基于基因 - 疾病网络的重叠社区发现算法研究*

戴彩艳 何 菊 胡孔法 丁有伟

（南京中医药大学信息技术学院 南京 210016）

〔摘要〕 介绍基因 - 疾病网络以及社区挖掘方法的基本概念，阐述基于基因 - 疾病网络的重叠社区发现交叉迭代方法过程及框架并进行实验验证，揭示基因 - 疾病之间存在的关系，为生物学和临床医学的诊断以及药物开发等方面提供理论指导。

〔关键词〕 基因 - 疾病网络；社团结构；交叉迭代算法；重叠社区

〔中图分类号〕 R - 056 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673 - 6036.2018.10.015

Study on Overlapping Community Discovery Algorithms Based on the Gene - disease Network *DAI Cai - yan, HE Ju, HU Kong - fa, DING You - wei, College of Information Technology, Nanjing University of Chinese Medicine, Nanjing 210016, China*

〔**Abstract**〕 The paper introduces the basic concepts of the gene - disease network and the community mining method, sets forth the overlapping community discovery cross iteration method process and framework based on the gene - disease network, carries out the experimental verification, as well as reveals the relation between genes and diseases, so as to provide theoretical instructions for diagnosis of biology and clinical medicine and drug development.

〔**Keywords**〕 gene - disease network; community structure; cross iteration algorithm; overlapping community

1 引言

目前人们已逐步完成人类基因组计划[1-2]和众多模式生物测序项目，产生大量生物序列数据[3]。但是数据与信息是无法等同的，在分析生物功能的过程中研究者们发现基因和蛋白质很少单独行动，相反其往往是通过多种基因和蛋白质共同作用对生物系统产生影响，于是出现成组基因的相关研究。基因研究进入后基因组时代，其主要目标是通过对基因组的功能分析来研究生物系统功能变化机制[4]。很多常见的人类疾病，如癌症、糖尿病和老年痴呆症等，都是复杂疾病，在分子水平上相当复

〔收稿日期〕 2018 - 04 - 14

〔作者简介〕 戴彩艳，讲师，博士；通讯作者：何菊，讲师。

〔基金项目〕 江苏省高校自然科学基金项目“基于动态蛋白质网络进行复合物挖掘及功能预测算法研究”（项目编号：18KJB520040）；国家自然科学基金项目“面向中医临床大数据的现代名老中医肺癌辨治规律并行挖掘策略及方法学研究”（项目编号：81674099）；“基于计算智能的心系基础证量化诊断方法学研究”（项目编号：81503499）。

杂。复杂疾病也被称为多因子疾病，原因在于多个方面，包括遗传、日常习惯以及环境因素，但是这些因素对于复杂疾病形成的影响尚未确定，还有待进一步研究。因为表型的最后结果会受到多基因和环境因素的影响[5]，所以表型的识别很难，导致对复杂疾病产生影响的基因很难被发现。针对不同的复杂疾病，相同的因素在其中所产生的影响也不同。一般某种复杂疾病都是由于多个基因共同作用产生异常引起[6]。对基因 - 疾病网络的研究[7]已成为研究热点，而其中的变化情况较为复杂，因此该研究富有意义的同时也存在着巨大的挑战[8]。

随着大量生物信息数据的出现，传统的遗传学方法已难以满足复杂疾病的研究需求，这就要求开发出新的生物信息学方法来快速准确地辨识这些复杂疾病产生的病理机制。基于基因 - 疾病网络的研究不仅能更好地理解基因与疾病之间的关联，而且从系统生物学角度来说其能够在诊断和治疗复杂疾病时提供一定的线索。本文基于基因 - 疾病网络，采用交叉迭代算法对其中存在的重叠社区进行挖掘，开发新的基因 - 疾病关联的有效方法。

