GO 功能类与基因差异表达的关联规则挖掘算法*

屠 康[1] 喻 辉[1] 郭 政[1,2]** 李 霞[1]

（[1]哈尔滨医科大学生物信息系，哈尔滨 150086；[2]哈尔滨基太生物芯片开发有限责任公司，哈尔滨 150086）

摘要 针对基因功能分类体系 Gene Ontology 的层次结构特点，修改关联规则挖掘算法 Apriori，开发“挖掘与基因差异表达关联的 GO 功能组合”软件（RuleGO）. RuleGO 以基因表达谱上的差异表达基因集合和不差异表达基因集合为输入，输出组合特征功能类与基因差异表达现象的关联规则，有助于解释基因差异表达现象的本质原因，如疾病发病机制、药物作用机理等. 将 RuleGO 和 OntoExpress 应用在结肠癌和腺癌表达谱数据集上，结果显示，RuleGO 比 OntoExpress 能发现更多的与差异表达现象关联的特征功能类，更能看到在 OntoExpress 上不能发现的组合特征功能类. 另外，结果显示，将规则的置信度和支持度要求设置较高时，一般只有组合功能类才能满足要求，这提示在基因表达谱分析中不宜采用单个角度的单个功能分类单元，考虑功能分类单元的组合可能更有意义.

关键词 基因表达谱，差异表达，特征功能类，关联规则，Gene Ontology

学科分类号 Q617

基因芯片技术[1]是高通量检测基因 mRNA 水平表达信息的生物新技术，根据基因表达谱筛选差异表达基因[2,3]，对于区分疾病亚型、研究疾病的发病机制或药物的作用机理等有指导意义. 然而，各种差异表达基因筛选方法都存在较大的假阳性和假阴性问题[4]. 近来人们开始关注差异表达基因的功能类[5]，即注释到该功能类内的大多数基因都差异表达的功能类，这类功能类与实验条件较相关，用户对它们的感兴趣程度较高，我们称为“特征功能类”. 对特征功能类进行分析，能获得比基因更高一层次的抽象结论，对理解疾病的发病机制或药物的作用机理等更有帮助.

OntoExpress[5]是一种代表性的挖掘特征功能类的软件，用户先确定感兴趣的前景基因，如差异表达的基因、表达相似性高的基因或对分类有鉴别意义的基因等，令剩余的基因构成背景基因. OntoExpress 考察每个基因功能类中前景基因的数目与背景基因数目之比. 该比率越大，用户对该功能类的感兴趣程度越高. OntoExpress 等软件的最大问题是它们独立地分析各功能类，不能发现各功能类组合的交互效应[6].

在本文中我们利用关联规则挖掘的思想，针对基因功能分类体系 Gene Ontology[7]的层次结构特点开发了“GO 功能类与基因差异表达的关联规则挖掘算法”（mining association rules of GO function classes and gene expression difference，RuleGO），旨在发掘与基因差异表达现象关联的单个特征基因功能类或多个特征功能类的组合. 算法输出一组形如“基因被注释到某个基因功能类（或某几个基因功能类），则该基因在不同的疾病类型中极可能差异表达”的关联规则. 这些规则能提示，在所研究的实验条件下整体有差异表达倾向的功能类或功能类组合，对理解疾病发病机制、指导药物开发有启示作用. 我们将 RuleGO 分别应用在结肠癌和腺癌数据集上，得到了很有启发性的结果.

1 RuleGO 前处理

1.1 确定阳性基因和阴性基因

用户根据具体情况和特定要求，选用合适的方法，确定差异表达的基因，构成阳性基因集合；确定不差异表达的基因，构成阴性基因集合. 注意，一个基因可能既不是差异表达基因，也不是不差异表达基因，因为本文的“不差异表达基因”是表达差异十分微小、在不同实验条件下表达值高度一致的基因. 处于差异表达和不差异表达两个极端之

* 国家自然科学基金资助项目（39970397，30170515），国家高技术“863”计划资助项目（2002AA222052，2003AA2Z2051），黑龙江科技攻关重点项目（GB03C602-4），哈尔滨市科技攻关项目（2003AA3CS113），黑龙江自然科学基金资助项目（F0177）和 211 工程“十五”建设项目.

