论著

文章编号:1000-5404(2006)13-1370-04

基于功能模块组织癌细胞系基因表达谱的关联规则

王晨光[1],郭 政[2],李 霞[2],王 健[2] (哈尔滨医科大学:[1]医用物理学教研室,[2]生物信息系,哈尔滨 150086)

提 要:目的 提出一种在功能模块化水平上组织大量基因表达谱关联规则的新方法。**方法** 以高频出现在关联规则后件中的基因为基础分类关联规则,再选择显著聚集前件基因的GO基因功能类作为功能模块关联规则的前件,获得功能模块表达关联规则。**结果** 根据NCI60数据,发现了一些与癌症相关的基因及相应的基因功能模块表达关联规则。**结论** 利用功能模块表达关联规则可以有效地挖掘疾病基因表达谱。

关键词:基因芯片;关联规则;gene ontology;癌症

中图法分类号:R318.04;R394 **文献标识码**:B

Functional modules based organization for association rules of gene expression profiles in cancer cell lines

WANG Chen-guang[1], GUO Zheng[2], LI Xia[2], WANG Jian[2] ([1]Department of Medical Physics, [2]Department of Bioinformatics, Harbin Medical University, Harbin 150086, China)

Abstract: Objective To apply apriori algorithm to mine the gene expression profiles will produce a huge number of association rules. A new approach to organize the mined association rules on the modularized gene functional category level is proposed. **Methods** The association rules containing the genes frequently appearing in the head parts of all the association rules were selected and grouped together. The gene functional categories, which were non-randomly enriched with the genes in the body parts of each grouped association rules, were found with statistical test and were used to form the body parts of the functional modular level association rules. **Results** By applying this new method to NCI60 data set, some cancer associated genes and their related functional category level expression association rules were discovered. **Conclusion** Gene expression profiles of diseases can be efficiently mined by association rules on the modularized gene functional category level.

Key words: gene chip; association rules; gene ontology; cancer

采用基因芯片技术,可高通量地并行检测成千上万个基因的表达水平,分析癌症等疾病的发病机制或药物的作用机制等。利用关联规则分析基因表达谱数据,可以发现一个基因是如何影响另外一个基因的表达,及在特定的细胞条件下表达的基因[1,2]。

由于基因表达谱涉及的基因数目很大,利用关联规则挖掘基因表达谱数据会产生大量的冗余与无关规则。常见的处理方法可分为前处理和后处理两种。前处理是一种基于约束的方法,即加入一些生物学先验的约束,筛选掉一些基因,把剩余的少数基因作为输入,以减少规则的产生数量。例如Berrar等[1]从NCI60数据中1 500多个基因中筛选出20个基因作为输入,再从生成的规则中选取一部分规则构造分类器,从而进行药物筛选。显然,如果采用过严的约束,这类方法将导致被筛选掉的基因中蕴涵的信息的丢失。后处理的一类方法是通过χ^2检验,两项分布检验等统计学方法检查规则前件与后件的相关性,筛选掉一些前后件不显著相关的规则。另外一类有意义的后处理方法是根据限定的生物学知识筛选已经挖掘到的关联规

基金项目:国家自然科学基金资助项目(30370798,30170515,30370388);国家高技术研究发展计划("863")资助项目(2003AA2Z2051;2002AA2Z2052);黑龙江科技攻关资助重点项目(GB03C602-4);哈尔滨市科技攻关资助项目(2003AA3CS113);黑龙江自然科学基金资助项目(F0177)

Supported by the National Natural Science Foundation of China (30370798, 30170515, 30370388), the National "863" Key Research Project of China (2003AA2Z2051, 2002AA2Z2052), the Foundation for the Key Sci & Tech Research Project of Heilongjiang Province (GB03C602-4), the Foundation for the Key Sci & Tech Research Project of Harbin (2003AA3CS113) and the Natural Science Foundaion of Heilongjiang Province (F0177)

