文章编号: 1001－2486(2009)04－0081－06

蛋白质相互作用网络的几种聚类方法综述*

王正华, 董蕴源, 王勇献
(国防科技大学 并行与分布处理国防科技重点实验室, 湖南 长沙 410073)

摘 要: 蛋白质相互作用网络是后基因组时代系统生物学研究的重要内容。针对蛋白质相互作用网络中的聚类问题, 介绍了几种代表性的聚类分析方法, 初步分析了这些方法的特点, 指出了当前研究工作的困难与挑战, 并对今后的研究方向作了展望。

关键词: 蛋白质相互作用网络; 谱聚类; 信息流模拟聚类; 整体聚类
中图分类号: Q811; TP391 文献标识码: A

Review on Several Clustering Methods in Protein-protein Interaction Network

WANG Zheng-hua, DONG Yun-yuan, WANG Yong-xian
(Key Laboratory of Science and Technology for National Defense of Parallel and Distributed Processing, National Univ. of Defense Technology, Changsha 410073, China)

Abstract: Protein-protein interaction network is one of the research hotspot in post-genome era. Several representative methods of clustering in protein-protein interaction network are reviewed. The characteristics of the methods are summarized and the difficulties and challenge are discussed. Finally, the development prospect is proposed.

Key words: protein-protein interaction network; spectral clustering; message passing clustering; ensemble clustering

随着人类基因组计划的完成, 由生物体动态产生并执行遗传程序的蛋白质逐渐进入人们的视线。蛋白质相互作用网络分析作为蛋白质组学的研究内容之一, 近年来越来越多地受到研究者的关注。研究发现, 蛋白质并不是单独行动, 它们在时间和空间上协调一致, 通过相互作用来创造、调控和维持细胞特定的功能[1]。不仅如此, 生物体的新陈代谢也和蛋白质之间的相互作用密不可分。同时, 蛋白质相互作用作为许多疾病的重大病因, 其机理尚未弄清。因此, 对蛋白质相互作用网络的研究可以帮助我们预测功能未知的蛋白质功能, 了解特定生物功能的分子机制。从大量的蛋白质相互作用网络数据中提取出功能模块, 即蛋白质相互作用网络的聚类研究, 是生物体行为理解、蛋白质功能预测和药物设计的基础。

聚类指将数据划分成有意义或有用的组(簇)。划分的基本原则是, 使得组内对象足够相似, 不同组间对象差别较大。组内的相似性越大, 组间差别越大, 说明聚类越好[2]。蛋白质相互作用网络被认为具有模块组织结构, 它由在拓扑结构或者功能上相对独立的子网构成。因此, 对蛋白质相互作用网络进行聚类分析可以揭示蛋白质相互作用网络的结构, 并且对聚类内功能未知蛋白质进行功能预测[3]。

1 蛋白质相互作用网络中的聚类方法

蛋白质相互作用网络中的聚类方法有很多种。传统的聚类方法有层次聚类方法(它又分为凝聚式聚类和分裂式聚类两大类)、基于划分的方法(典型的有 K－均值和 K－中心方法)和基于密度的聚类方法(例如 DBSCAN)。之后又出现了基于距离的聚类方法和基于图的聚类方法。本文重点关注蛋白质相互作用网络模块分析中出现的一些新的聚类方法。

* 收稿日期: 2008－12－23
基金项目: 国家自然科学基金资助项目(60603054, 60773021); 湖南省自然科学基金资助项目(08JJ4021)
作者简介: 王正华(1962－), 男, 教授, 博士生导师。

1.1 谱聚类方法

谱聚类方法是一种基于矩阵特征向量的方法,也是一种能根据顶点之间的权值对图进行划分的方法。它利用矩阵理论和线性代数理论来研究图的邻接矩阵,根据矩阵的谱来确定图的某些性质[4]。

谱聚类方法最早可以追溯到Donath和Hoffman的工作,他们提出利用邻接矩阵的特征向量对图进行划分[5]。同年,Fiedler发现利用图的拉普拉斯矩阵的第二特征向量(次小特征值所对应的特征向量)可以完成图的二分[6]。近些年,谱聚类法逐渐成为最流行的聚类方法之一,并在机器学习、数据挖掘、模式识别、图像分析和生物信息学领域得到广泛的应用。此方法容易实现,并且得到的结果远远好于传统的k-平均聚类算法,主要是因为它能在任意形状的样本空间上聚类,收敛于全局最优解,并且可以处理大规模稀疏数据的聚类问题。