** 通讯联系人.

Tel：0451-86669617，Fax：0451-86615933

E-mail：markgz@0451.vip.com

收稿日期：2004-01-17，接受日期：2004-03-30

间的基因（本文称之为过渡态基因）不包括在分析中.

目前判断基因是否差异表达的方法大致可以分为4类：a. t-统计量推断，这是最简单，最常用的差异基因判断方法，它基于两点假设：基因在同类实验条件下的表达值观测数据服从正态分布；不同基因的上述正态分布中的方差相等. b. 回归分析，它不基于上述两点假设，但要求有一定样本含量. c. 非参数检验，构造合适的统计量 Z_i，通过随机重排（permutation）构造 Z_i 的经验分布，进而确定基因表达属于经验分布（不差异分布）的 p 值；d. 方差分析（ANOVA），适用于多种因素多水平交互作用的情况. 4类方法各有利弊，适用范围也各不一样. 用户应当根据数据质量、样本数量、实验影响因素等实际情况选择合适的差异表达筛选方法，确定差异表达基因和不差异表达基因，并将它们作为 RuleGO 的输入.

本实验中采用了 t-统计量推断方法. 我们用“类别”属性（gene type）表示基因所隶属的基因类，阳性基因的类别值为1，阴性基因的类别值为0.

1.2 背景知识体系：Gene Ontology（GO）

Gene Ontology（GO）[7]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[8]问题. GO 中一个结点又称为一个概念，它表示一个功能类.

1.3 基因功能注释

由基因的 IMAGE 克隆号[9]或 GenBank 编号[10]，经过 GenBank[10]、Unigene[11]、LocusLink[12]3个数据库，将基因注释到它能被注释的最具体的功能类中，此步称为“狭义注释”. 由于 GO 上同一路径的父子概念的内涵有包含关系，仿照 Havidsten[13]，将狭义注释进行推广（称为广义注释）：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.

生物学家希望挖掘出 GO 概念来解释不同实验条件下差异表达的本质原因，过于抽象的概念（如 biological process，cell，metabolism 等）意义不大，因而我们让用户输入一个层数阈值（l_{cut}），基因对 l_{cut} 层以上结点的广义注释被视作无效，从而过于抽象的功能概念不出现在结果规则中.

2 RuleGO：基于 GO 修改 Apriori 算法

我们利用关联规则的基本思想来挖掘与基因差异表达现象关联的特征功能类或其组合（表1）.

Table 1 The mapping of objects in RuleGO to objects in traditional association rule mining task

Objects in traditional association rule mining task	Objects in RuleGO
Transaction	genes
Item	GO terms and the gene positive type
Universe，U	The union of Go terms and gene classes
The item appears in the transaction	The gene is annotated to the GO term，or the gene type label is 1
The item doesn't appear in the transaction	The gene is not annotated to the GO term，or the gene type label is 0

类似传统关联规则挖掘中的做法，我们把基因的广义注释信息和基因的类别信息组织成基因注释矩阵（GenAnno），它对应于传统关联规则挖掘中的交易矩阵. GenAnno 的每行表示一个基因，各列为 GO 概念或基因的类别，矩阵中的元素只能取1或0，分别表示基因是否广义注释到某个 GO 概念，或基因是否是阳性基因.

基因注释矩阵（GenAnno）是标准 Apriori 算法[14]的输入格式，标准 Apriori 是关联规则挖掘的代表性算法，它搜索所有的频繁项集（交易数超过频数阈值 $min\_\ occ$ 的各种商品项组合），$min\_\ occ$ 由交易总数 n 与支持度阈值 $min\_\ sup$ 决定（$min\_\ occ = n \times min\_\ sup$）.