2 基本概念

2.1 基因 - 疾病网络

基因 - 疾病网络含有两种不同的节点类型，即疾病和基因，其网络结构，见图 1。上半部分圆圈代表基因 - 疾病网络中的基因，下半部分方框代表疾病，两种不同类型节点之间的连线代表已知的疾病与基因之间的关联关系。近年来随着高通量生物技术的产生，出现大规模的基因疾病网络数据。目前关于基因疾病交互作用的研究远远不足，但现有数据在一定程度上能够体现与已知生物过程相关的拓扑特征[9]，为从系统生物学角度理解人类疾病的病理机制提供良好的平台。在已发表的文献中有大量的基因 - 疾病生物学数据，然而目前很少有人研究疾病、基因之间的联系，所以这项工作是具有挑战性的任务[10-11]。现在很多研究已表明具有相同或相似表型的疾病在基因网络、拓扑距离上具有相似性。因此能够认为在基因疾病网络上与已知致病基因的功能在拓扑结构上重叠度比较大，那么该基因也可能是导致这种疾病的关键基因。即便当前的基因 - 疾病网络仍存在缺陷，有些链接可能是假阳性的，但是从大规模的生物信息数据中提取出来的有用信息也可以显示出与生物特性有关的网络拓扑结构[12-13]。因此在基因 - 疾病网络进行社区发现的操作可以将相同或相似表型的疾病及其各自对应的基因划分到同一社区中，下一步就可以推断未知疾病的致病基因。

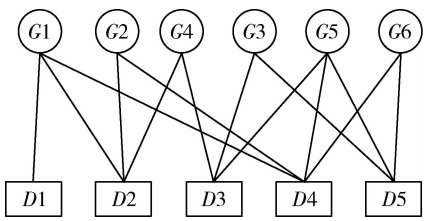

图 1 基因 - 疾病网络结构

2.2 社区挖掘方法

近年来出现很多网络社区挖掘算法，针对不同学科的数据采用侧重点不同的社区划分方法。按其基本原理可分为基于划分、基于模块性优化、基于标签传播的社区划分方法等。基于划分的社区挖掘算法中最有代表性的是 GN 算法[14]。该算法在复杂网络社区挖掘研究中有非常重要的地位，因为其首先提出网络结构中存在的社区概念，首次发现复杂网络中普遍存在的社区结构，为其他研究者对社区问题的研究打下基础，于是出现从各个角度对复杂网络社区挖掘的方法。总而言之研究者们都是在该算法的基础上提出很多改进算法[15-17]。基于模块度优化的社区挖掘算法中衡量方法最有代表性的是 Newman[18]提出的。其选取候选解的搜索方法是：选取两个已有的社区进行合并，刚开始候选解中每个社区只有 1 个节点；在迭代过程中选取使得模块度函数 Q 值增量最大的社区并对其进行合并；最终候选解只与 1 个社区相对应时就标志着算法结束。该算法使用自底向上的层次聚类过程，输出层次化的聚类树，将对模块度函数 Q 值最大的社区划分作为最后社区划分的结果。而基于标签传播方法属于启发式算法，与基于模块性方法不同的是该方法不存在特定的目标函数。使用该方法遵循的规则是：

在具有社区结构的网络中对于任意节点都应该与它的大部分邻居处于同 1 个社区。Raghavan 等人提出具有代表性的标签传播算法[19]。该算法的流程为：初始化时为每个节点设定 1 个唯一的标签；在迭代过程中每个节点选取其多数邻居的标签对各自的标签进行更新；当网络中所有节点的标签都与其多数邻居的标签相同时就认定算法结束。这时稠密子图中的所有节点通过标签达到一致，于是便形成社区。上述研究很多都不是针对生物数据展开的研究且存在不足：将两部网络中的数据投影到单部网络中，网络数据存在信息缺失；难以发现基因 - 疾病网络中存在的重叠社区，也就是两类数据社区划分后形成的关系是 1 对 1 的；需要事先给出各类数据聚类的数目；要求两种不同类型节点的聚类数目相等。

