基于图元向量的差异共表达分析研究

肖碧玉,李先斌,刘文斌
(温州大学物理与电子信息工程学院,温州,浙江 325035)

摘 要: 差异分析对于揭示生命体的生长、发育和衰老过程及疾病发生具有重大的意义,基于网络的差异分析方法已经成为系统生物学的一个研究热点.Przulj 提出的图元及图元向量作为一描述网络局部结构信息的方法,已经在网络分析方法面取得了很多重要的结果.本文在图元向量的基础上提出了二种节点变化的差距度量方法,通过聚类可以分别挖掘网络中模块内变化基因簇和模块间变化基因簇.应用 AGEMAP 数据库中 12 个小鼠组织基因表达数据的结果表明:大部分聚类簇都高度显著富集与衰老相关的 GO 条目.

关键词: 图元向量; 差异网络; 小鼠衰老
中图分类号: **文献标识码:** A **文章编号:** 0372-2112 (2015)10-2009-05
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2015.10.019

Mining Differential Co-expression Clusters Based on Graphlet Orbits

XIAO Bi-yu, LI Xiang-bin, LIU Wen-bin
(*Department of Physics and Electronic information engineering, Wenzhou University, Wenzhou, Zhejiang* 325035, *China*)

Abstract: Differential analysis is a major method to understand the process of biological evolution and the progress of diseases. Recently, differential analysis based on graph has been a hot area in system biology. Since Przulj proposed the conception of graphlet and graphlet orbits, both of them have been used in various network analyses. In this paper, we proposed two measures based on graphlet orbits, and we used them to mine within module differential co-expresssion clusters and module to module differential co-expression clusters. Applications on the data of mice for 16 months and 24 months have shown that most of the clusters are significant enrichment in some GO terms related with aging.

Key words: graphlet orbit; differentially network; aging

1 引言

系统生物学的研究表明:生命体的生长、发育、衰老过程以及疾病的发生与基因之间相互作用的变化密切相关.随着以微阵列技术为代表的各种高通量技术的飞速发展,人们可以同时观察到一个细胞中成千上万个基因的活动状况.如何利用这些数据挖掘出其中的潜在变化,揭示衰老及疾病的发生发展,已经成为系统生物学研究的一个热点和难点[1,2].基因差异表达的分析大致可以分为三个层次:(1)基因表达水平的变化,如 Jacob 等发现小鼠衰老期各组织的基因表达水平变化存在较大的差异,可归为神经组织、血管组织、类固醇反应组织三种衰老类型[3];(2)基因之间相互作用关系的变化.如 Remondini 通过比对网络度的分布,发现致癌基因 c-myc 的活性与网络结构变化有密切关系[4],Voy 等通过网络差异分析确定小鼠受辐射影响的基因簇[5],Oldham 等比对人类和黑猩猩差异网络的拓扑重叠,挖掘出与进化有关的基因集合[6],Zhang 提出差异相关子网络(DDN),检测两个差异转录网络拓扑变化的显著性[7].Southworth 等构建基因带权差异网络,挖掘与小鼠衰老密切相关的基因模块[8];(3)基因簇间关系的变化.如 Tesson 等提出了 DiffCoEx 算法,可以同时挖掘基因簇及基因簇间的变化[9].此外,在差异表达分析方面,Fang 研究了二组样本中表达水平高低的变化,提出了三种基因表达谱差异模式及其关系,这种模式特征的研究对于差异模式的挖掘具有重要的指导意义[10,11].

虽然差异表达的研究已经取得了很多成果,但是由于网络结构的复杂性,现有的算法往往仅能挖掘具有某种特征的差异模式.差异模式研究的关键是如何刻画网络的局部拓扑结构变化.Przulj 提出图元及图元向量研

收稿日期:2013-01-06;修回日期:2013-03-19;责任编辑:郭游
基金项目:国家自然科学基金(No.61572367,No.61272018);浙江省自然科学基金(No.R1110261);浙江省新苗人才计划(No.2012R424051)

究生物网络与随机网络的拓扑结构差异、癌症相关基因、生物网络的进化树[12~15]等.本文将研究图元向量在差异模式的挖掘中的应用,并利用 AGEMAP 数据库中小鼠的基因表达数据进行了分析.