Apriori 算法利用了频繁项集的 Apriori 性质：频繁项集的所有非空子集必然也是频繁的，不频繁项集并上别的项形成的新项集一定是不频繁的.

$$\left.\begin{array}{l} ItemSet \subseteq U,\ Item \subseteq U, \\ Occ(ItemSet) < min\_\ sup \end{array}\right\} \Rightarrow$$

$$Occ(ItemSet \cup Item) < min\_sup$$

注：*Occ*（*ItemSet*）表示项集 *ItemSet* 在交易矩阵中出现的频数

Apriori 算法逐层（level，项集包含的项的个数）搜索频繁项集，在搜索过程中利用 Apriori 性质，如果一个 k 层项集至少有一个子集不包含在已经确定的 $k-1$ 层频繁项集池中（the pool of（$k-1$）th-level frequent Itemsets），则该 k 层项集不是频繁集，无须再进一步扫描交易矩阵考察其频繁性. 反之，如果该项集的所有子集都被包括在 $k-1$ 层频繁项集池中，则将之加入到 k 层备选频繁项集池中（pool of candidate kth level frequent Itemsets）中. 然后扫描交易矩阵，根据置信度阈值 *min_ conf* 确定 k 层频繁项集池，这些 k 层频繁项集将有助于预过滤 $k+1$ 层的不频繁项集.

将关联规则挖掘思想运用在基因表达谱分析上，要求我们对标准 Apriori 算法进行一些修改. 我们期望发现的规则形式为“基因功能概念组合⇒基因差异表达”. 基因的类别是一个特殊的项，它只能出现在规则的后件中，所以我们搜索频繁项集的空间不包括基因的表达类别属性. 另外，我们关心哪些概念与基因的差异表达现象有关联，不关心哪些概念与基因不差异表达现象关联，因而我们只在类别值为 1（差异表达）的基因中搜索频繁项集.

标准 Apriori 算法不考虑各商品项之间的关系，而在我们的应用中，各 GO 概念之间互有关系，被组织在 GO 的有向无环图中，各项之间的关系必须被考虑. 另外，用户对挖掘出的功能关联规则有一些特别要求，如不希望看到过于抽象的概念，不希望得到概念相互包含的规则等. 我们结合 GO 的结构特点和具体的应用要求，进一步修改了 Apriori 算法，一方面减少了功能关联规则之间的冗余性，另一方面，也使算法速度显著提高（成百上千倍），使 RuleGO 能以网页的形式在几分钟内向用户返回结果.

2.1 适应 Ontology 结构的动态频数阈值技术

各 GO 概念构成一个分类体系（ontology），在 ontology 上位于同一路径上的概念在内涵上相互包含，位于高层的概念较抽象，位于低层的概念较具体. 对基因进行广义注释时，抽象概念比具体概念更容易被注释到，即在基因注释矩阵中，抽象概念的出现概率先验地大于具体概念的出现概率，用固定的频数阈值（static *min_ occ*）来统一评判不同抽象层次的功能概念则对具体概念不公平. RuleGO 的频数阈值随着 ontology 层数加深逐渐减小，这是一种适应 ontology 结构的动态频数阈值技术（dynamic *min_ occ*）.

我们让用户确定 ontology 上 l_{cut} 层概念的频数阈值 *min_ occ_* l_{cut}，则 ontology 上第 l 层上的单概念的频数阈值 *min_ occ_* l 与 *min_ occ_* l_{cut} 满足如下关系：

$$\frac{min\_occ\_l}{min\_occ\_l_{cut}} = \frac{l\text{层上平均每个概念注释的基因数目}}{l_{cut}\text{层上平均每个概念注释的基因数目}}$$

从上式可得到 l_{cut} 层下任意单个功能概念（1-项集）的最低频数要求. 当考察包含 k 个概念的 k-项集 ItemSet 是否频繁时，我们需要对该具体的 k-项集设定频数阈值. 设某 k-项集 ItemSet =（C_1，C_2，…C_k），其中 k 个概念在 ontology 上所处的层数依次是（l_1，l_2，…l_k），则该 ItemSet 的频数阈值 $min\_occ(ItemSet) = \underset{i=1,2,\cdots,k}{\text{Max}}(min\_occ\_l_i)$.

