 与 1 之间. 那么，目前对算法的挑战就成为可以处理带权重的网络数据. CPM 算法采用设定一个阈值，并把低于这个阈值的交互边都舍去，从而把有权图转化为无权图来处理，但是这样做会舍弃一些真实的交互信息，因为某些真实的交互边或许具有较小的权值，从而破坏了真实的网络结构. 基于此，文献[57]ICPM 考虑到子图强度，它定义为模块内边的权值的几何平均，从而使模块内可包含低于设定阈值的交互边. 由此可见，CPM 算法处理带权重的网络所得到的结果是 ICPM 算法处理带权重网络所得结果的一个子集. 正是由于 ICPM 考虑到子图强度，才会识别出比 CPM 更多的模块.

## 6　其它聚类算法

除了前面提到的几类算法外，还有很多其它的算法，如 MCL 算法、基于主分量分析的一致性聚类算法、混合算法 CMG 等等. 目前，又不断地有新的聚类算法出现[59-61].

### 6.1　MCL 算法

MCL 算法（Markov Cluster Algorithm）是一种快速、简单、易扩展的聚类算法，它可以应用于有权有向图，其核心思想是模拟图中的随机游走过程. 所依据的理论是“在图 $G$ 中的随机游走，如果它访问图 $G$ 的一个稠密子图，除非子图中的大部分节点都被访问过，否则不会走出该稠密子图”. MCL 算法定义了两种简单的矩阵操作（expansion 和 inflation）来模拟这一过程. expansion 操作是矩阵乘法，用来模拟随机游走的扩展，使矩阵具有齐次性. 若对矩阵

做一次 expansion 运算，相当于在图中又随机游走了($e-1$)步，其中 $e$ 为 expansion 参数的取值；inflation 操作是对于矩阵中第 $i$ 行，第 $j$ 列的元素 $m_{ij}$，用下面这个式子进行更新：$\frac{m_{ij}^r}{\sum_j m_{ij}^r}$，其中 $r$ 是 inflation参数，$\sum_j m_{ij}^r$ 表示 $m_{ij}$ 所在列所有元素的 $r$ 次幂之和. inflation 操作模拟随机游走的收缩，增大稠密子图内部随机游走的概率，而减小稠密子图之间随机游走的概率. MCL 算法就是通过这两个操作的交叉迭代进行，把一个图划分成若干稠密子图[15-16].

在将 MCL 算法应用于 PPI 网络的聚类时[16]，inflation 参数的选择非常关键. inflation 参数表示的是聚类的粒度，它越大，则稠密子图的数目越多，子图直径就越小. 通过多次实验比对得出，在一般情况下，inflation 参数 $r$ 取值为 1.8 时，聚类效果最好. 然而这也不是绝对的，对于不同的 PPI 网络，inflation参数的取值会有一定的差异[62]，比如在文中第 8 节分析与比较部分，针对我们选用的 PPI 数据，inflation 参数选取 2 时结果比较好.

MCL 算法能够快速、准确地识别蛋白质相互作用网络中的蛋白质复合体，这些复合体之间可以有重叠. 而且与其它已有的一些模块划分算法相比，如 RNSC、SPC 以及 MCODE 算法，MCL 算法对于网络的变化，具有非常好的鲁棒性，抗噪声能力强[62]. 因此，在从 PPI 网络中提取蛋白质复合体方面，具有很大的优势. 但是，它是一种随机性聚类算法，聚类结果受给定条件的影响较大.

MCL 算法的实现软件可从网站 http://micans.org/mcl/[63] 获得.

### 6.2 基于主分量分析的一致性聚类算法

目前所研究的蛋白质相互作用数据中存在噪音信息，同时蛋白质相互作用网络拓扑结构比较复杂，致使传统的聚类算法在提取蛋白质复合体方面，不能得到好的聚类结果. 因此，Asur 等人[64] 在 2007 年提出了一种基于主分量分析(Principal Component Analysis，PCA)的一致性聚类算法，来解决蛋白质网络聚类中存在的问题.

首先提出两种基于 PPI 网络拓扑属性的相似性标准：基于聚集系数的标准和基于介数的标准，目的是为 PPI 网络中的边赋权重，用于反映边所对应的相互作用的可靠性. 相应地，权重较小的边就隐含说明它们很可能是噪音信息. 聚类算法可以利用这些权重来去除噪音边，得到有意义的划分.