作者简介:王晨光(1966-),男,黑龙江省哈尔滨市人,副教授,主要从事生物信息学方面的研究,电话:(0451)86674918

通讯作者:郭 政,电话:(0451)86615933

收稿日期:2005-06-16;**修回日期**:2005-12-30

则。例如，Alexander 等[2]提出了一种类似查询的策略，使生物学家可以通过语法词汇的组合选择自己感兴趣的规则。然而这是一种被动的策略，它过度地依赖生物学家的先验知识。

基因具有系统性表达与模块化发挥功能的性质，功能相关的基因倾向于一致表达，这为我们从基因功能模块层次分析基因之间的内在联系及这些基因与疾病可能的关联等问题提供了基础。据此，我们以基因功能分类体系 gene ontology(GO)[3]为指导，对应用关联规则 Apriori 算法[4]于表达谱数据产生的大量规则，从基因功能模块化的角度进行筛选与再组织，进而挖掘出基因功能类层次上的表达关联信息。通过对 NCI60 数据的分析，发现了一些与癌症病变机制密切相关的信息。

1　材料与方法

1.1　基因表达谱数据

本次实验采用 NCI60 数据(http://genome-www.stanford.edu/nci60)。Ross 等[5]采用 cDNA 微阵列技术检测了 9 类、60 个人类癌细胞系的 8 000 多个基因的表达水平，以细胞系混合物为共同的对照，以两者基因表达水平的比值反映基因表达的差异情况。其中 60 余个癌细胞包括中枢神经瘤 6 例、肾癌 8 例、卵巢癌 6 例、乳腺癌 7 例、白血病 8 例、结肠癌 7 例、黑色素瘤 8 例、肺癌 9 例与前列腺癌 2 例。

我们沿用 Ross 等[5]采用的方法，以 *Ratio* 值表示基因在癌细胞系中的表达水平与该基因在对照组中的表达水平的相对比，然后选择那些在所有样本细胞系中至少有 4 个 $|Ration|\geq 7$ [即 $|log_2(Ration)|\geq 2.8$] 的基因作为差异表达基因，共得到 1 161个差异基因，它们是在 60 个人类癌细胞系中总体表达变异较大的基因。

1.2　方法

1.2.1　GO 基因功能注释体系　　GO[3]是使用有控制的词汇表和严格定义的概念关系，以有向无环图方式表示的跨越原核生物与真核生物各物种的基因功能分类体系。GO 将基因功能划分为分子功能(molecular function)、生物过程(biological process)与细胞组成(cellular component)3 个维度，纠正了传统功能分类体系中常见的维度混淆问题。

由基因的 IMAGE 克隆号或 GenBank 编号，我们将基因注释到它能被注释的最具体的 GO 功能类中，此步称为“狭义注释”。由于 GO 上同一路径上的父子概念的内涵有包含关系，我们将狭义注释推广为广义注释：基因被狭义注释到 GO 上某个结点，则被广义注释到该结点到根结点(Gene Ontology)路径上的所有结点。本次实验采用了 GO 的 2004.4 的版本。

1.2.2　基因表达水平离散化与关联规则的生成　　以每个细胞系作为一次交易(transaction)，其中 1 161 个差异表达基因对应于传统关联规则挖掘中的商品项(item)。对 1 161 个差异表达基因中的每个基因，我们根据其在每个细胞系中的表达水平进行离散化：当其表达值的 $|log_2(Ration)|\geq 1.5$ 时，则认为该基因在该细胞系中差异表达，取值为 1；否则，认为该基因在该细胞系中不差异表达，取值为 0。我们以行代表癌细胞系，列代表基因，从而形成 0-1 矩阵，它对应于传统关联规则挖掘中的交易矩阵。将离散化以后的矩阵作为 Apriori 关联规则挖掘算法的输入，选取最小支持(support)与最小置信(confidence)度，生成规则形如 A→B(A，B 表示基因的集合)，以发现一组基因的差异表达如何影响另外一组基因的差异表达。