2 图元及图元向量相关概念

图元就是一个小的子图,图 1 列出了包含 2、3、4 个节点的非同构图元.为了刻画节点的拓扑等价性,Przulj 把图元中具有相同拓扑位置的节点标记为相同的记号(白色、黑色、灰色).然后对其中具有不同拓扑位置的节点唯一标号.图 1 共包含三种图元的 15 个不同的拓扑位置.它们出现的频率即为图元向量.由于这些标号可能存在维度信息冗余(如标号 14 与 3,7 与 2),Milenkovi 按照标号大的图元向量包括标号小的图元向量的个数对每一维图元向量定义一个权值 w_k($k=0,\cdots,14$).

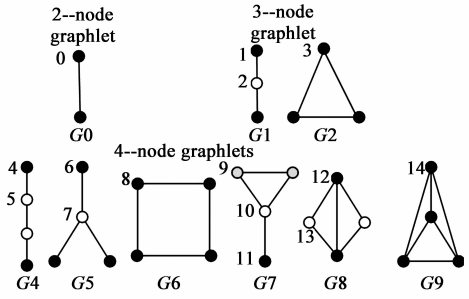

图1 图元及图元向量

图元向量细致地刻画一个节点与其近邻的拓扑关系.当网络的局部结构发生变化时,每个节点与其邻节点之间形成的各种图元的数量将会发生相应的变化.显然,一个节点的局部结构发生的变化越大,其图元向量的变化就可能越大.给定二个网络 $G(V,E)$和 $G(V,E')$,我们将节点 $v\in V$ 的图元向量差异度定义为

$$D_{vv'}=\frac{\sum_{k=0}^{14} w_k(\log(v_k+1)-\log(v_k'+1))}{w\log(\max\{v_k,v_k'\}+2)} \quad (1)$$

其中 v_k 和 v_k' 分别表示节点 v 在 $G(V,E)$和 $G(V,E')$中的第 k 维图元向量,$w=\sum_{k=0}^{14} w_k$. 度是反映节点在网络中地位或重要性的一个非常重要的指标.因此,我们定义节点 $v\in V$ 的局部结构差异度为

$$D_{vv'}'=d_v * D_{vv'} \quad (2)$$

d_v 为顶点 v 在一个网络中的度.直观上,同一个网络中簇的节点变化,应该具有一定的相似性,即对于一个簇中的节点 $u,v\in V'$,他们的局部结构差异度可能基本相当;另一个就是它们图元向量差向量具有相似性.因此,下面我们给出二个衡量节点变化的差距的定义:

$$L_{uv}^1=|D_{uu'}'-D_{vv'}'| \quad (3)$$

$$L_{uv}^2=D_{\Delta u\Delta v} \quad (4)$$

其中 $\Delta u,\Delta v$ 分别为节点 u,v 的在二个网络中的图元向量的差向量.

3 差异网络研究

由于引起网络结构差异的原因各种各样,下面我们通过二个简单的例子来观察图元向量在差异分析中的应用.

3.1 模块内变化

图 2 为一个有 5 个节点的模块在两种情况下的连接情况.如果使用度或传统的稠密子图模块方法,我们得到的模块将是 1,3,4,5,无法捕捉到节点 2 所发生的巨大变化.表 1 中列出图 2 中 5 个节点在两个网络中的图元向量的差向量和图元向量差异度($D_{vv'}$),可以看出,虽然节点 2 与节点 1,3,4,5 的各维图元向量的变化(差向量)在 0,1,2,6,7 这些维度的变化不一致,但是它们在网络 G_A 中的变化程度却基本相当.利用节点局部结构变化差异度,1、2、3、4、5 可以作为一个模块被整体挖掘到.

