

文章编号: 1673 1689(2011)01-0095-06

基于图聚类的蛋白质相互作用网络功能模块探测

梅娟¹, 何胜², 李炜疆²

(1. 无锡城市职业技术学院 电子信息工程系, 江苏 无锡 214153; 2. 江南大学 生物工程学院, 江苏 无锡 214122)

摘 要: 作者采用了基于模块性的图聚类算法来探测蛋白质相互作用网络中的集团, 在具有 2 617 个节点 11 855 个相互作用的酵母蛋白质相互作用网络中探测出 177 个集团, 同时采用慕尼黑信息中心(Munich Information Center, MIPS)的层次功能注释对其进行了注释, 并且验证得到的集团的确是内部连接紧密的子图。

关键词: 蛋白质相互作用网络; 模块性; 聚类; 功能模块

中图分类号: Q 811.4

文献标识码: A

Detection of Communities in the Yeast Protein-Protein Interaction Network Based on Graph Clustering

MEI Juan¹, HE Sheng², LI Wei jiang²

(1. Department of Electronic and Information Technology, Wuxi City College of Vocational Technology, Wuxi 214153, China; 2. School of Biotechnology, Jiangnan University, Wuxi 214122, China)

Abstract: Interaction detection methods can led to the discovery of thousands of protein-protein interactions, and discerning relevance within large-scale data sets is important for bioinformation. As an important means for knowledge discovery, graph clustering attracts much attention in analysis of protein-protein interaction networks. Here, a modularity-based method was used to find communities of protein-protein interaction networks. Using this method, 177 communities were detected from a network involving 11,855 interactions among 2617 proteins in yeast and annotated according to MIPS hierarchical functional categories. The results validated that these communities are indeed densely connected in sub-graphs.

Key words: protein-protein interaction network, modularity, clustering, functional module

自 1995 年 12 月完成第一个微生物(流感嗜血杆菌)全基因组测序以来^[1], 越来越多原核和真核生物的基因组被成功测序^[2]。后基因组时代的一个焦点就是挖掘测序完成的酵母基因组的基因产物的功能地位。蛋白质几乎不独立地在生化层次

起作用, 而是通过与其他蛋白质相互作用形成一个集体去执行某个特定的细胞任务^[3], 因此, 探测蛋白质相互作用的高通量技术吸引了许多研究者的关注。近年来, 以酵母双杂交^[4]、串联亲和纯化^[5]、蛋白质芯片^[6]、质谱分析^[7]和噬菌体显示^[8]等为代

收稿日期: 2010-01-10

作者简介: 梅娟(1980-), 女, 江苏盐城人, 工学博士, 讲师, 主要从事人工智能和生物信息学研究。Email: meijuanwx@gmail.com

表的高通量蛋白质组技术的发展和运用,使得蛋白质相互作用数据呈指数增长。不同研究组已经成功测定了幽门螺杆菌(*H. pylori*)^[9]、酵母(*S. cerevisiae*)^[10]、线虫(*C. elegans*)^[11]、果蝇(*D. melanogaster*)^[12]等模式生物的部分蛋白质相互作用。大量的相互作用数据已经被收集整理在不同的公共数据库中。例如, DIP(Database of Interacting Proteins)^[13]数据库中收集了4 963个蛋白间的17 570个相互作用(Scere20081014),然而,其中接近一半蛋白质的功能是未知的。因此,需要有效的计算方法来处理这些急速增长的复杂数据。

相互作用数据为系统地分析大的生命系统的结构提供了一个很好的机会,借此能够理解一些基本组成部分,比如蛋白复合物,细胞通路,蛋白相互作用和功能模块。相应的细胞功能和生化过程都是通过功能模块中的蛋白相互作用体现出来的,复杂网络的模块结构对功能分析是至关重要的^[14-15]。鉴别蛋白质相互作用网络中这样的功能模块有助于理解这些基本细胞网络的结构和功能。因此,开发有效的计算方法探测功能模块是一个巨大的但必不可少的挑战。聚类作为探索关系数据组织结构的重要手段能够帮助我们更好地理解蛋白质相互作用网络的拓扑结构以及其组分间的相互关系。并且,每个集团的主要功能可以由它成员(蛋白质)的功能来推断,集团中未被注释的成员的功能可以用已被注释的成员的功能进行注释^[16]。