接着，将这两个相似性标准应用于 3 个传统的图聚类算法[64]：重复的二分法、直接 $k$ 路划分、多级 $k$ 路划分，得到 6 种基础聚类结果.

最后，利用一致性技术将这 6 种不同的聚类结果进行合并，得到一个有意义的一致性聚类.

基于 PCA 的一致性技术由 3 个步骤组成：

1. 模块提纯(Cluster Purification).
2. 减少维数(Dimensionality Reduction).
3. 一致性聚类(Consensus Clustering).

除此之外，还可以利用基于聚类可靠性的权重构建一个新图，同时通过软聚类(soft clustering)得到某些蛋白质存在于多个复合体中[64].

一致性聚类的质量由 3 种有效性标准来估算：基于拓扑结构的模块性(Modularity)[42]，基于信息论的归一化互信息[65] 和基于特定域的聚类评分[66]. 通过这 3 种标准对聚类结果的评估说明[64]，基于 PCA 的算法除了在可扩展性方面的优势外，还能够得到高有效性的一致性聚类.

### 6.3 CMG 算法

通过分析 GN 算法与 MCODE 算法的特性，我们提出了一种新的结合 MCODE 算法和 GN 算法的 CMG(Combing MCODE with GN)算法[49] 来识别蛋白质网络中的复合体. 首先，利用 MCODE 算法对大规模的蛋白质相互作用网络进行粗划分，得到较大、较稀疏的蛋白质模块. 接着，利用 GN 算法对 MCODE 算法划分所得的结果进行进一步划分，得到更稠密、更有意义的蛋白质模块. 在算法中，还定义了模块密度、大小以及用来衡量模块意义的 $P$ 值等参数作为过滤条件，用以保证实验结果更有意义.

将聚类算法 CMG 应用到酵母的蛋白质相互作用网络中，可以获得与 MIPS 中已知蛋白质复合体进行比对具有极高匹配率的蛋白质模块. 仿真结果表明，我们的聚类算法提供了一种高效、可靠、易扩展的识别蛋白质模块算法. 它可以被用于预测未知蛋白质的功能和确定关键蛋白质，同时还有助于理解细胞中分子模块的功能相关性.

整个 CMG 算法的时间复杂度是多项式 $O(nmh^3)$，其中 $n$ 是顶点的个数，$m$ 是边的个数，$h$ 是所输入的图中节点邻居区域中所包含节点的平均个数. 这个时间复杂度主要由 MCODE 步骤决定. 在 GN 步骤中，GN 算法作用于 MCODE 算法的聚类结果中，因此时间复杂度是 $O(cs^3)$，其中 $c$ 是 MCODE 步骤得到的模块数，$s$ 是 $c$ 个模块中最大模

块的节点数.在实际应用中，$s$ 远远小于整个网络的节点数 $n$.因此，CMG 算法的时间复杂度为 $O(nmh^3)+O(cs^3)=O(nmh^3)$.

## 7　基于蛋白质相互作用网络比对算法

除了通过聚类的方法挖掘蛋白质相互作用网络中的复合体和功能模块外，还有另外一类通过网络比对的方法挖掘网络中被称为保守模式(conserved pattern)的蛋白质模块.所谓保守模式可以理解为不同物种在进化过程中，从共同祖先那里继承下来保留在自己物种中的模块结构.发掘蛋白质相互作用网络中的保守蛋白质复合体，已成为网络比对方法的一个非常重要的应用.2005 年，Sharan 等人[30-31]提出了一个 log ratio likelihood 概率模型，借助这个概率模型进行生物网络比对的相似性打分，通过搜索算法查找相似度得分较高的子图，作为比对的局部最优解，并且将该方法应用在发掘保守蛋白质复合体中.这里[30-31]找到的蛋白质复合体与通过聚类方法找到的蛋白质复合体有所不同.聚类方法是在单个物种的蛋白质相互作用网络中发现复合体，因此发现的复合体不一定具有保守性.而网络比对方法是通过在多个物种的蛋白质相互作用网络中发现它们共有的复合体，因此发掘的蛋白质复合体具有保守性.