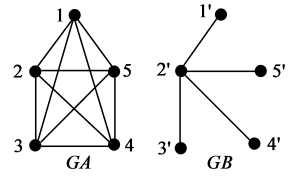

图2 差异网络模块基因变化实例

表 1 图 2 中 5 个节点在两个网络中的图元向量的差向量和图元向量差异度($D_{vv'}$)

维度 / 节点	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	$D_{vv'}$
1	3	−3	0	6	0	0	−3	0	0	0	0	0	0	0	4	0.29
2	0	0	−6	6	0	0	0	−4	0	0	0	0	0	0	4	0.26
3	3	−3	0	6	0	0	−3	0	0	0	0	0	0	0	4	0.29
4	3	−3	0	6	0	0	−3	0	0	0	0	0	0	0	4	0.29
5	3	−3	0	6	0	0	−3	0	0	0	0	0	0	0	4	0.29

3.2 模块间变化

图 3 为一个有 4 个模块的网络在两种情况下的连接情况.其中节点 6、10、17、22,主要起连接模块与模块之间的关系,它们的变化将引起整个模块发生质的蜕变,挖掘这样的变化节点具有重要的意义.表 2 中列出了图 3 中各节点在两个网络的图元向量差向量和差异度($D_{vv'}$),可以看出 6、10、17、22 在二个网络的图元向量差向量非常相似.特别在第 10 维,它们与其他的节点的区别更明显.

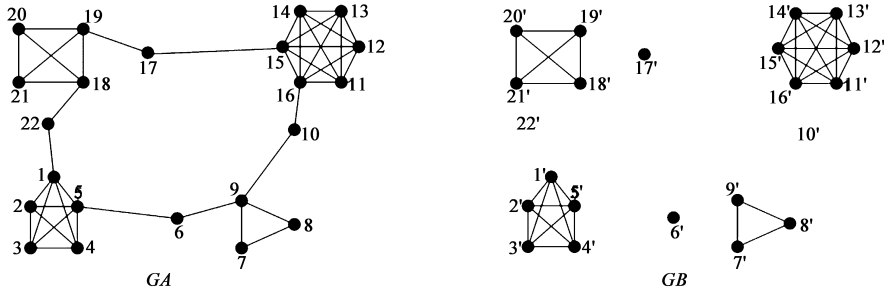

图3　差异网络中连接模块基因与模块基因的变化基因

表 2　图 3 中各节点在两个差异网络的图元向量差向量和差异度($D_{vv'}$)

维度 / 节点	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	$D_{vv'}$
1	1	5	2	2	6	4	5	0	0	0	0	3	6	0	4	0.49
2	0	4	2	6	4	2	0	0	0	0	0	6	0	0	4	0.42
3	0	4	2	6	4	2	0	0	0	0	0	6	0	0	4	0.42
4	0	4	2	6	4	2	0	0	0	0	0	6	0	0	4	0.42
5	1	5	2	2	6	4	5	0	0	0	0	3	6	0	4	0.49
6	2	2	7	1	0	2	7	2	0	0	7	0	0	0	0	0.46
7	0	2	2	1	1	2	0	1	0	0	0	2	0	0	0	0.34
8	0	2	2	1	1	2	0	1	0	0	0	2	0	0	0	0.34
9	2	4	2	4	1	9	6	0	2	0	0	0	2	0	0	0.45
10	2	2	8	10	2	2	2	0	0	1	1	0	0	0	0	0.47
11	0	5	2	10	10	2	0	0	0	0	0	8	0	0	10	0.43
12	0	5	2	10	10	2	0	0	0	0	0	8	0	0	10	0.43
13	0	5	2	10	10	2	0	0	0	0	0	8	0	0	10	0.43
14	0	5	2	10	10	2	0	0	0	0	0	8	0	0	10	0.43
15	1	6	2	5	10	4	6	0	0	0	0	4	10	0	10	0.51
16	1	6	2	5	10	4	6	0	0	0	0	4	10	0	10	0.51
17	2	2	8	1	0	2	8	0	0	0	13	0	0	0	0	0.41
18	1	4	2	0	3	5	4	0	0	0	0	2	3	0	1	0.45
19	1	4	2	0	3	5	4	0	0	0	0	2	3	0	1	0.45
20	0	3	2	3	3	2	0	0	0	0	0	4	0	0	1	0.39
21	0	3	2	3	3	2	0	0	0	0	0	4	0	0	1	0.39
22	2	2	7	1	0	2	7	0	0	0	9	0	0	0	0	0.41