蛋白质相互作用网络的数学模型是无向图,图的节点表示蛋白质,连接节点的边表示相应蛋白间的相互作用。因此,对蛋白相互作用数据的聚类问题简化为图论问题(或称为图聚类)^[17]。例如, maximal clique 方法通过寻找整个图中的完全连通子图来探测功能模块^[18],然而由于构成有生物学意义的集团的蛋白质之间的相互联系没有那么紧密,因此这个方法只能找出少量集团。为了弱化 maximal clique 方法的严格要求, Quasi Clique^[19], MCODE^[20]等方法通过优化密度函数来探测联系紧密的子图而不是完全连通子图。RNSC^[21]和 HCS^[22]方法通过寻找最小边切集来探测集团。MCL^[23]方法通过迭代使用 expansion 和 inflation 这两个操作分别促进连接紧密的区域和减弱连接稀疏的区域。Samantha 和 Liang^[24]采用了一个统计方法,他们提出了一个基于随机抽样的概率模型,用来评价网络中任意给定的两个蛋白质之间是不是显著地拥有大量相同的相互作用蛋白质。如果两个蛋白质所拥有的共同相互作用蛋白质的数

量远远超出随机出现的可能,那么就认为这两个蛋白质之间在功能上有关联。

作者采用了以“模块性”这个刻画集团结构强弱的指标为目标函数的图聚类算法 CD 来探测酵母蛋白相互作用网络中的集团,并且验证了得到的集团的确是内部连接紧密的子图。

1 数据和方法

1.1 数据

高通量数据中存在着大量的假阳性和假阴性^[25],而且不同研究组得到的相互作用数据之间只有很少的重叠部分^[26]。为了度量它们的精确度和鉴别它们的偏歧, von Mering 等^[25]评估了已经发表的5 400多个酵母蛋白中的80 000多条相互作用,并且给每个相互作用一个置信度的打分。为降低假阳性对数据分析的干扰,剔除了低置信度的相互作用数据,保留了中度和高度的置信度的数据,最终得到的数据集包含了2 617个蛋白的11 855个相互作用。

1.2 蛋白质相互作用网络的构建

蛋白质相互作用网络用无向图 $G(V, E)$ 来表示,其中节点集合 $V = \{P_1, P_2, \dots, P_n\}$,包含了每个蛋白作为节点,边集合 $E = \{(P_i, P_j) \mid \text{蛋白 } P_i \text{ 和 } P_j \text{ 间具有相互作用}\}$ 。对称的 n 维邻接矩阵用 A 表示,且 $A_{ij} = \begin{cases} 1 & (P_i, P_j) \in E \\ 0 & (P_i, P_j) \notin E \end{cases}$ 。

1.3 算法

从直观上看,网络的集团是指内部联系紧密的子网络,集团之间的联系相对比较稀疏。为了对网络的这种集团结构有一个定量描述, Newman 和 Girvan^[27]提出了模块性(modularity)这个评价指标。在这个框架下,较大的模块性值对应着网络一个较强的集团结构,因此,探测网络的集团转化为寻找最大模块性的优化问题。

1.3.1 模块性的定义 对于一个给定了划分,且由 n 个节点和 m 条边组成的网络,它的模块性定义为:

$$Q(c) = \frac{1}{2m} \sum_{i,j=1}^n (A_{ij} - \frac{d_i d_j}{2m}) \delta(c_i, c_j),$$

其中, A 为这个网络的邻接矩阵,当节点 i 和节点 j 之间有边时, A 中的元素 A_{ij} 等于1,否则为0; d_i 为节点 i 的度; δ 为 kronecker delta 函数,当 x 等于 y 时, $\delta(x, y)$ 等于1,否则为0; c_i 为节点 i 所在集团的编号。

1.3.2 基于模块性的 CD 算法 笔者使用的方法

是组前阶段开发的基于模块性优化的 CD 算法^[28], 下面作简单介绍。

基于模块性的优化算法的目标是能够找到一个整型向量 c 使得 Q 值最大化。CD 算法的主要操作是单点移动。为了表示的方便, 可用一个最小变化算子 T_{ka} , 来描述这个操作。该算子作用于集团的编号向量 c 从而产生新的集团结构 c' 。也就是说, 节点 k 移动到集团 α 可以表示为:

$$c' = T_{ka}c,$$

其中,

$$c'_j = \begin{cases} \alpha, & \text{当 } j = k; \\ c_k, & \text{其他所有 } j. \end{cases}$$