1.2.3　寻找关联规则中高频出现的基因　　根据挖掘到的关联规则，我们进一步寻找关联规则的前件或后件中高频出现的基因。通过此方法找到的基因是在癌症中与多种基因都密切相关的基因，此种基因有可能在正常细胞转变成癌细胞的过程中参与一些关键功能。算法的计算机伪代码如下：

```
输入：rules，mincount
输出：前件中出现频率大于 mincount 的基因，后件中出现频率大于 mincount 的基因
while all rules do begin
  for each gene in rule do
    if gene in leftside then genename.count++ in geneL
    else genename.count++ in geneR.
end
for each gene in geneR，geneL do
if gene.count > mincount * totalrulesnumber
then output gene
```

rules 表示 Apriori 算法生成的所有关联规则；geneL 与 geneR 分别为前件与后件中出现的基因的集合。算法输出在前件或后件中出现频率大于最小阈值 mincount 的基因。

1.2.4　根据 GO 模块化组织关联规则　　以在关联规则中高频出现的各个基因为基础，我们可以对涉及高频出现的各个基因的关联规则按基因功能类进行模块化组织：①以在关联规则后件中高频出现的各个基因为基础，通过分别寻找它们的前件中的基因在 GO 中非随机聚集的相关功能类节点，形成规则：功能类→gene。②以前件中高频出现的基因为基础，分别对它们后件中的基因进行类似的模块化处理，可以形成规则：gene→功能类。

我们也可以不以在关联规则中高频出现的各个基因为基础，更一般地对全体关联规则按基因功能类进行模块化组织。例如，我们可以采用如下的模块化方法：①寻找前件中的基因在 GO 中非随机聚集的相关功能类节点，再分别寻找这些相关功能类节点的后件中的基因在 GO 中非随机聚集的相关功能类节点，从而形成规则：功能类→功能类；②寻找后件中的基因在 GO 中非随机聚集的相关功能类节点，再分别寻找这些相关功能类节点的前件中的基因在 GO 中非随机聚集的相关功能类节点，从而形成规则：功能类→功能类。本研究只限于讨论以在关联规则后件中高频出现的各个基因为基础，对生成的规则进行模块化组织，形成规则：功能类→gene。

1.3　统计学方法

对在后件中高频出现的某个基因所涉及的关联规则，我们采用超几何分布检验注释到 GO 中的每个功能单元中的前件基因是否非随机地聚集，所得显著聚集前件基因的功能单元可能是与后件中高频出现的某个基因的表达相关的功能类。假设整个数据集中共有 N 个基因在 GO 中得到注释，其中在功能单元 F 中注释有 M 个基因；共有 K 个前件基因在 GO 中得到注释，其中在功能单元 F 中有 X 个基因，则在 K 个前件基因中有 X 个基因出现在功能单元 F 内的随机概率可以通过超几何分布计算[6]：

$$P(X=x|N,M,K)=\frac{\binom{M}{x}\binom{N-M}{K-x}}{\binom{N}{K}}$$

因此，在 K 个基因中至少有 X 个基因出现在功能单元 F 内的随机概率 $P(X\geqslant x)$ 为：

$$p=1-\sum_{i=0}^{x-1}\frac{\binom{M}{i}\binom{N-M}{K-i}}{\binom{N}{K}}$$

由于通常情况下 N 足够大，因此我们使用二项分布计算相应的概率：

$$p=1-\sum_{i=0}^{x-1}\binom{K}{i}\left(\frac{M}{N}\right)^{i}\left(1-\frac{M}{N}\right)^{K-i}$$

我们选择显著聚集前件基因（$P<0.005$）的功能结点作为候选的与后件中高频出现的某个基因的表达相关的功能类。我们没有采用严格的多重统计检验修正显著性水平，以提高选择表达相关的功能类的敏感性。