通过上述方法确定的 k-项集的频数阈值是 k 个概念中频数阈值最大概念的频数阈值，这是一种严谨保守的做法，且保证了 RuleGO 算法的反单调性[14]，使我们可以沿袭标准 Apriori 按层（项集中包含的项的个数）搜索频繁项集，大大减少了搜索时间.

2.2 同一规则中不会同时出现祖孙概念

在有向无环图 GO 上，两个概念如果处在同一条从根结点到叶结点的路径上，我们称它们相互构成祖孙关系. 为使功能关联规则精炼，如果一个规则中同时出现有祖孙关系的概念，我们保留子孙概念，舍弃祖先概念. 这是因为，人们一般想用具体的功能过程、精细的细胞定位来解释生物现象，不希望得到的特征功能类过于广泛.

2.3 子规则与父规则的取舍

如果两个规则覆盖的基因完全相同，且它们分别包含相互构成祖孙关系的两个概念，我们把包含子孙概念的规则 R_1 称为子概念规则，把包含祖先概念的规则 R_2 称为父概念规则.

$$R_1: g \in C_1 \cup C_2 \Rightarrow DE(g) = 1$$

$$R_2: g \in C_1' \cup C_2 \Rightarrow DE(g) = 1$$

$$H(C_1, C_1')$$

$$Coverage(R_1) = Coverage(R_2)$$

注：$g \in C$ 表示基因 g 被注释到功能概念 C，$g \in C_1 \cup C_2$ 表示基因 g 同时被注释到两个功能概念 C1

和 C2，$DE(g)=1$ 表示基因 g 的类别为 1，即基因 g 差异表达. $H(A, B)$ 表示概念 A 和概念 B 在 ontology 中位于同一条从根结点到叶结点的路径上，且 B 更靠近根结点. *Coverage*（R）表示规则 R 覆盖（cover）的基因集合，即能注释到规则 R 前件中所有功能概念的基因集合.

由于子规则能够覆盖（cover）全部父规则覆盖的基因，且它的内涵更丰富，我们取子规则 R_1 而舍父规则 R_2. 本操作使搜索速度大大加快.

2.4 规则相互包含时的取舍

如果两个规则覆盖的基因完全相同，且其中一个规则的所有概念包含了另一规则的所有概念，我们把包含概念多、限制条件多的规则称为包含规则 R_1，把包含概念少、限制条件少的规则称为被包含规则 R_2. 包含规则与被包含规则能覆盖（cover）相同的基因，且前者的内涵更丰富、限制条件更多，我们在算法中取包含规则 R_1 而舍被包含规则 R_2.

$$R_1: g \in C_1 \cup C_2 \cup C_3 \Rightarrow DE(g) = 1$$
$$R_2: g \in C_1 \cup C_2 \Rightarrow DE(g) = 1$$
$$Coverage(R_1) = Coverage(R_2)$$

2.5 实现

目前我们采用了基于网络浏览器的 B/S 结构来实现 RuleGO，服务端的架构为 CGI + MATLAB + SQL Server，用户通过网页将数据提交给服务器 CGI 程序，该程序调用后台 MATLAB 辅助运算，同时从 SQL Server 中读取 Gene Ontology 以及基因注释的相关信息，最终结果用网页形式返回给用户.

该软件的具体位置在 http://www.biocc.net/K/RuleGO/index.htm.

3 数值实验

3.1 基因表达谱数据集

a. 结肠癌数据集：Affymetrix 公司的结肠癌基因表达谱实验数据（http://www.sph.uth.tmc.edu/hgc），原实验点有 6 500 个寡聚核苷酸探针组的 DNA 芯片，样本包括 40 例结肠癌组织和 22 例正常组织，在本研究中，我们仅采用 ALON[15] 筛选出的 2 000 个基因表达谱数据进行分析.

b. 腺癌数据集：Badea[16] 从 23 100 个 cDNA 克隆、73 个组织样本的 cDNA 芯片数据集上选择出 918 * 41 的子数据集，并且将 41 个腺癌组织样本分为两大类. 我们选用了 Badea 的数据集.