谱聚类方法是揭示复杂联系中结构信息的一个有力工具。Bu和Lu等人利用谱聚类方法对蛋白质网络进行分析,他们借鉴Web页面分析思路,提出了一种可挖掘芽殖酵母蛋白质相互作用网络中近似全连接和近似二分的网络模式方法[7-8]。这样在蛋白质相互作用网络中挖掘出了隐含的拓扑结构,给预测功能未知的蛋白质功能提供了一种新的方法。文献[9]改进了他们的算法,在啤酒酵母、秀丽线虫和大肠杆菌的蛋白质相互作用网络中有效地挖掘出了近似二分模式。Kamp等人使用谱分析法研究果蝇的蛋白质相互作用网络,发现离散的谱密度与蛋白质相互作用网络的拓扑结构有着明显的对应关系,特别是环状结构[10]。针对传统的谱平分法需要预先确定聚类个数的缺点,文献[11]使用改进后的基于Normal矩阵的谱平分法,得到较好的结果。Sen等人在连通矩阵上使用奇异值分解来获得特征向量,他们发现在同一特征向量中的蛋白质并不具有相同的度,但却更倾向于发生相互作用[12]。

经典的谱聚类算法流程如算法1所示[13]。假设给定数据点$x_1, x_2, \cdots, x_n$,它们的相似性定义为$s_{ij}=s(x_i, x_j)$,相似性矩阵$\boldsymbol{S}=(s_{ij})_{i,j=1\cdots n}$。

算法1 非正规化的谱聚类算法

输入:相似性矩阵$\boldsymbol{S}\in\mathbf{R}^{n\times n}$,需要构建的聚类个数$k$

1. 构建一个相似性图,$\boldsymbol{W}$是其权重邻接矩阵;
2. 计算非正规化的拉普拉斯算子L;
3. 计算L的前k个特征向量$v_1, \cdots, v_k$;
4. 构建矩阵$\boldsymbol{V}\in\mathbf{R}^{n\times k}$,使向量$v_1, \cdots, v_k$作为列;
5. For $i=1, \cdots, n$, $y_i\in\mathbf{R}^k$表示矩阵$\boldsymbol{V}$的第i行;
6. 对在$\mathbf{R}^k$中的节点$(y_i)_{i=1,\cdots,n}$用k均值的方法进行聚类,聚类结果$C_1, \cdots C_k$;

输出:聚类$C_1, \cdots, C_k$

尽管谱聚类算法具有坚实的理论基础——谱图理论,并且在实践中也取得了很好的效果,但是它仍然存在一些关键问题需要解决:

(1)谱聚类方法对初始化敏感,需要预先确定聚类的个数。当确定的聚类数大于实际聚类数时,得到的结果很差。因此如何自动地确定聚类数目是一个关键的问题。

(2)如何构造邻接矩阵$\boldsymbol{W}$。在谱聚类方法中,有多种构造邻接矩阵的方法,这些方法可能会导致不同的结果。如何在众多的邻接矩阵中选择一个适合问题的邻接矩阵是一个难题。

(3)计算复杂性问题。使用谱聚类方法不可避免地要计算矩阵的特征值和特征向量。由于求解非稀疏矩阵的所有特征向量的标准解法的时间复杂度为$O(n^3)$,因此通常这种计算的代价很大,特别是当应用于海量数据时,常常会遇到超出计算机内存的情况。

1.2 信息流模拟聚类方法

信息流模拟聚类方法可以被称为信息传递的聚类方法,它其实是"信念传递"方法的应用。信念传递方法是近年来在通讯理论和推理领域新发展起来的一种方法。

我们可以这样理解信息传递方法[14]:每个数据点都希望能够找到与自己最相似的标本点。假设数据点A选择数据点B作为自己的标本点,同时数据点B也需要寻找自己的标本点。每个数据点都会在

寻找自己的标本点的过程中产生约束条件, 这样, 所有的数据点就建立起一个关于约束的网络。当每个约束条件都被满足的时候, 网络的相似性达到最大。接下来各个节点向其相邻节点发送信息, 经过多次的信息传递过程, 每个节点都会找到自己的标本节点。信息流模拟聚类算法的过程[15] 如算法 2 所示。