则, 节点 k 移动到集团 α 后, 模块性的变化量为:

$$D_{ka}(c) = \frac{(d_k(a) - d_k(c_k))}{m} + \frac{d_k(d(c_k) - d(a) - d_k)}{2m^2} + \frac{A_{kk}}{m}, a = c_k,$$

其中, $d_i(\alpha)$ 表示节点 i 在集团 α 中邻居的个数(也就是有连接的节点个数), $d(\alpha)$ 表示集团 α 中所有节点的度之和。从上式可以看出, 由单点移动引起的模块性变化量的计算量是 $O(1)$ 。由于这个操作的高效性, 所以 CD 算法主要依赖于它。

对于一个给定的网络, 最初将所有节点随机分配到任意多的集团中, 为了改进这个初始划分, 依次不断地将单个节点移动到使 Q 值增长最快的集团中, 直至 Q 值不能再增大为止。在这个过程中, 节点逐渐归并到部分集团中, 出现了不断增加的空集团, 导致非空集团的个数不断减少, 称之为收缩过程; 经历收缩过程后, 所有的节点处于较合适的集团中, 但这个状态只是相对于单节点移动的一个局部最优解。为了寻求更好的集团结构, 使所有节点以扰动概率 p 离开原来所在的集团, 即由 p 来控制对收缩过程输出的集团结构的破坏程度。这个过程中非空集团的个数迅速变多, 因此称之为膨胀过程。整个算法流程就是将收缩和膨胀过程迭代多次求最优值的过程, 其形式化定义如下:

CD algorithm:

(初始化)

1. 随机分配所有节点到 $c_{\max}$ 个集团中;
分别记录当前的集团结构和 Q 值为 c_{best} 和 Q_{best}

2. repeat
(收缩过程)

3. repeat

4. for each node i

5. 移动节点 i 到集团 $\beta = \arg \max_{\alpha} \Delta_{\alpha}$

6. endfor

7. until Q 不能再增大为止

8. 如果 $Q > Q_{\text{best}}$, 更新 c_{best} 和 Q_{best}
(膨胀过程)

9. for each node i

10. 以扰动概率 p 来决定节点 i 是否移动到其他集团中

11. endfor

12. until NITER times

在上面的算法流程中, NITER 表示整个收缩和膨胀过程迭代的次数, $\arg \max_{\alpha} \Delta_{\alpha}$ 表示使 Δ_{α} 取最大值的变量 α 。

1.4 集团的定量定义

从直观上看, 集团是网络的一部分, 它内部的连接相比其与外部的连接更加稠密。为了能更好地评价集团探测算法的性能表现, 需要有更精确的定义来描述集团, 为此, 文献中给出了多种定量的定义。节点 i 的度用 k_i 表示, 且 $k_i = \sum_j A_{ij}$ (邻接矩阵明确且完整地反映了网络的拓扑结构, 因为蛋白质相互作用网络的数学模型是不带权的无向图, 所以邻接矩阵的元素或为 0 或为 1)。设节点 i 属于子图 $G_1 (G_1 \subset G)$, 则节点 i 的度分成两部分: $k_i(G_1) = k_i^{\text{in}}(G_1) + k_i^{\text{out}}(G_1)$ 。其中, $k_i^{\text{in}}(G_1) = \sum_{j \in G_1} A_{ij}$ 表示节点 i 与子图 G_1 中的节点之间的连接数, $k_i^{\text{out}}(G_1) = \sum_{j \notin G_1} A_{ij}$ 表示节点 i 与子图 G_1 外的节点之间的连接数。

当 $\sum_{i \in G_1} k_i^{\text{in}}(G_1) > \sum_{i \in G_1} k_i^{\text{out}}(G_1)$ 时, 子图 G_1 被认为是集团。这时, 子图 G_1 内的度之和大于其与外部的度之和。