3 基于基因 - 疾病网络的重叠社区发现算法

3.1 交叉迭代算法

在本文中算法大体分为 3 步。第 1 步将原有的基因 - 疾病网络映射成基因网络和疾病网络；第 2 步对得到的基因网络或疾病网络进行聚类，得到初始社区；第 3 步是将前两步的初始社区进行交替迭代。在迭代的每一步，固定某一边的顶点（设为 X 部分）的社区划分，通过计算社区的链接度来判断 Y 部顶点应归属的社区，从而将 Y 部的顶点进行社区划分。在下一步的迭代中再固定 Y 部顶点的社区划分，将 X 部的顶点进行社区划分。重复这样的交叉迭代过程直到满足结束条件为止。由于第 3 步的迭代是以第 2 步挖掘出来的初始社区为基础，也就是在第 2 步时最终社区个数已确定，这就对第 2 步初始社区挖掘的准确率有较高的要求，因为初始社区的质量直接影响最后挖掘结果的正确性。

3.2 算法过程

假设 1 个基因 - 疾病网络由 X 节点 $\{x_1, x_2, \cdots, x_8\}$ 和 Y 节点 $\{y_1, y_2, \cdots, y_6\}$ 组成。X 节点之间没有边连接，而是与 Y 节点互相连接，见图 2。（1）第 1 步，遍历网络图。要得到与 X 相关的节点需对 Y 节点进行遍历，首先是根据与 y_1 相关的节点，与 y_1 相连的节点之间的度设为 1，见图 3。然后是 y_2，与 y_2 相连的 X 节点分别是 x_1、x_2。这时需要判断之前得到的图中有没有这两个节点。如果两个节点都在，则其度增加 1；如果有 1 个节点在，则在这个节点基础上增加 1 个新的节点，其度为 1；如果都不在，则创建一个新的图，其度为 1，见图 4。以此类推在对上图 Y 节点遍历结束后得到与 X 节点的图，见图 5。（2）第 2 步，按照图中节点度的大小进行聚类。得到与 X 相关的聚类为 $\{x_1, x_2, x_4\}$、$\{x_3, x_5\}$、$\{x_6, x_7, x_8\}$。按照同样的方法，遍历 X 节点，得到与 Y 节点相关的一部图，根据图中的度进行聚类，得到与 X 相关的聚类为 $\{y_1, y_2, y_4\}$，$\{y_3, y_5\}$，$\{y_6\}$。（3）第 3 步，彼此迭代。找出稳定的基因 - 疾病社区。这时已达到稳定。最后通过图对应的邻接矩阵可以得到划分好的社区，见图 6。得到社区后可以计算出社区的模块度，以便和其他算法比较，验证算法的有效性。