2 结果

对于原始的含有 1 161 个差异表达基因的表达谱，我们选取最小支持度为 40%，最小置信度为 80%，产生了5 661条规则。我们取 mincount = 15%，选择在关联规则中的前件或者后件中高频出现的基因，结果见表 1。最小支持度与最小置信度的选择对结果会有较大的影响，作为初步研究，我们通过多次实验发现：最小支持度为 30% 时，会引入大量无用关联规则信息，计算时间会明显增加，所以我们选取最小支持度为 40%。而最小置信度取 90% 以上时，后件基本上为单一的基因，所以我们取最小置信度为 80%，以便于进一步模块化分析。

表 1 在关联规则中高频出现的基因

Cloneid	前件	后件	Gene symbol
510467	1062	1083	CDKN2A
510534	1022	1111	TACSTD1
363981	1075	850	SPP1
254006	975	——	SCN7A

基因 CDKN2A，通常称其为 p16，它的作用与 p53 相似，Lukas等[7]发现 CDKN2A 具有抑制肿瘤生成的作用。基因 TACSTD1 编码肿瘤关联的 Ca 离子信号传导蛋白，人类细胞表面抗原。基因 SPP1 编码的蛋白同属于细胞外基质蛋白，它存在于胞外，与细胞表面基膜关系密切，有促进细胞繁殖，支持细胞结构等功能。SPP1 编码骨桥蛋白（OPN），在癌症病变过程中与其他基因高度相关，有可能参与多种功能。Crawford 等[8]通过诱导敲除 OPN 的小鼠产生鳞状上皮细胞癌，发现普通细胞来源的 OPN 起到巨噬细胞诱导物的作用，而肿瘤细胞产生的 OPN 能够起到抑制巨噬细胞吞噬，促进肿瘤细胞生长和转移的作用。Liaw 等[9]提出 OPN 是一个多功能蛋白，它的过表达与癌症的形变和转移密切相关。Brown 等[10]发现 OPN 在人类的多种癌症中都大量表达。基因 SCN7A 编码 Na 离子通道蛋白，只在前件中高频出现，它在癌症的病变过程中可能起到一种辅助作用。离子通道相关基因被证实在多种癌症中表达，但其作用机制尚不清楚。

在以上 4 个基因中，SCN7A 只在前件中高频出现，CDKN2A、TACSTD1、SPP1 在前件与后件中都高频出现（在全部规则的前件与后件中出现的频率都不小于 15%）。我们从后件中包含高频出现的基因 CDKN2A、TACSTD1 或 SPP1 的 3044 条关联规则中，通过分别寻找它们的前件中的基因在 GO 中非随机聚集的相关功能类节点（$P<0.005$），从而形成 16 条功能类层次的表达关联规则：功能类→基因（表 2）。这些显著聚集前件基因的功能类可能是与后件中高频出现的基因的表达相关的功能类，提示在这些功能类中差异表达的基因有较大的可能性会伴随后件中相应基因的差异表达。

表 2 功能类层次的表达关联规则

功能类	基因
[GO:0046849, bone remodeling, biological_process],	
[GO:0007275, development, biological_process],	
[GO:0001503, ossification, biological_process],	
[GO:0005886, plasma membrane, cellular_component],	
[GO:0005905, coated pit, cellular_component],	
[GO:0016020, membrane, cellular_component]	CDKN2A
[GO:0007283, spermatogenesis, biological_process],	
[GO:0007276, gametogenesis, biological_process],	
[GO:0016477, cell migration, biological_process],	
[GO:0048232, male gamete generation, biological_process],	
[GO:0000003, reproduction, biological_process],	
[GO:0019953, sexual reproduction, biological_process],	TACSTD1
[GO:0005543, phospholipid binding, molecular_function],	
[GO:0005509, calcium ion binding, molecular_function],	
[GO:0005901, caveola, cellular_component],	
[GO:0005578, extracellular matrix, cellular_component],	SPP1