4　实验

本文的生物数据来自 AGEMAP 数据库(http://cmgm.stanford.edu/~kimlab/aging_mouse/),其中包括 C57BL6 小鼠 16 个组织在 1、6、16、24 个月的 8932 个基因表达数据,由于其中 4 个组织的数据存在较大噪声和不完整性,我们使用 Adrenal Glands(1)、Cerebellum(2)、Cerebrum(3)、Eye(4)、Gonads(5)、Heart(6)、Hippocampus(7)、kidney(8)、Lung(9)、Muscle(10)、Spinal Cord(11)、Thymus(12)等 12 个组织的数据.共表达网络的构建如下:首先计算基因间的 Pearson 相关系数 r,然后将其转化为另一变量 $r' = \sqrt{(n-2)r^2/(1-r^2)}$($n$ 表示表达谱数据个数).r'是服从自由度为 $n-2$ 的 t 分布.当 r'大于设定的 p-value 对应的 t 分布表的值时,两个基因之间就加一条边.本文除 Eye 组织的 p-value 取值为 1E-04,Spinal Cord 组织为 1E-07、其余均取 5E-06.

David 数据库可以对小鼠基因簇的功能进行分析,如果一个基因簇有 50%以上的基因显著的共享一个或多个 GO 项,则认为其是一个显著富集簇.本文主要是根据 Southworth 等验证的 401 条与小鼠衰老相关的 GO 项判断一个簇是否与衰老有关,GO 项的 P-value 取 0.05.通常认为 16 至 24 个月为小鼠的衰老期,下面我们通过二种方法挖掘这两个月的差异基因簇,并进行功能分析.

4.1　模块内基因变化分析

在实际的生物网络中,模块内基因的度都相对较大,并且几乎都发生了较大的变化.我们首先标记($D'_{vv'} - D'_{v',v}$)大于等于 5 或小于等于 -5 的基因分别为 16 个月 24 个月发生变化的基因;然后利用 L^1_{uv} 进行 K-means 聚类,分别得到 16 或 24 个月的差异聚类簇;最后,对每个聚类簇的 GO 功能进行分析.结果如图 4 和图 5,可以看出大部分基因簇都显著富集与一个或多个 GO 项并且与衰老相关.

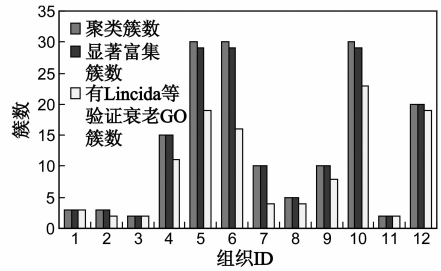

图4　小鼠12组织16个月模块内基因变化簇统计结果

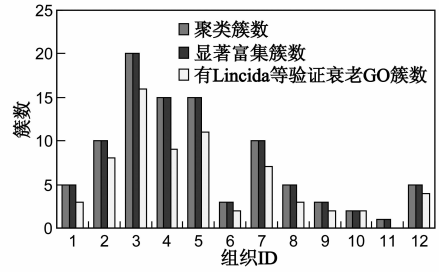

图5　小鼠12组织24个月模块内基因变化簇统计结果

4.2 模块间基因变化分析

Tijana 等认为节点差异度 $D_{vv'}$ 大于 0.4 就表示节点的两个图元向量差别明显[14].因此,我们首先提取节点差异度 $D_{vv'}$ 大于等于 0.4 的基因;然后,根据差异度 $D_{vv'}$ 的分布,利用 $L^2_{uu'}$ 进行 K-means 聚类;对个聚类簇的 GO 功能分析如图 6.与前面的结果类似,可以看出大部分基因簇都与衰老相关.