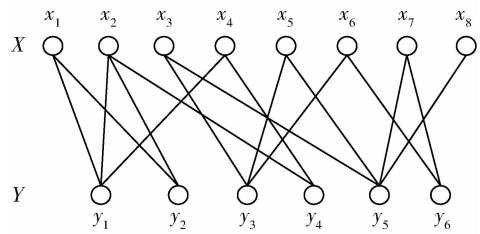

图 2 基因 - 疾病网络

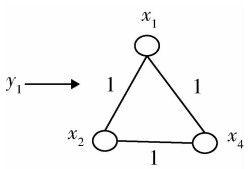

图 3 遍历 y_1 节点得到的 X 节点

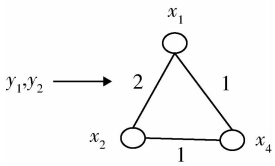

图 4 遍历 y_1 y_2 节点得到的 X 节点

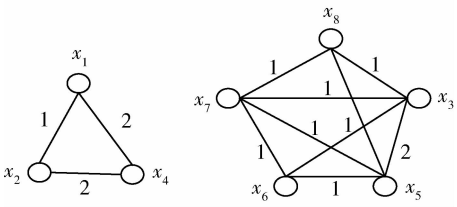

图 5 Y 节点遍历结束后得到与 X 相关节点

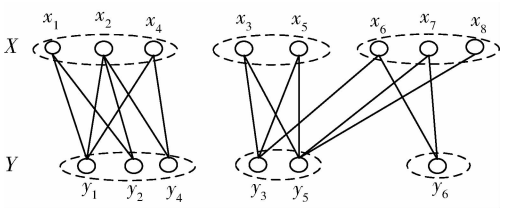

图 6 划分好的社区

3.3 算法框架

采用基因 – 疾病网络的重叠社区挖掘算法。输入：G：基因 – 疾病网络，其关系矩阵为 R；输出：基因 – 疾病网络社区挖掘结果。

开始

1 通过基因 – 疾病网络中两种类型节点之间的关系，将原有的整体网络映射成两部分：基因网络和疾病网络；

2 对基因网络中的社区进行初始化；

/ * 通过聚类方法得到 L 个初始社区：$T = \{X_1 \ldots X_L\}$ * /

3 For $i = 1$ to L do

4 repeat

5 将对应的疾病网络中的节点划分到已有基因社区 X_i 中；

/ * 求 Y_i 使得疾病网络中待划分的节点与已有基因社区 X_i 连接的边最多 * /

6 将对应的基因网络中的节点划分到已有疾病社区 Y_j 中；

/ * 求 X_j 使得基因网络中待划分的节点与已有疾病社区 Y_j 连接的边最多 * /

7 Until 节点所在社区不再变化；

8 End for i

9 输出基因 – 疾病网络社区挖掘结果；

结束

该算法存在以下优点：（1）投影图带权重，网络中本身存在的信息不易缺失。（2）可以发现基因 – 疾病网络中存在的重叠社区，也就是一种疾病可以存在于多个基因社区中，一种基因也可能对应多个疾病社区。（3）不需要事先给出各类数据聚类的数目。（4）不要求两种不同类型节点的聚类数目相等。

4 实验分析

4.1 环境

为评估所提出的重叠社区发现算法（Overlapping Community Detection Algorithm，Olap）在基因 – 疾病网络中社区挖掘时的性能，在真实基因 – 疾病数据集上进行一系列实验。对该算法的实验结果与其他 4 种算法 CN[20]、JC[21]、DD_{ALL}[22] 以及 LPCS[23] 进行比较。实验在内存为 4G、CPU 为 1.9GHz 的机器上运行，程序采用能够快速处理生物信息学数据的 Python 语言编写。

4.2 数据

本文所使用的基因疾病网络[23]有 4 种类型的链接。根据本体论对疾病进行编码和分类，按照 HUGO 对基因进行命名。4 种类型的链接分别是遗传基因关联、性状链接、基因 – 基因交互以及磷酸基序查找，分别缩写为 G，P，PPI 以及 F。该网络的组成成分，见表 1。其中代表疾病的节点有 703 个，代表基因的节点有 1 132 个，4 种不同类型的边条数分别是 10 483，74 523，2 450 以及3 279。

表 1 基因 – 疾病网络中节点和边

疾病	基因	G	P	PPI	F
703	1 132	10 483	74 523	2 450	3 279

4.3 结果

在基因 – 疾病网络上 5 种算法测试得到的 AUC 值，见表 2，每行代表在 1 种链接上的挖掘结果。每行得分最高的 AUC 被加粗字体进行强调，次高得分的 AUC 被加下划线进行强调。5 种算法中 Olap 的 AUC 得分在 4 种关系中都是最高的。在基因 – 疾病网络上通过实验得到的 5 种算法的精度、召回率