算法 2 信息流模拟聚类算法

步骤 1. 计算每个节点对之间的信息流出现的概率

步骤 2. 基于每个节点上累计出现的信息流的概率, 为每个节点选择聚类的标本点

步骤 3. 根据已选择的标本点形成初始聚类

步骤 4. 根据各个初始聚类结果之间的相似性, 合并初始的聚类结果

Geng, Ali 等人撰写了一系列文章[16-18], 介绍了消息传递聚类算法(MPC)。MPC 算法同时考虑了数据点之间的局部和全局的距离关系, 允许各个数据点之间进行通信, 利用消息传递来描述并行和自发的聚类过程, 因此可以得到比较精确的聚类结果。MPC 算法的结构灵活, 可以处理多种不同类型的数据; 它同时提供可扩展的结构和版本(MPC-AFS, MPC-STO, MPC-SEMI)来满足各种不同聚类的要求, 例如随机聚类融合[19]、无监督和半监督的消息传递聚类[20] 等。

Hwang, Cho 等人提出了 STM 算法[21], 首次在蛋白质相互作用网络中利用信号传递思想对蛋白质进行聚类并发掘出功能模块。使用修改后的爱尔兰分布模拟网络中的信号传递模型。STM 算法通过蛋白质之间的最短路径传递出现概率, 模拟网络中每个节点对其他节点的信号传递, 这样就反映了该节点的生物和拓扑属性。该方法允许得到的蛋白质聚类有重叠。2008 年, 他们在此基础上, 利用近似全路径算法(QAP)在蛋白质节点之间传递出现概率, 发展了 CASCADE 算法[15]。CASCADE 算法既可以发现密度较大的聚类连接, 也可以发现较稀疏的聚类连接, 同时丢弃蛋白质的数目较少。实验结果表明, CASCADE 算法的性能要优于 STM 算法。

Frey 和 Dueck 提出了"affinity propagation"(相似性传播)算法[22]。该算法基于节点之间的相似性, 在节点之间交换信息, 直到出现标本点, 此时相应节点也逐步融合在一个聚类中。该算法初始时认为所有节点都可能是标本点, 不需要预先确定聚类的个数, 逐步确立标本点, 这样错误率低, 同时该算法运算时间少。Mézard 在充分肯定该方法的基础上, 在寻找标本点的问题上提出了新的看法, 并解释了寻找标本点的过程[14]。Iqbal 与 Freitas 等人, 同样应用基于信息传递的信念传递方法, 在蛋白质相互作用的结构域网络中寻找最可能具有相互作用的结构域信息[23]。该方法在蛋白质相互作用网络中推断域间相互作用, 并利用推断出的域间相互作用预测新的蛋白质相互作用, 可以有效地避免错误预测, 并且可以系统地解决数据中存在的矛盾问题。

信息流模拟的方法可以用来解决许多问题, 如误差修正、神经网络中的学习、计算机视觉等, 其应用于聚类分析也十分有效。该方法不强调节点密度和聚类内的连接性, 避免了规模很小甚至只有一个节点的聚类结果, 但同时这样的放松条件也使得到的聚类结果不再满足"聚类内高度连接, 聚类间极少连接"的前提。由于需要考虑全部节点的信息流传递, 该方法的时间复杂度比较高。另外该方法在聚类过程中会丢弃一些蛋白质。

1.3 整体聚类方法

研究人员发现, 蛋白质相互作用网络与其他的生物网络一样, 通常具有无尺度特性和小世界特性[24]; 对于这种特性的网络, 使用传统的聚类算法往往很难得到令人满意的结果(例如会形成一个庞大的超级聚类和若干个由孤节点组成的小聚类)。在此情形下, 整体聚类法就成为增强简单聚类算法性能的有力工具[25-27]。它的目标就是把多个独立的不同聚类融合成为单一的全面聚类[28], 从而提高对无尺度网络聚类的质量。整体聚类算法的基本流程[28] 如算法 3 所示。

Asur 等人首次将整体聚类法应用于无尺度图的聚类——芽殖酵母的蛋白质相互作用网络[28]。他们使用基于网络拓扑结构的距离度量对蛋白质网络进行初始聚类, 然后使用基本聚类方法获得一系列的聚类结果, 最后在该聚类集合上使用整体聚类方法得到最后的聚类结果。实验结果表明, 使用整体聚类法得到的结果要优于传统聚类法。之后, 他们又发表了一系列的文章和技术报告, 对整体聚类法做了