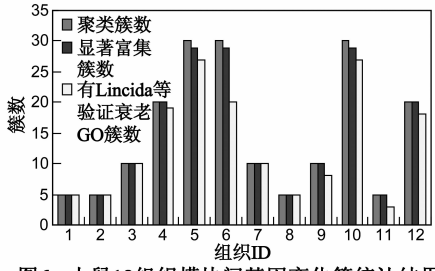

图6 小鼠12组织模块间基因变化簇统计结果

通过 GO 分析,我们发现差异簇主要显著富集以下 GO 项: GO: 0005488、GO: 0005515、GO: 0005622、GO: 0005623、GO: 0005737、GO: 0008152、GO: 0009987、GO: 0043226、GO: 0043227、GO: 0043229、GO: 0043231、GO: 0044424、GO:0044464、GO:0044237、GO:0044238 等.说明这些功能的差异与衰老具有密切的关系.尽管仍然有些基因簇按照 Southworth 验证的 401 条衰老 GO 项与衰老无关,这可能与目前有关衰老相关的 GO 项还不完善有关,如有些文献还发现了其它一些与小鼠衰老有关的 GO 项[16,17].

此外, David 数据库的注释与 AGEMAP 数据库往往存在不一致.如 Thymus 组织模块内的一个基因簇(其中包括 196 个基因),在 David 数据库中注释衰老 GO: 0005488~ binding 条目的富集度为 52.6%,显著性为 0.027.此簇中的 Mm.20935 基因不含 GO: 0005488 ~ binding 条目功能,但在 AGEMAP 提供的数据中则显示此基因具有 GO:0005488 ~ binding 功能.另外,Thymus 组织模块间基因变化簇(其中包括 356 个基因),在 David 数据库中注释衰老 GO:0008152 ~ metabolic process 条目的富集度为 41.2%,显著性为 8.37 – 05.显示此簇中的 Mm.103728 基因不具 0008152 ~ metabolic process 条目功能,同样 AGEMAP 数据库显示该基因具有 0008152 ~ metabolic proces 功能.因此,那些与衰老无关的基因簇可能还有待进一步实验验证.

4.3 二种方法的结果比较

下面我们讨论二种基因簇的连接关系,内连接是指模块内基因相互之间作用的边数量,总连接是指模块内基因之间边数量与模块间边数量的总和.表 3 给出了 *Lung* 组织中的两个簇 *A* 和 *B* 的连接情况.第一种方法得到的簇 *A* 共包含 236 个基因,在 24 个月时簇内有 1377 条边,基因间的连接非常紧密;而在 16 个月时簇内基因连接却只有 20 条,是典型的模块内连接变化;用第二种方法得到的簇 *B* 共包含 87 个基因.在 16 个月时簇内基因连接为 10 条,总链接为 193 条;24 个月时簇内基因连接为 7 条,总链接为 131 条,可见主要变化是簇间的连接.

表 3 小鼠 Lung 组织两种基因簇连接的比较

	簇 A(236 个基因)		簇 B(87 个基因)	
时 间	16 个月	24 个月	16 个月	24 个月
内连接	20	1377	10	7
总连接	372	6349	193	131

5 结论

差异模块的挖掘对于认识生命现象的进化、衰老及疾病的产生等生物问题具有重要的意义.由于差异簇内部各基因之间关系的变化具有相似性,本文利用图元向量这一结构,提出了二种度量网络节点变化差距的方法,并将其应用于差异模块的挖掘.在 AGEMAP 数据库中 12 个小鼠组织的基因表达数据的分析结果表明,本文提出的二种方法可以分别挖掘模块内部变化及模块间变化基因簇.并且在 DAVID 数据库的 GO 分析结果表明,挖掘的模块都显著的富集于一些 GO 条目,而且其中大部分都与衰老有关.

图元向量作为刻画网络局部拓扑结构的一种参数,在差异分析中应该有广阔的应用前景.但是,图元向量的计算复杂度将随图元规模的增大急剧增加,如何克服这一局限将有待进一步研究.此外,由于差异模式产生的原因及机制各种各样,深入细致的分析生物网络中的差异模式的类型,以及并确定相应的识别方法将是本文后续的工作.