1.5 给集团赋予功能和集团的 P 值

对于聚类得到的集团, 其中可能含有多种蛋白功能, 为某集团赋予功能的最简单的方法是用该集团中大部分蛋白所共有的功能来注释它。或者, 我们可以采用超几何分布来计算一个集团对某个特定功能的蛋白的富集程度。对于蛋白数目为 s 的集团, 设其含有某功能的蛋白 t 个, 并设整个蛋白质相互作用网络中共含有 N 个蛋白, 该功能类共含有 C 个蛋白, 则这样的集团随机出现概率的 P 值是:

$$P = 1 - \sum_{i=1}^t \frac{\binom{c}{i-1} \binom{N-c}{s-i+1}}{\binom{N}{s}}$$

当 P 值越小(接近 0) 时, 说明该集团能够随机出现这种功能的概率就越低, 则可能更有生物学意义。这里用集团的最小 P 值对应的蛋白功能给该集

团赋予主要功能。

2 结果和讨论

将 CD 算法应用于 1.1 节构建的蛋白相互作用数据集。算法的参数设置如下: $c_{\max} = 500$, NITER= 500, $p = 0.7$ 。对于得到的聚类结果,按照 1.4 节给出的集团定义进行验证,去掉不满足集团定义的类,最终共得到 177 个集团。这些集团共包含了 2 170 个蛋白, 占所有蛋白的 82.92%, 这表明 CD 算法得到的聚类结果对于酵母蛋白的功能有较广的覆盖面。

首先从拓扑结构的角度来评价这 177 个集团的合理性。为了达到这个目的,提出连接密度这个指标。某一子图 G_1 的连接密度等于这一子图内部所有度之和 $\sum_{i \in G_1} k_i$ 除以组成该子图的所有节点的度之和。显然,子图的连接密度的取值范围是 0 到 1。根据集团的定量定义,集团的连接密度的取值范围是 0.5 到 1。对于聚类结果中某个特定的集团,我们考虑如何选择参照物来说明其合理性,其中一个简单的做法是从整个蛋白相互作用网络中随机选取与该集团含有相同数目的蛋白,计算由这些蛋白组成的子图的连接密度。然而,随机选取的蛋白间不一定有相互作用,因此,这种做法不足以令人信服。这里使用了一种较严格的方法,即从该集团中随机选取小部分的蛋白与集团外相同数目的蛋白交换,并且要求这些交换的蛋白间是有相互作用的,然后再观察交换后该集团连接密度的变化情况。

图 1 描绘了聚类结果中所有集团的连接密度以及交换后的连接密度,虚线表示直线 $y = x$ 。图中的数据点是 20 次独立实验的平均值。从图中

可以看出,当集团中 15% 的节点被交换后,绝大部分集团的连接密度显著减小至 0.5 以下,这表明它们已经不能成为集团。当更多的节点被交换 (30%),集团的连接密度减少更多。从这个实验发现,CD 算法得到的集团的确是内部联系紧密的局部子图,因此能够作为酵母蛋白相互作用网络的候选功能模块。

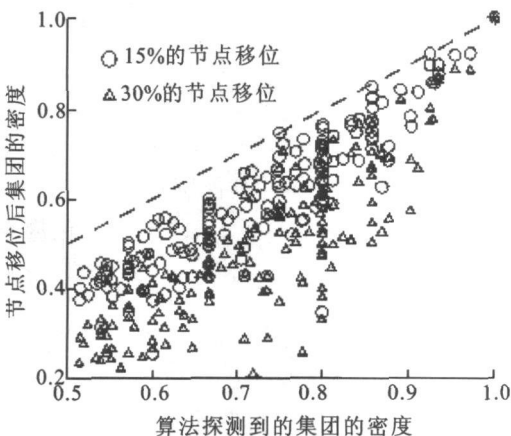


图 1 聚类结果中集团连接密度与部分蛋白变换位置后连接密度的散点图

Fig. 1 A scatter plot of the connectivity densities of the modules and the ones after replacement of proteins

对于聚类得到的 177 个集团,计算了它们对应的 P 值,并采用了慕尼黑信息中心 (Munich Information Center, MIPS) 的层次功能注释对其进行了注释。在计算 P 值的过程中,注意到 MIPS 的注释允许一个蛋白有多于一种的功能。结果显示其中 175 个集团都被赋予一种功能,而其他两个集团被赋予两种功能。表 1 列出了这些集团的详细信息 (由于篇幅有限,只列出部分集团)。