进一步的研究。2007 年, 他们提出了整体聚类算法的框架结构(图 1), 应用于蛋白质相互作用网络, 证明该方法可以有效地寻找到具有生物学意义的功能聚类, 并且能够发现同一蛋白质具有多种功能的情况[29]。

算法 3 *EnsembleClustering*(G, CA, k)

输入: 无尺度图 $G=(V, E)$, 聚类数目 k

输出: $C^{Cl}=C_1^{CA}\cup\cdots\cup C_k^{Cl}$

for $i=1:|SimMetrics|$

  do for $j=1:|BaseAlgorithms|$

    do 使用每个相似性度量与每个基本算法来获得 k 个聚类结果

      $C^{i*j}=C_1^{i*j}\cup\cdots\cup C_k^{i*j}$

  end for

end for

把聚类结果转化成表示矩阵 $\boldsymbol{M}=$ represent(C^{1*1}, C^{1*2}, …, $C^{|SimMetrics|*|BaseAlgorithms|}$)

使用 CA 算法对 $\boldsymbol{M}$ 进行聚类, $C^{CA}=C_1\cup\cdots\cup C_k$

return C^{CA}

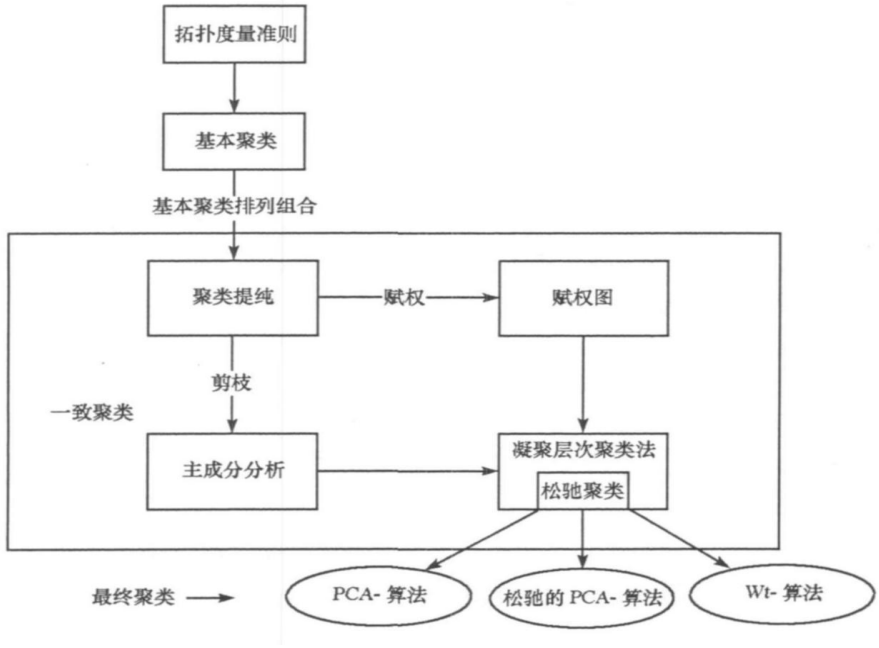

图 1 整体聚类框架[29]

Fig. 1 The framework of ensemble clustering[29]

Greene 等人使用非负矩阵分解法(NMF-like)对已有高质量的酵母蛋白质相互作用网络数据进行聚类分析[30]。该方法可以发现聚类之间的重叠以及蛋白质的多功能情况, 使用直观的树形结构表示数据(他们还提供了"NMF—树形浏览器"软件)。

尽管使用整体聚类方法在具有无尺度网络特性的蛋白质相互作用网络中取得了良好的效果, 但整体聚类法也存在着一些不可忽视的缺点。首先, 整体聚类法缺乏全局目标函数, 它使用各种标准, 在每一步局部地确定需要合并的聚类; 此外, 该方法的时间复杂度和空间复杂度较大, 也限制了它的应用。

2 蛋白质相互作用网络聚类分析中存在的困难与挑战

尽管研究者提出了许多新方法对蛋白质相互作用网络进行聚类分析, 但是当前蛋白质相互作用网络聚类分析仍存在一些难以克服的困难与挑战, 主要问题如下:

首先是原始实验数据的质量问题。目前国际公开数据库中的蛋白质相互作用数据主要是通过酵母双杂交法和质谱法等实验方法测定的, 这些实验方法各有其局限性。因此通过实验测定的数据完整性严重不足, 引入的噪声数据过多, 数据的假阳性率和假阴性率都较高; 此外, 由于测定原理的不同, 不同实验方法所引入的误差也不同, 这就导致不同的实验方法测定的蛋白质相互作用数据的吻合度较差。所有这些都给后续的研究工作带来很大的影响。

第二, 蛋白质相互作用网络本身的无尺度特性问题。蛋白质相互作用网络同其他典型的生物学网络一样, 具有较显著的无尺度特性, 即只有少数节点的度较大, 而其他大部分节点的度较小。因此使用传统聚类方法往往会得到严重不均衡的聚类结果: 只有少数(通常只有一个或几个)规模庞大的聚类, 而其他为数众多的聚类只包含了少数的结点(可能只有一个节点)。因此, 如何结合蛋白质相互作用网络自身独特的拓扑结构信息, 改进已有的聚类算法, 是一个亟待解决的问题。

第三, 聚类结果的评价与比较标准难以界定问题。由于模块边界以及蛋白质相互作用网络的层次结构并不十分明确, 利用现有的聚类方法所得到的聚类结果并不相同, 因此选择合适的评价与比较标准就成为一个难题。现有的评价与比较标准为数众多, 但是没有形成 benchmark。没有统一的评价与比较标准, 得到的聚类结果难以令人信服。

3 展 望

纵观近些年来蛋白质相互作用网络聚类研究的发展, 人们借助复杂网络、图论以及信息论等理论和方法取得了丰硕的成果。但随着研究的深入和认识的加深, 在蛋白质相互作用网络的聚类研究中还有许多亟待解决的问题, 蕴藏着极大的研究潜力。

虽然目前借助聚类分析对蛋白质相互作用网络的拓扑结构已经有了一定程度的了解, 但对于这种结构的形成机理以及进化和自然选择在蛋白质相互作用网络拓扑结构形成方面所起的作用还不十分清楚。因此将聚类分析与生物的进化过程相结合研究蛋白质相互作用网络的拓扑结构形成将是今后的研究趋势之一。

其次, 随着系统生物学研究的不断深入和生物学网络整合程度的不断加深, 蛋白质相互作用网络的聚类研究也将更多地考虑基因调控、信号转导、代谢等其他网络的信息和网络间的联系。因此整合各种生物网络的信息, 并根据生物体的特点, 考虑网络中存在的正负反馈环的影响, 从系统生物学角度进行蛋白质相互作用网络的聚类研究将成为今后研究的一个重要方向。

参 考 文 献:

[1] Von Mering C, Krause R, Snel B, et al. Comparative Assessment of Large-scale Data Sets of Protein-protein Interactions[J]. Nature, 2002, 417: 399—403.

[2] Tan P N, Kumar V. 数据挖掘导论[M]. 范明, 范宏建, 等译. 北京: 人民邮电出版社, 2006.

[3] Lin C, Hwang W C, Pei P J, et al. Knowledge Discovery in Bioinformatics: Techniques, Methods and Application[C]//Wiley Series on Bioinformatics: Computational Techniques and Engineering. John Wiley &Sons, Inc., 2006.

[4] 高琰, 谷士文, 唐玉进, 等. 机器学习中谱聚类方法的研究[J]. 计算机科学, 2007, 34(2): 201—203.

[5] Donath W E. Lower Bounds for the Partitioning of Graphs[J]. IBM J. Res. Develop., 1973, 17: 420—425.

[6] M F. Algebraic Connectivity of Graphs[J]. Czechoslovak Math. J., 1973, 23: 298—305.

[7] Bu D, et al. Topological Structure Analysis of the Protein-protein Interaction Network in Budding Yeast[J]. Nucleic Acids Research, 2003, 31(9): 2443—2450.

[8] Lu H, et al. The Interactome as a Tree-an Attempt to Visualize the Protein-protein Interaction Network in Yeas[J]. Nucleic Acids Research, 2004, 32(16): 4804—4811.

[9] 董蕴源, 王正华, 王勇献. 啤酒酵母、秀丽线虫和大肠杆菌蛋白质相互作用网络的近似二分模式分析[J]. 上海交通大学学报,

2008, 42(5): 701—706.