3 讨论

关联规则挖掘是分析基因表达谱的一种有用的工具，可以发现一个基因是如何影响另外一个基因的表达，及在特定的细胞条件下表达的基因[9]。针对关联规则挖掘中产生大量规则的问题，我们以高频出现在关联规则后件的基因为基础，将前件映射到 gene ontology 上，从基因功能类角度对大量规则进行组织，进一步从大量规则中抽取那些涉及基因功能类的方向性规则。根据这些方向性的规则，可以从不同功能层次水平透视基因表达之间的关系，进一步考察其中涉及到的基因与基因间的关联规则，分析我们感兴趣的细节。如我们在方法中所述，还可以更一般地对全体关联规则按基因功能类进行模块化组织。我们有必要进一步分析不同的对基因表达关联规则进行功能模块化组织的合理途径与意义。另外一个可以改进的方向是进一步将差异表达基因为上调与下调表达，以挖掘更细致的表达关系。

基因表达的高变异与基因芯片技术的检测误差对基因表达谱分析结果的可靠性与稳定性有很大的影响。另外，差异表达基因的选择标准、基因表达值的离散化方式及数据挖掘方法的参数设置（如关联规则中的支持度与置信度等）都会对基因表达谱的分析结果

产生影响。所以，基因表达谱数据挖掘分析的一个关键问题是寻求稳健的（抗基因表达变异、检测误差及差异表达基因的选择标准等）分析结果。由于本研究提出的基因表达谱关联规则的功能模块化组织方法的分析结果来自于对涉及大量基因的关联信息的综合，对单个基因的表达值及预先选入的单个关联规则的依赖性较小，因此对各种随机因素的影响具有一定的稳健性。我们将进一步研究这种功能模块化组织方法的稳健特性。

参考文献：

[1] BERRAR D, DUBITZKY W, GRANZOW M, *et al*. Analysis of Gene Expression and Drug Activity Data by Knowledge-based Association Mining[C]//Proceedings of Critical Assessment of Microarray Data Analysis Techniques(CAMDA'01), 2001: 25 - 28.

[2] ALEXANDER T, GEDIMINAS A. Handling very large numbers of association rules in the analysis of microarray data[C]//KDD. New York: ACM, 2002: 396 - 404.

[3] ASHBURNER M, BALL C A, BLAKE J A, *et al*. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium[J]. Nat Genet, 2000, 25(1): 25 - 29.

[4] AGRAWAL R, SRIKANT R. Fast algorithms for mining association rules in Large Databases [C]//Proceedings of the 20th International conference on Very Large Databases (VLDB). Santiago: Morgan Kaufmann Publishers, 1994: 487 - 499.

[5] ROSS D T, SCHERF U, EISEN M B, *et al*. Systematic variation in gene expression patterns in human cancer cell lines[J]. Nat Genet, 2000, 24(3): 227 - 235.

[6] KHATRI P, DRAGHICI S, OSTERMEIER G C, *et al*. Profiling gene expression using onto-express[J]. Genomics, 2002, 79(2): 266 - 270.

[7] LUKAS J, PARRY, D, AAGAARD L, *et al*. Retinoblastoma-protein-dependent cell-cycle inhibition by the tumour suppressor p16[J]. Nature, 1995,375(6531): 503 - 506.

[8] CRAWFORD H C, MATRISIAN L M, LIAW L. Distinct roles of osteopontin in host defense activity and tumor survival during squamous cell carcinoma progression *in vivo*[J]. Cancer Res, 1998, 58(22): 5206 - 5215.

[9] LIAW L, CRAWFORD H C. Functions of the extracellular matrix and matrix degrading proteases during tumor progression [J]. Braz J Med Biol Res, 1999, 32(7): 805 - 812.

[10] BROWN L F, PAPADOPOULOS-SERGIOU A, BERSE B, *et al*. Osteopontin expression and distribution in human carcinomas[J]. Am J Pathol, 1994, 145(3): 610 - 623.