以及 F－measure，见图 7、图 8、图 9。从图 7 中可以发现在基因－疾病网络上进行社区挖掘时，Olap 在 5 种算法中的精度最高。如算法 JC 的精度较 Olap 对应的精度低 10% 左右。从图 8 中可以看到 Olap 在处理关系 G 、P 以及 F 时，召回率最好，而在处理关系 PPI 时，召回率略低于最高的算法。从图 9 中可以发现 Olap 在处理关系 G、P 以及 F 时 F－measure 最理想，在处理关系 PPI 时 Olap 的 F－measure 略低于 LPCS，但是优于其他 3 种算法。

表 2 在基因－疾病网络上各算法的 AUC 值

数据集	算法				
	CN	JC	DD_{ALL}	LPCS	Olap
G	0.922	0.936	0.940	0.946	0.958
P	0.901	0.768	0.906	0.922	0.938
PPI	0.768	0.776	0.831	0.822	0.833
F	0.811	0.823	0.844	0.847	0.851

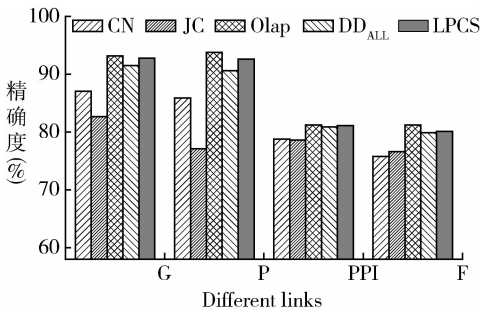

图 7 5 种算法在基因－疾病网络上的精度

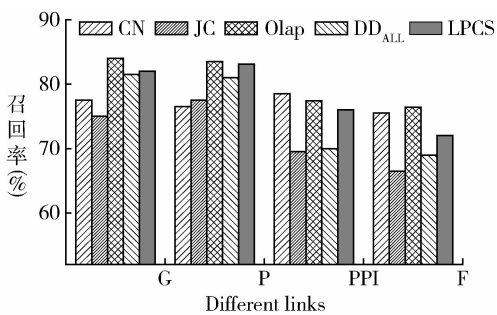

图 8 5 种算法在基因－疾病网络上的召回率

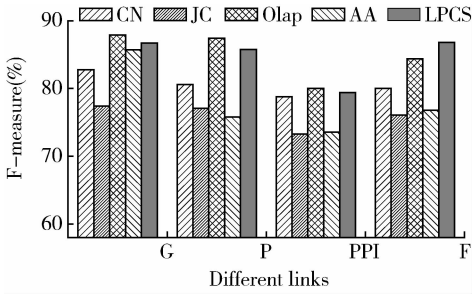

图 9 5 种算法在基因－疾病网络上的 F－measure

5 结语

本文提出基于基因－疾病网络进行重叠社区发现的新方法，首先对网络中的每类数据进行聚类，按照聚类结果彼此迭代。当各个节点所属的类不再发生变化时可以认定社区划分结束。为验证该算法的正确性和效果，对真实数据集进行社团划分的实验。结果表明该算法效果较为理想，可以在不丢失信息以及不规定聚类数目的基础之上发现不同类型节点组成的社区，可以找出社区之间 1 对多的关系。

参考文献

1 Radin J. Human Genome Diversity Project: history [M]. Holland: Elsevier, 2015: 306－310.

2 Rizzo J M, Buck M J. Key Principles and Clinical Applications of "next－generation" DNA Sequencing [J]. Cancer Prevention Research, 2012, 5 (7): 887－900.

3 Green E D, Watson J D, Collins F S. Human Genome Project: twenty－five years of big biology [J]. Nature, 2015, 526 (7571): 29.

4 Huss J. Methodology and Ontology in Microbiome Research [J]. Biological Theory, 2014 , 9 (4): 392.

5 Jones MB, Highlander SK, Anderson EL, et al. Library Preparation Methodology can Influence Genomic and Functional Predictions in Human Microbiome Research [J]. Proceedings of the National Academy of Sciences of the United States of America, 2015, 112 (45): 14024－14039.