表 1 各个集团的功能注释

Tab. 1 Functional annotation of each protein community

ID	Proteins	MIPS Scheme	MIPS Description	Density
1	Q0045, Q0250, Q0275, YAR008W, YBR022W, YBR024W, YBR037C, YCL004W, YDL036C, YDL235C, YDR044W, YDR076W, YDR127W, YDR196C, YDR268W, YDR338C, YER043C, YER141W, YFL033C, YFL045C, YFR047C, YGL037C, YGL059W, YGL094C, YGL143C, YGL156W, YGL256W, YGR112W, YGR169C, YGR276C, YHL016C, YHR028C, YHR032W, YHR179W, YHR206W, YIL042C, YIL147C, YJL121C, YJR125C, YKR031C, YLR006C, YLR105C, YLR107W, YLR142W, YLR165C, YLR281C, YLR382C, YMR234W, YNL123W, YNL218W, YNL305C, YOL080C, YOL097C, YOL114C, YOR111W, YOR131C, YOR143C, YOR219C, YOR278W, YOR330C, YOR356W, YPL132W, YPL160W, YPL171C, YPL172C, YPR035W	30.05.01.10	two σ component signal transduction system (sensor kinase component)	0.623 1

续表 2

ID	Proteins	MIPS Scheme	MIPS Description	Density
2	Q0085, Q0130, YBL013W, YBL099W, YBR039W, YBR127C, YBR136W, YDL004W, YDL101C, YDL141W, YDL185W, YDL227C, YDR277C, YDR298C, YEL027W, YEL051W, YER070W, YER172C, YGL236C, YGL251C, YGR020C, YGR124W, YGR180C, YGR246C, YGR271W, YHR025W, YHR026W, YHR060W, YIL066C, YJL026W, YJR121W, YLR289W, YLR447C, YML096W, YMR054W, YOR028C, YOR270C, YOR332W, YPL234C, YPR145W	34. 01. 01. 03	homeostasis of protons	0. 797
3	YMR186W, YNL064C, YOR027W, YPL240C	14. 01	protein folding and stabilization	0. 6
4	YAL003W, YGL244W, YGR140W, YGR285C, YHR064C, YJR060W, YKL049C, YKL081W, YKL089W, YMR168C	10. 03. 04. 05	chromosome segregation/ division	0. 571 4
5	YAL009W, YDL116W, YGL092W, YGL100W, YGR057C, YGR161G C, YHR004C, YIL109C, YJR042W, YKL057C, YKL176C, YKR082W, YLR208W, YNL006W, YOR160W	42. 10. 05	nuclear membrane	0. 838 7
...	...	...	...	...
...	...	...	...	...
176	YDR074W, YER062C, YHR043C, YHR044C, YIL053W, YLL026W, YML100W, YMR261C, YPL201C, YPR026W	32. 01	stress response	0. 928 6
177	YDR047W, YDR080W, YER014W, YGL067W, YGR223C, YML115C, YOR089C, YOR176W	01. 20. 19. 01	metabolism of porphyrins	0. 736 8

3 结 语

采用基于模块性的聚类算法 CD 分析构造的酵母蛋白相互作用网络, 对聚类得到的结果按照集团的定量定义进行验证, 最终得到 177 个功能集团, 采用慕尼黑信息中心(Munich Information Center,

MIPS) 的层次功能注释对其进行了注释, 使用超几何分布来计算一个集团对某个特定功能的蛋白的富集程度。研究表明, 得到的集团的确是内部连接紧密的子图, 能够作为酵母蛋白相互作用网络的候选功能模块, 从中挖掘相关的功能信息是下一步研究的主要任务。

参考文献(References):

[1] Fleischmann R D, Adams M D, White O, et al. Whole genome random sequencing and assembly of Haemophilus influenzae Rd[J]. **Science**, 1995. 269(5223): 496– 512.

[2] Letovsky S, Kasif S. Predicting protein function from protein/protein interaction data: a probabilistic approach[J]. **Bioinformatics**, 2003, 19(Suppl 1): 197– 204.

[3] Gavin A C, Bosche M, Krause R, et al. Functional organization of the yeast proteome by systematic analysis of protein complexes[J]. **Nature**, 2002. 415(6868): 141– 147.

[4] Fields S, Song, O. A novel genetic system to detect protein protein interactions[J]. **Nature**, 1989. 340(6230): 245– 246.

[5] Rigaut G, Shevchenko A, Rutz B, et al. A generic protein purification method for protein complex characterization and proteome exploration[J]. **Nat Biotechnol**, 1999, 17(10): 1030– 1032.