## 8　分析与比较

### 8.1　数据集

本文选用数据库 DIP(Database of Interaction Protein：http：//dip.doe-mbi.ucla.edu/)2007 年发布的酵母蛋白质相互作用数据，作为分析比较几种经典聚类算法特性的数据源.该数据集包含 4932 个蛋白质，17491 个相互作用.为了评估各种聚类算法在检测蛋白质复合体和功能模块方面的有效性，我们利用具有蛋白质功能注解的 MIPS(Munich Information Center for Protein Sequences：ftp：//ftpmips.gsf.de/yeast/)中 CYPD(Comprehensive Yeast Genome Database)数据库中的数据集[52].本文功能注解表选用 funcat-2.1_data_20070316，用于分析和比较的蛋白质复合体列表选用 MIPS 中的 complexcat_data_18052006.

### 8.2　实验结果

前面主要介绍了算法的基本思想，所采用的关键技术，并分析了算法在具体应用中的优点和存在的缺点.在这一部分，将着重讨论几种算法在具体分析给定蛋白质相互作用网络中所表现出来的特性，希望能为大家在今后的应用中提供有价值的参考.我们主要针对以下 6 种算法进行比较分析：Newman 快速算法、RNSC 算法、CPM 算法、MCODE 算法、MCL 算法以及 CMG 算法.

不同的聚类算法采用不同的模块衡量标准，因此得到的结果也会大不相同.针对同一蛋白质相互作用数据，表 1 从模块数量、模块规模范围、匹配率、运行时间等 4 个方面，给出了 6 种算法操作结果的比较.计算模块数量时的过滤条件是 3，即只保留模块规模大于 3 个蛋白质的模块.模块规模范围是指保留下来的模块中蛋白质数量的变化范围.此时选用的 $P$ 值阈值为 0.005，即模块的 $P$ 值小于 0.005，才被认为是有意义的.对于匹配率项，是指将算法的操作结果与 MIPS 数据库中已知蛋白质复合体进行比对时，$P$ 值最小且蛋白质匹配值大于等于 50% 的有意义模块数在总的模块数中所占的比例.也就是说，这里只统计有意义而且模块与已知蛋白质复合体的匹配值大于等于 50% 的结果.因为 $P$ 值越小且匹配值越高，说明检测到的蛋白质复合体越接近真实的复合体，也就越有意义.最后还比较了各个算法的运行时间效率，运行环境是 2.4 GHz 奔腾 4 处理器.

从表 1 中我们能很清楚地看到，从模块数量来说，MCL 算法分析结果中满足大于 3 个蛋白质的模块最多，有 396 个，而 Newman 快速算法最少，只有 28 个，而且 Newman 快速算法的匹配率最低只有 10.3%.虽然 Newman 快速算法的运行效率最高只有 15 s，但是在结果有效性方面它却是这 6 种算法中最差的.这是由于 Newman 快速算法目标函数 $Q$ 的局限性[45]导致的.Newman 算法将本不属于同一模块的蛋白质聚到了一起.如由 Newman 算法检测到的 5-蛋白质模块，“YIL008W YHR111W YKL149C YGR072W YFL020C”，与 MIPS 数据库中已知模块进行功能比对时，只有蛋白质“YKL149C”和“YGR072W”具有功能“01.03.16.01 RNA degradation”，其它的 3 个蛋白质并不具备这样的功能注解，因此尽管该 5-蛋白质模块满足了 $P$ 值的条件(小于 0.005)，但其匹配率却只有 40%，即小于 50%，因此也被认为是无意义的模块而被舍弃.MCL 算法虽然模块数量最多，运行效率也较高(71 s)，但是它的匹配率却只有 43%.CPM 算法匹

配率较高(84.3%),但是由于紧密连接子网比较严格的特性,模块数量只有 51 个,漏掉了许多有意义的模块. MCODE 算法由于不能保证找到的模块都是稠密子图,匹配率只有 52.6%. RNSC 算法运行时间较长(74 s),但却检测到了 159 个模块,匹配率达到了 79.2%. CMG 算法检测到了 188 个规模在 4～16 个蛋白质之间的模块,匹配率最高,达到了 90.4%,说明它能检测到更多有意义的模块. CMG 算法之所以可以达到这么高的匹配率,是因为它可以将 MCODE 算法检测到的稀疏模块做进一步划分,提高检测结果的匹配率. 如表 2 所示,模块 1(模块 $ID=1$)是由 MCODE 算法检测到的具有 19 个蛋白质的模块,其中仅有 6 个蛋白质共同具有功能注解"34.01.01.03 homeostasis of protons",因此匹配率仅为 31.6%(粗体蛋白质表示匹配上的蛋白质). CMG 算法将模块 1 做了进一步分解,得到 4 个更为稠密的子模块,即模块 2、模块 3、模块 4 以及模块 5. 功能相近的蛋白质被进一步细分到一个模块中,得到了 4 个更有意义的蛋白质模块.