6 Widmann P, Reverter A, Fortes MRS, et al. A Systems Biology Approach Using Metabolomic Data Reveals Genes and Pathways Interacting to Modulate Divergent Growth in Cattle [J]. Bmc Genomics, 2013, 14 (1): 798.

7 Chu J H, Hersh C P, Castaldi P J, et al. Analyzing Networks of Phenotypes in Complex Diseases: methodology and applications in COPD [J]. BMC Systems Biology, 2014, 8 (1): 78.

8 Zuzana Z, Adam B, Ruzena T, et al. T Neven . POPE Study: rationale and methodology of a study to phenotype patients with COPD in Central and Eastern Europe [J]. International Journal of Chronic Obstructive Pulmonary Disease, 2016, 11 (Issue 1): 611.

9 Coneva V, Simopoulos C, Casaretto J A, et al. Metabolic and Co - expression network - based Analyses Associated with Nitrate Response in rice [J]. Bmc Genomics, 2014, 15 (1): 1 - 14.

10 Emre G, Jörg M, Marc V, et al. Network - based in Silico-drug Efficacy Screening [J]. Nature Communications, 2016 (7): 10331.

11 Bergman J, Mitrikeski P T, Bŏrcic K. Dominant Epistasis Between Two Quantitative Trait Loci Governing Sporulation Efficiency in Yeast Saccharomyces Cerevisiae [J]. Food Technology & Biotechnology, 2015, 53 (4): 367.

12 Ucisikakkaya E, Leatherwood J K, Neiman A M. A Genome - wide Screen for Sporulation - defective Mutants in Schizosaccharomyces Pombe [J]. G3 - Genes Genomes Genetics, 2014, 4 (6): 1173 - 1182.

13 Rahmani H, Blockeel H, Bender A. Using a Human Disease Network for Augmenting Prior Knowledge About Diseases [J]. Intelligent Data Analysis, 2015, 19 (4): 897 - 916.

14 Girvan M, Newman M E J. Community Structure in Social and Biological Networks [J]. Proc. Natl. Acad. Sci. USA, 2002, 99 (12): 7821 - 7826.

15 Shan J, Shen D, Nie T , et al. An Efficient Approach of Overlapping Communities Search [C]. Hanoi: International Conference on Database Systems for Advanced Applications, 2015.

16 Zhou L, Yang P, Wang L, et al. An Approach for Overlapping and Hierarchical Community Detection in Social Networks Based on Coalition Formation Game Theory [J]. Expert Systems with Applications an International Journal, 2015, 42 (24): 9634 - 9646.

17 Sah P, Singh LO, Clauset A, et al . Exploring Community Structure in Biological Networks with Random Graphs [J]. Bmc Bioinformatics , 2014 , 15 (1): 220

18 Newman M E J, Girvan M. Finding and Evaluating Community Structure in Networks [J]. Physical Review E-Statistical, Nonlinear and Soft Matter physics, 2004, 69 (2): 026113.

19 Raghavan UN, Albert R, Kumara S. Near Linear Time Algorithm to Detect Community Structures in Large - scale Networks [J]. Ai Communications, Physical Review E Statistical Nonlinear & Soft Matter Physics, 2007, 76 (3 Pt 2): 036106.

20 Lü L Y, Zhou T. Link Prediction in Complex Networks: a survey [J]. Physica A Statistical Mechanics & Its Applications, 2011, 390 (6): 1150 - 1170.

21 Zhou T, Lü L Y, Zhang Y C. Predicting Missing Links via Local Information [J]. Eur Phys J B, 2009, 71 (4): 623 - 630.

22 Chungmok Lee, Minh Pham, Myong K, et al. A Network Structural Approach to the Link Prediction Problem [J]. INFORMS Journal on Computing, 2015 (27): 249 - 267.

23 Wang Z, Wu Y, Li Q, et al. Link Prediction Based on Hyperbolic Mapping with Community Structure for Complex Networks [J]. Physica A: Statistical Mechanics and its Applications, 2016 (450): 609 - 623.