参考文献

[1] 覃桂敏,高琳,周晓锋.一种基于统计的生物网络模体发现算法[J].电子学报,2009,37(11):2420 – 2426.
QINGui-min, GAO Lin, ZHOU Xiao-feng. Non-Treelike Network Motif Detection Algorithm[J]. Acta Electronica Sinica, 2009,37(11):2420 – 2426.

[2] 赵建邦,董安国,高琳.一种用于生物网络数据的频繁模式挖掘算法[J].电子学报,2010,38(8):1803 – 1807.
ZHAOJian-bang, DONG An-guo, GAO Lin. An Algorithm for Frequent Pattern Mining in Biological Networks[J]. Acta Electronica Sinica, 2010,38(8):1803 – 1807.

[3] Jacob MZ, Suresh P. AGEMAP: A Gene Expression Database for Aging in Mice[J]. PLOS Genetics, 2007, 3(11):2326 – 2337.

[4] Remondini D, Connell BO, et al. Targeting c-Myc-activated genes with a correlation method: Detection of global changes in

large gene expression network dynamics[J]. Proc Natl Acad Sci USA,2005,102(19):6902 – 6906.
[5] Voy BH.,Scharff J A,et al. Extracting gene networks for low-dose radiation using graph theoretical algorithms[J]. PLoS Comput Biol,2006,2(7):757 – 768.
[6] Oldham MC, Horvath S, et al. Conservation and evolution of gene coexpression networks in human and chimpanzee brains [J]. Proceedings of the National Academy of Sciences, 2006, 103(47):17973 – 17978.
[7] Zhang B, Li H, et al. Differential dependency network analysis to identify condition-specific topological changes in biological networks[J]. Bioinformatics, 2009 25(4):526 – 532.
[8] Southworth LK, Owen AB, Ki SK. Aging Mice Show a Decreasing Correlation of Gene Expression within Genetic Modules[J]. PLOS Genetics, 2009, 5(12):2264.
[9] Tesson B, Breitling R, et al. Diffcoex: a simple and sensitive method to find differentially coexpressed gene modules[J]. BMC Bioinformatics, 2010, 11(1):497.
[10] Fang G, Pandey G, et al. Mining Low-Support Discriminative Patterns from Dense and High-Dimensional Data. Knowledge and Data Engineering[J]. IEEE Transactions, 2012, 24(2), 279 – 294.
[11] Fang G, Kuang R, et al. Subspace differential coexpression analysis: Problem Definition and a General Approach[C]. Pacific Symposium on Biocomputing, 2010, 15:145 – 156.
[12] Przulj N. Biological network comparison using graphlet degree distribution[J]. Bioinformatics, 2007, 23(2):177 – 183.
[13] Przulj N, Derek G, et al. Modeling Interactome: Scale-Free or Geometric[J]. Bioinformatics, 2004, 20(18):3508 – 3515.
[14] Milenkovic T, Przulj N. Uncovering Biological Network Function via Graphlet Degree Signatures[C]. Cancer Inform, 2008, 6:257 – 273.
[15] Kuchaiev O, Milenkovic T, et al. Topological network alignment uncovers biological function and phylogeny[J]. Journal of the Royal Society Inferface, 2011, 5(17):1341 – 1354.
[16] Fu CX, Hickey M, et al. Tissue specific and non-specific changes in gene expression by aging and by early stage CR [J]. Mechanisms of Ageing and Development, 2006, 127 (12), 875 – 936.
[17] Jamie L, Barger, et al. A Low Dose of Dietary Resveratrol Partially Mimics Caloric Restriction and Retards Aging Parameters in Mice[J]. PLoS ONE, 2008, 3(6):2264.

作者简介

肖碧玉　女,研究生,1987 年出生于湖南衡阳,主要研究方向为生物信息学、数据挖掘.

李先斌　男,研究生,1988 年出生于江西德安,主要研究方向为生物信息学、数据挖掘.

刘文斌(通信作者)　男,教授,博士,1969 年出生于陕西韩城.2004 年获华中科技大学博士学位,目前感兴趣的研究领域为生物信息学、数据挖掘、DNA 计算等.获得省部级奖励 4 项,主持国家省部级项目 6 项,发表学术论文 40 余篇.

E-mail:wbliu6910@126.com