**表 1 6 种算法分析酵母蛋白质相互作用网络的结果比较**

| 算法名称 | 模块数量/个 | 模块规模范围(蛋白质) | 匹配率≥50%的模块数百分比/% | 运行时间/s |
|---|---|---|---|---|
| Newman快速算法 | 28 | 4～1117 | 10.3 | 15 |
| RNSC 算法 | 159 | 4～38 | 79.2 | 74 |
| CPM 算法 | 51 | 4～39 | 84.3 | 19 |
| MCODE 算法 | 38 | 4～156 | 52.6 | 46 |
| MCL 算法 | 396 | 4～97 | 43.0 | 71 |
| CMG 算法 | 188 | 4～16 | 90.4 | 51 |

注:模块数量指的是模块规模大于 3 的模块总数;模块规模范围是指模块中蛋白质数量的变化范围;匹配率≥50%的模块数百分比是指模块与 MIPS 数据库中一个或者多个功能相匹配的匹配率不小于 50%的模块数量在总的模块数中所占的比例;运行时间是指算法分析给定的数据源所花费的时间,运行环境为 2.4GHz 奔腾 4 处理器

**表 2 CMG 算法操作结果实例**

| 模块 ID | 蛋白质模块 | 匹配率/% | 功能注解 |
|---|---|---|---|
| 1 | **YDL185W YGR020C YEL051W YBR127C YJR033C YDR202C** YMR106C YPR175W YDR121W YNL262W YHR012W YJL053W YJL154C YOR069W YOR132W YOL145C YJR138W YBR009C YPR110C | 31.6 | 34.01.01.03 homeostasis of protons |
| 2 | **YDL185W YGR020C YEL051W YBR127C YJR033C YDR202C** | 100.0 | 34.01.01.03 homeostasis of protons |
| 3 | YMR106C **YPR175W YDR121W YNL262W** | 75.0 | 10.01.03.05 extension/ polymerization activity |
| 4 | **YHR012W YJL053W YJL154C YOR069W YOR132W** | 100.0 | 20.09.07 vesicular transport (Golgi network, etc.) |
| 5 | **YOL145C** YJR138W **YBR009C YPR110C** | 75.0 | 16.03.01 DNA binding |

因此,根据我们的分析比较结果可以推断,在实际检测蛋白质相互作用网络中复合体和功能模块的应用中,比较行之有效的算法是 CMG 算法、CPM 算法以及 RNSC 算法,效果较差的是 Newman 快速算法.

## 9 结束语

本文对蛋白质复合体和功能模块预测算法进行了综述和评价,主要对典型的聚类模块预测算法以及新提出的聚类算法进行了研究和分析. 除了本文分析的聚类算法外,近两年还出现了很多新的预测算法,如 Rosvall 等人[28]提出了一种新的基于信息论的模块预测算法,Martin 等人[29]也提出了一种处理有权有向网络的基于信息论的模块预测算法. 由于篇幅有限,本文不可能涵盖所有预测蛋白质模块结构的算法. 但是通过对一些有代表性的聚类算法的分析总结,仍然可以发现模块预测研究目前还存在很多问题,比如目前还没有模块的统一定义,不同的算法都采用自己的定义,这样得到的模块也是各式各样的,不利于各个算法之间性能的对比. 同时由于所研究的蛋白质数据不完整,如蛋白质的相互作用没有时间和空间的信息,这样不利于对预测所得的蛋白质模块进行复合体和功能模块的区分. 而且现有的蛋白质数据存在很大的噪音,影响了预测算法检测结果的正确性,因而如何消除噪音数据也是一个需要解决的问题. 现有的大部分算法还只是针对无权无向图,并未涉及对有权有向网络的处理,但是真实的网络却往往需要通过有权有向图模型才能准确地刻画它们的特性.