- [6] Zhu H, Bilgin M, Bangham R, et al. Global analysis of protein activities using proteome chips[J]. **Science**, 2001. 293 (5537): 2101– 2105.
- [7] Ho Y, Gruhler A, Heilbut A, et al. Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry[J]. **Nature**, 2002. 415(6868): 180– 183.
- [8] Smith G P. Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface[J]. **Science**, 1985. 228(4705): 1315– 1317.
- [9] Rain J C, Selig L, De Reuse H, et al. The protein-protein interaction map of *Helicobacter pylori*[J]. **Nature**, 2001. 409 (6817): 211– 215.
- [10] Uetz P, Giot L, Cagney G, et al. A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae* [J]. **Nature**, 2000. 403(6770): 623– 627.
- [11] Li S, Armstrong C M, Bertin N, et al. A map of the interactome network of the metazoan *C. elegans*[J]. **Science**, 2004. 303(5657): 540– 543.
- [12] Giot L, Bader J S, Brouwer C, et al. A protein interaction map of *Drosophila melanogaster*[J]. **Science**, 2003. 302 (5651): 1727– 1736.
- [13] Xenarios I, Salwinski L, Duan X J, et al. DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions[J]. **Nucleic Acids Res**, 2002. 30(1): 303– 305.
- [14] Guimera R, Nunes Amaral L A. Functional cartography of complex metabolic networks[J]. **Nature**, 2005. 433(7028): 895– 900.
- [15] Pereira Leal J B, Enright A J, Ouzounis C A. Detection of functional modules from protein interaction networks[J]. **Proteins**, 2004. 54(1): 49– 57.
- [16] Bolten E, Schliep A, Schnekenner, S, et al. Clustering protein sequences ——structure prediction by transitive homology [J]. **Bioinformatics**, 2001. 17(10): 935– 941.
- [17] 梅娟, 王正祥, 石贵阳, 李炜疆. 复杂生物网络分析的图聚类方法研究进展[J]. 食品与生物技术学报, 2008, 27(5): 15 – 20.
- MEI Juan, HE Sheng, WANG Zheng xiang, SHI Gui yang, LI Wei jiang. Clustering protein sequences through modularity optimization[J]. **Journal of Food Science and Biotechnology**, 2008, 27 (5): 15– 20. (in Chinese)
- [18] Spirin V, Mirny L A. Protein complexes and functional modules in molecular networks[J]. **Proc Natl Acad Sci USA**, 2003. 100(21): 12123– 12128.
- [19] Bu D, Zhao Y, Cai L, et al. Topological structure analysis of the protein-protein interaction network in budding yeast [J]. **Nucleic Acids Res**, 2003. 31(9): 2443– 2450.
- [20] Bader G D, Hogue C W. An automated method for finding molecular complexes in large protein interaction networks[J]. **BMC Bioinformatics**, 2003. 4: 2.
- [21] King AD, Przulj N, Jurisica I. Protein complex prediction via cost-based clustering[J]. **Bioinformatics**, 2004. 20(17): 3013– 3020.
- [22] Hartuv E, Shamir R. A clustering algorithm based on graph connectivity[J]. **Information Processing Letters**, 2000. 76 (4– 6): 175– 181.
- [23] Enright A J, van Dongen S, Ouzounis C A. An efficient algorithm for large scale detection of protein families[J]. **Nucleic Acids Res**, 2002. 30(7): 1575– 1584.
- [24] Samanta M P, Liang S. Predicting protein functions from redundancies in large scale protein interaction networks[J]. **Proc Natl Acad Sci USA**, 2003. 100(22): 12579– 12583.
- [25] von Mering C, Krause R, Snel B, et al. Comparative assessment of large scale data sets of protein-protein interactions [J]. **Nature**, 2002. 417(6887): 399– 403.
- [26] Deane C M, Salwinski L, Xenarios I, Eisenberg D. Protein interactions: two methods for assessment of the reliability of high throughput observations[J]. **Mol Cell Proteomics**, 2002. 1(5): 349– 356.
- [27] Newman M E, Girvan M. Finding and evaluating community structure in networks[J]. **Phys Rev E**, 2004. 69: 026113.
- [28] Mei J, He S, Shi G Y, Wang Z X, et al. Revealing network communities through modularity maximization by a contraction-dilation method[J]. **New Journal of Physics**, 2009. 11: 043025.