（编辑　栾　嘉）

（上接 1366 页）

制，如神经激素反馈调节或血管床本身内在的耐受机制。传统观念认为硝酸甘油耐药与：①磷酸二酯酶活性增高使 cGMP (cyclic guanosine monophosphate 环乌苷酸)降解增多；②细胞内巯基耗竭；③硝酸甘油的转化受抑制等机制有关[2]。但最新的大量研究表明，硝酸甘油连续治疗过程中过氧化物产生明显增加可能是形成硝酸甘油耐受性更重要的机制[1]。有机硝酸酯类是通过在血管内皮细胞和平滑肌细胞内的谷胱甘肽转移酶的催化下释放出 NO，NO 进而增加 cGMP 导致血管扩张，从而产生降压和治疗心绞痛的作用。但这种作用随着治疗时间的延长（尤其是连续给药）而逐渐减弱。Munzel 等[1] 在对硝酸甘油耐受性的研究中发现，血管局部氧自由基的产生明显增加，氧自由基可直接与 NO 结合使其失活，也可通过间接抑制硝酸甘油的转化减少 NO 的产生，而抑制或减少 cGMP 生成而发生耐受。硝酸甘油治疗中也伴随有神经内分泌的激活，如肾素-血管紧张素-醛固酮系统的激活和儿茶酚胺的增加，造成血管收缩力增强，抵消了硝酸甘油的血管舒张[3]。而更为重要的是，这些神经内分泌因子，尤其是血管紧张素Ⅱ，通过激活膜相关还原型烟酰胺腺嘌呤二核苷酸氧化酶增加血管过氧化物的产生，使硝酸甘油的血管扩张作用受抑制，并且随药物使用时间的延长，这种作用也更加突出[3,4]。以上研究表明氧自由基产生增加和激活是长时间使用硝酸甘油耐受性形成的重要原因。

还原型谷胱甘肽是在人类细胞质中合成的一种肽，它通过激活多种酶（如巯基酶），从而促进糖、蛋白质、脂肪等物质的代谢，而更为重要的功能是通过巯基与氧自由基的结合，使其转化为容易代谢的酸类物质而加速氧自由基的排泄，是保护细胞稳定及维持细胞生理功能的重要物质[5]。本研究就是利用还原型谷胱甘肽的抗氧化作用，使硝酸甘油使用过程中产生的氧自由基被得以及时清除，从而达到抑制硝酸甘油耐受性的目的。本研究的最终结果也证实了这一点。

本研究表明抗氧化剂还原型谷胱甘肽对硝酸甘油耐受性有抑制作用，从而为临床解决硝酸甘油的耐受性问题提供了一条新的途径，但其机制仅是综合相关文献后作出的一种理论推测，尚有待进一步的研究。

关键词：谷胱甘肽；硝酸甘油；耐受性
中图法分类号：R969.2　　文献标识码：B

参考文献：

[1] MUNZEL T, SAYEGH H, FREEMAN B A, *et al*. Evidence for enhanced vascular superoxide anion production in nitrate tolerance. A novel mechanism underlying tolerance and cross-tolerance[J]. J Clin Invest, 1995, 95(1): 187 - 194.

[2] KURZ S, HINK U, NICKENIG G, *et al*. Evidence for a causal role of the renin-angiotensin system in nitrate tolerance [J]. Circulation, 1999, 99(24): 3181 - 3187.

[3] HIRAI N, KAWANO H, YASUE H, *et al*. Attenuation of nitrate tolerance and oxidative stress by an angiotensin II receptor blocker in patients with coronary spastic angina[J]. Circulation, 2003, 108(12): 1446 - 1450.

[4] TAYLOR A L, ZIESCHE S, YANCY C, *et al*. Combination of isosorbide dinitrate and hydralazine in blacks with heart failure[J]. N Engl J Med, 2004, 351(20): 2049 - 2057.

[5] CARDOSO S M, PEREIRA C, OLIVEIRA C R. The protective effect of vitamin E, idebenone and reduced glutathione on free radical mediated injury in rat brain synaptosomes[J]. Biochem Biophys Res Commun, 1998, 246(3): 703 - 710.

（编辑　张大春）