随着蛋白质相互作用网络数据量的不断增大,我们需要一个快速、精确以及鲁棒性好的方法来完

成生物模块的识别.因此,模块检测仍然是一个热门话题,它吸引了各个领域的人投入到这方面从事研究工作.模糊理论具有处理模糊性和不确定性的优势,因此若将其引入聚类分析中,将来很可能成为蛋白质相互作用网络中检测模块结构的另一途径.希望通过本文的工作,为 PPI 网络中蛋白质模块的预测及其它的生物网络数据,如基因调控网络、代谢网络模式的挖掘和分析提供有价值的参考.

**致　谢**　在此,我们向对本文的工作给予支持和建议的同行表示感谢!

## 参　考　文　献


[1] Hartwell L H, Hopfield J J et al. From molecular to modular cell biology. Nature, 1999, 402: C47-C52

[2] Barabasi A L, Oltvai Z N. Network biology: Understanding the cell's functional organization. Nature Reviews Genetics, 2004, 5(2): 101-114

[3] Ravasz E, Somera A L et al. Hierarchical organization of modularity in metabolic networks. Science, 2002, 297 (5586): 1551-1555

[4] Rives A W, Galitski T. Modular organization of cellular networks. Proceedings of the National Academy of Sciences, 2003, 100(3): 1128-1133

[5] Xenarios I, Salwinski L et al. DIP, the database of interacting proteins: A research tool for studying cellular networks of protein interactions. Nucleic Acids Research, 2002, 30 (1): 303-305

[6] Spirin V, Mirny L A. Protein complexes and functional modules in molecular networks. Proceedings of the National Academy of Sciences, 2003, 100(21): 12123-12128

[7] Bader G D, Hogue C W. An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics, 2003, 4(2): 20-29

[8] Xiong H, He C et al. Identification of functional modules in protein complexes via hyperclique pattern discovery//Proceedings of the Pacific Symposium on Biocomputing. Big Island, Hawaii, USA, 2005: 221-232

[9] Chen J C, Yuan B. Detecting functional modules in the yeast protein-protein interaction network. Bioinformatics, 2006, 22(18): 2283-2290

[10] Pržulj N, Wigle D et al. Functional topology in a network of protein interactions. Bioinformatics, 2004, 20(3): 340-348

[11] Hartuv E, Shamir R. A clustering algorithm based on graph connectivity. Information Processing Letters, 2000, 76(4-6): 175-181

[12] Li W, Liu Y et al. Dynamical systems for discovering protein complexes and functional modules from biological networks. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2007, 4(2): 233-250

[13] King A D, Pržulj N et al. Protein complex prediction via cost-based clustering. Bioinformatics, 2004, 20(17): 3013-3020

[14] Arnau V, Mars S et al. Iterative cluster analysis of protein interaction data. Bioinformatics, 2005, 21(3): 364-378

[15] van Dongen S. Graph clustering by flow simulation [Ph. D. dissertation]. Centers for Mathematics and Computer Science (CWI), University of Utrecht, Amsterdam, 2000

[16] Enright A J, Dongen S V et al. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Research, 2002, 30(7): 1575-1584

[17] Pereira-Leal J B, Enright A J et al. Detection of functional modules from protein interaction networks. Proteins: Structure, Function, and Bioinformatics, 2004, 54(1): 49-57

[18] Girvan M, Newman M E J. Community structure in social and biological networks. Proceedings of the National Academy of Sciences, 2002, 99(12): 7821-7826

[19] Newman M E J. Fast algorithm for detecting community structure in networks. Physical Review E, 2004, 69(6): 066133

[20] Radicchi F, Castellano C et al. Defining and identifying communities in networks. Proceedings of the National Academy of Sciences, 2004, 101(9): 2658-2663

[21] Luo F, Yang Y et al. Modular organization of protein interaction networks. Bioinformatics, 2007, 23(2): 207-214

[22] Dunn R, Dudbridge F et al. The use of edge-betweenness clustering to investigate biological function in protein interaction networks. BMC Bioinformatics, 2005, 6(1): 39

[23] Yoon J, Blumer A et al. An algorithm for modularity analysis of directed and weighted biological networks based on edge-betweenness centrality. Bioinformatics, 2006, 22(24): 3106-3108

[24] Palla G, Derényi I et al. Uncovering the overlapping community structure of complex networks in nature and society. Nature, 2005, 435(7043): 814-818

[25] Palla G, Derényi I et al. Uncovering the overlapping modular structure of protein interaction networks. FEBS Journal, 2004, 272(Suppl. 1): 434