[10] Kamp C, Christensen K. Spectral Analysis of Protein-protein Interactions in Drosophila Melanogaster[J]. Physical Review, 2005, 71(4).

[11] 董蕴源, 王正华, 王勇献. 利用基于Normal矩阵的谱平分法挖掘酵母蛋白质相互作用网络中的社团[J]. 激光生物学报, 2008, 17(1): 13—18.

[12] Sen T Z, Jernigan R L. Functional Clustering of Yeast Proteins from the Protein-protein Interaction Network[J]. BMC Bioinformatics, 2006, 7: 355.

[13] Luxburg U V. A Tutorial on Spectral Clustering[R]. Max Planck Institute for Biological Cybernetics, 2006.

[14] Mézard M. Where Are the Exemplars[J]. Science, 2007, 315: 949—951.

[15] Hwang W C, Zhang A D, Ramanathan M. CASCADE: A Novel Quasi All Paths-based Network Analysis Algorithm for Clustering Biological Interactions[J]. BMC Bioinformatics, 2008, 9: 64.

[16] Geng H M, Ali H. A New Approach to Clustering Biological Data Using Message Passing[C]//Proceedings of the 2004 IEEE Computational Systems Bioinformatics Conference (CSB 2004), Stanford, CA, USA: IEEE Computer Society, 2004.

[17] Geng H M, Ali H. A New Clustering Algorithm Using Message Passing and Its Applications in Analyzing Microarray Data[C]//Proceedings of the Fourth International Conference on Machine Learning and Applications (ICMLA'05), IEEE Computer Society, 2005.

[18] Geng H M, Ali H. Message Passing Clustering (MPC): A Knowledge-based Framework for Clustering under Biological Constraints[J]. Int. J. Data Mining and Bioinformatics, 2008, 2(2): 95—120.

[19] Geng H M. A New Clustering Strategy with Stochastic Merging and Removing Based on Kernel Functions[C]//Proc. Workshop and Abstract of IEEE Computer Society Bioinformatics Conf. (CSB'05), Stanford, CA, USA: IEEE Computer Society, 2005.

[20] Geng H M, Bastola D, Ali H. On Clustering Biological Data Using Unsupervised and Semi-supervised Message Passing[C]//Proc. IEEE 5th Symposium on Bioinformatics and Bioengineering (BIBE'05), Minneapolis, MN, USA, 2005.

[21] Hwang W C, Zhang A D, Murali R. A Novel Functional Module Detection Algorithm for Protein-protein Interaction Networks[J]. Algorithms for Molecular Biology, 2006, 1: 24.

[22] Frey B J, Dueck D. Clustering by Passing Messages between Data Points[J]. Science, 2007, 315: 972—977.

[23] Iqbal M, Johnson C G, Vergassola M. Message-passing Algorithms for the Prediction of Protein Domain Interactions from Protein-protein Interaction Data[J]. Bioinformatics, 2008, 24(18): 2064—2070.

[24] Barabasi A L, Oltvai Z N. Network Biology: Understanding the Cell's Functional Organization[J]. Nat. Rev. Genet., 2004, 5(2): 101—113.

[25] Topchy A, Punch W. A Mixture Model for Clustering Ensembles[C]//Proc. SIAM Conf. on Data Mining, 2004.

[26] Topchy A, Jain A K, Fred A. Analysis of Consensus Partition in Cluster Ensemble[C]//IEEE International Conference on Data Mining (ICDM), 2004.

[27] Gionis H M, Tsaparas P. Clustering Aggregation[C]//21st International Conference on Data Engineering (ICDE'05), 2005.

[28] Asur S, Parthasarathy S, Ucar D. An Ensemble Approach for Clustering ScaleFree Graphs[C]//Proceedings of Workshop OnLink Analysis, Dynamics and Static of Large Network (LinkKDD'06). Philadelphia, Pennsylvania, USA, 2006.

[29] Asur S, Ucar D, Parthasarathy S. An Ensemble Framework for Clustering Protein-protein Interaction Networks[J]. Bioinformatics, 2007, 23: 29—40.

[30] Greene D, Cagney G, Krogan N, et al. Ensemble Non-negative Matrix Factorization Methods for Clustering Protein-protein Interactions[J]. Bioinformatics, 2008, 24(15): 1722—1728.