[26] Zhang S, Ning X et al. Identification of functional modules in a PPI network by clique percolation clustering. Computational Biology and Chemistry, 2006, 30(6): 445-451

[27] Zhang C, Liu S et al. Fast and accurate method for identifying high-quality protein-interaction modules by clique merging and its application to yeast. Journal of Proteome Research, 2006, 5(4): 801-807

[28] Rosvall M, Bergstrom C T. An information-theoretic framework for resolving community structure in complex networks. Proceedings of the National Academy of Sciences, 2007, 104(18): 7327-7331

[29] Martin R, Bergstrom C T. Maps of random walks on complex networks reveal community structure. Proceedings of the National Academy of Sciences, 2008, 105: 1118-1123

[30] Sharan R, Ideker T et al. Identification of protein complexes by comparative analysis of yeast and bacterial protein interaction data. Journal of Computational Biology, 2005, 12(6): 835-846

[31] Hirsh E, Sharan R. Identification of conserved protein complexes based on a model of protein network evolution. Bioinformatics, 2007, 23(2): e170-176

[32] Spirin V, Mirny L A. Protein complexes and functional modules in molecular networks. Proceedings of the National Academy of Sciences, 2003, 100: 12123-12128

[33] Gavin A C, Krause R et al. Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature, 2002, 415(6868): 141-147

[34] Gavin A C, Aloy E et al. Proteome survey reveals modularity of the yeast cell machinery. Nature, 2006, 440(7084): 631-636

[35] Ho Y, Gruhler A et al. Systematic identification of protein complexes in saccharomyces cerevisiae by mass spectrometry. Nature, 2002, 415(6868): 180-183

[36] Kumar A, Snyder M. Protein complexes take the bait. Nature, 2002, 415: 123-124

[37] Krogan N J, Cagney Q et al. Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature, 2006, 440(7084): 637-643

[38] Xie Zhou, Wang Xiao-Fan. An overview of algorithms for analyzing community structure in complex networks. Complex Systems and Complexity Science, 2005, 2(3): 1-12(in Chinese)
(解邹, 汪小帆. 复杂网络中的社团结构分析算法研究综述. 复杂系统与复杂性科学, 2005, 2(3): 1-12)

[39] Wasserman S, Faust K. Social Network Analysis: Methods and Applications. Cambridge: Cambridge University Press, 1994

[40] Freeman L. A set of measures of centrality based upon betweeness. Sociometry, 1977, 40(1): 35-41

[41] Yang Q, Lonardi S. A parallel edge-betweenness clustering tool for protein-protein interaction networks. International Journal of Data Mining and Bioinformatics, 2007, 1(3): 241-247

[42] Newman M E J, Girvan M. Finding and evaluating community structure in networks. Physical Review E, 2004, 69(2): 026113

[43] Brandes U, Delling D et al. On modularity clustering. IEEE Transactions on Knowledge and Data Engineering, 2008, 20(2): 172-188

[44] Muff S, Rao F et al. Local modularity measure for network clusterizations. Physical Review E, 2005, 72: 056107

[45] Fortunato S, Barthelemy M. Resolution limit in community detection. Proceedings of the National Academy of Sciences, 2007, 104(1): 36-41

[46] Fortunato S. Quality functions in community detection//Proceedings of the SPIE 6601. Florence, Italy, 2007: 660108

[47] Ruan J, Zhang W. Identifying network communities with a high resolution. Physical Review E, 2008, 77(1): 016104

[48] Li Z, Zhang S et al. Quantitative function for community detection. Physical Review E, 2008, 77(3): 036109

[49] Yu L, Gao L et al. A hybrid clustering algorithm for identifying modularity in protein-protein interaction networks. International Journal of Data Mining and Bioinformatics, 2010, 5(4): 600-615

[50] Lin X, Gao L et al. MOHCS: Towards mining overlapping highly connected subgraphs. Computing Research Repository, 2008, abs/0806.3215

[51] Adamcsek B, Palla G et al. CFinder: Locating cliques and overlapping modules in biological networks. Bioinformatics, 2006, 22(8): 1021-1023

[52] Guldener U, Munsterkotter M et al. CYGD: The comprehensive yeast genome database. Nucleic Acids Research, 2005, 33(Suppl. 1): D364-368

[53] Ruepp A, Zollner A et al. The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. Nucleic Acids Research, 2004, 32(18): 5539-5545

[54] Blatt M, Wiseman S et al. Superparamagnetic clustering of data. Physical Review Letter, 1996, 76(18): 3251-3254

[55] Bader G D, Hogue C W. Analyzing yeast protein-protein interaction data obtained from different sources. Nature Biotechnology, 2002, 20(10): 991-997

[56] Watts D J, Strogatz S H. Collective dynamics of 'small-world' networks. Nature, 1998, 393(6): 440-442

[57] Sun P, Gao L. Fast algorithms for detecting overlapping functional modules in PPI networks//Proceedings of the IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology CIBCB 2009. Nashville, USA, 2009: 247-254

[58] Gao L, Sun P et al. Clustering Algorithms for detecting functional modules in protein interaction networks. Journal of Bioinformatics and Computational Biology, 2009, 7(1): 217-242

[59] Qi Y, Balem F et al. Protein complex identification by supervised graph local clustering. Bioinformatics, 2008, 24(13): i250-268

[60] Chua H, Kang N et al. Using indirect protein-protein interactions for protein complex predication. Journal of Bioinformatics and Computational Biology, 2008, 6(3): 435-466

[61] Friedel C C, Krumsiek J et al. Bootstrapping the interactome: Unsupervised identification of protein complexes in yeast. Journal of Computational Biology, 2009, 16(8): 971-987

[62] Brohee S, van Helden J. Evaluation of clustering algorithms for protein-protein interaction networks. BMC Bioinformatics, 2006, 7(488): 1-19

[63] van Dongen S. Graph clustering via a discrete uncoupling process. Siam Journal on Matrix Analysis and Applications, 2008, 30(1): 121-141

[64] Asur S, Ucar D et al. An ensemble framework for clustering protein-protein interaction networks. Bioinformatics, 2007, 23(13): i29-40

[65] Strehl A, Gosh J. Cluster ensembles — A knowledge reuse framework for combining multiple partitions. Journal on Machine Learning Research, 2002, 3(3): 583-617

[66] Ashburner M, Ball C A et al. Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. Nature Genetics, 2000, 25(1): 25-29

**YU Liang**, born in 1979, Ph. D. candidate, lecturer. Her research interests include data mining, modularity analysis in complex network, clustering and computational biology.

**GAO Lin**, born in 1964, Ph. D., professor, Ph. D. supervisor. Her research interests include bioinformatics, data mining in biological data, graph theory and intelligence computation.

**SUN Peng-Gang**, born in 1982, Ph. D.. His research interests include data mining algorithms, modularity analysis in complex network and bioinformatics.

**Background**

This work is supported by the National Key Natural Science Foundation of China (No. 60933009), Specialized Research Fund for the Doctoral Program of Higher Education (No. 200807010013), the Fundamental Research Funds for the Central Universities (No. K50510030006), and National Natural Science Foundation of China (No. 61072103).

As a crucial level of biology hierarchy, functional modules encompass groups of genes or proteins involved in common elementary biological functions. Identifying these functional modules in biological networks is important to understand the organization and interaction of the cellular processes they represent. Also they are useful in system level understanding of biological organization which is a key objective of the post-genomic era.

The identification of functional modules in protein interaction networks can be successfully accomplished through the use of cluster analysis. Cluster analysis is invaluable in elucidating network topological structure and the relationships among network components. Clustering seeks to identify groups of proteins that are more likely to interact with each other than with proteins outside the group. There are many different types of clustering approaches available for modularization of protein interaction networks. Partition-based approaches have been applied to biological networks. One partition-based clustering approach, the Restricted Neighborhood Search Clustering (RNSC) algorithm, determines the best partition using a cost function. In addition, other approaches have been applied to biological networks. For example, the Markov Clustering Algorithm (MCL) finds clusters using iterative rounds of expansion and inflation that, respectively, prefer the strongly connected regions and weaken the sparsely connected regions.

In this paper, we mainly reviewed some typical clustering algorithms. Firstly, we made a classification to these algorithms according to their mathematical properties. We analyzed them from several aspects, such as the idea of algorithm, key technology, advantages and disadvantages. Secondly, we briefly introduced algorithms based on protein interaction networks comparison for mining conserved patterns. Finally, combining with a protein interaction dataset, we made a comparison and analysis to some clustering algorithms from efficiency and matching rate of the prediction results, which provides a useful reference for mining and analysis modules of biological networks.