3.2 阳性基因集和阴性基因集的确定

对两个数据集，我们首先将每张芯片的每个数据减去列均值再除以列标准差，使芯片数据得到标准化[17]，然后对所有参与实验的基因做两类不同条件下表达值的成组 t 检验[4]，得出每个基因的 t 检验 P 值，最后我们采用 $P<0.005$ 的条件来选择显著差异表达基因，采用 $P>0.5$ 的条件来选择显著不差异表达基因[16].

3.3 应用 RuleGO 挖掘基因差异表达的功能规则

我们经验性地限制广义注释的最高层数 $l_{cut}=5$，设置信度阈值 $min\_conf=0.9$，以确保规则的可靠性；采用动态频数阈值技术，确定 l_{cut} 层的支持度 $min\_sup\_l_{cut}=0.1$. 两个数据集的总基因数 n 不同，根据 $min\_occ\_l_{cut}=n\times min\_sup\_l_{cut}$，计算得两个数据集在 l_{cut} 层上的频数阈值分别是 $min\_occ\_l_{cut}$（colon）=16 和 $min\_occ\_l_{cut}$（AC）=4.

3.4 运用 OntoExpress 算法进行对照实验

OntoExpress 算法逐个考察 GO 的各个功能结点，检验该功能结点作为一个整体的差异表达显著性. 我们在上述两个数据集上运行 OntoExpress，与 RuleGO 结果比较，取 OntoExpress 假设检验的显著性水平 $\partial=0.05$.

4 实验结果

4.1 挖掘出的主要功能规则及其生物学意义

我们在大肠癌数据集和腺癌数据集上分别挖掘出 116 条和 39 条功能关联规则，其中大多数的规则由至少两个功能概念构成. 我们把两套数据集上挖掘出的所有规则示于网页 http://www.biocc.net/K/RuleGO/index.htm

为检验这些功能关联规则是否与结肠癌的发病有关，我们在 PubMed 上进行关键词检索，得到不少文献支持. 结肠癌的挖掘结果中多条规则都包含了“核糖体结合”或“核糖核蛋白复合物”功能概念，而 Chu[18] 在结肠癌组织中发现，9 种构成核糖核蛋白体（RNP）复合物细胞的 RNA 中，有 7 种与胸苷酸合成酶有高亲合性，说明结肠癌组织中核糖核蛋白复合物确实是一个值得研究的对象. 另一条关联规则涉及到“GO：0006261：DNA dependent DNA replication”，而 Klingler[19] 也指出，10% ~15% 的人类结肠癌与 DNA 复制中的误配修补系统（MMR）缺陷有关.

结肠癌的挖掘结果中大多数规则都涉及到

“RNA 结合”或“核糖核蛋白体”概念. 但我们注意到，规则“GO：0015935：small ribosomal subunit，GO：0006412：protein biosynthesis”甚至将细胞位置进一步缩小，定位到核糖体小亚基，这为结肠癌的进一步研究提供了一个方向. 另外，8 个基因同时被注释到功能概念“GO：0003677：DNA binding，GO：0003723：RNA binding，GO：0006139：nucleobase，nucleoside，nucleotide and nucleic acid metabolism”，它们无一例外地都差异表达，这个现象使人们再次注意到 DNA 代谢、RNA 代谢在结肠癌组织中的异常.

4.2 通过锐化基因两类分界凸现实验相关概念

RuleGO 摒弃过渡状态的基因，采用极端差异表达和极端不差异表达的基因构成阳性基因集合和阴性基因集合，这类似于图形学中的锐化作用，使得一些与实验条件有一定相关性的概念得以被发现，和 OntoExpress 对照实验相比，典型的例子见表 2.

Table 2 Examples of rules with higher confidence in RuleGO than in OntoExpress due to the elimination of transitional genes in RuleGO

Name of the concept	OntoExpress confidence	RuleGO confidence	Dataset
GO：0030529：ribonucleoprotein complex	26/45 = 0.58	25/27 = 0.93	Colon cancer
GO：0006261：DNA dependent DNA replication	5/11 = 0.45	5/5 = 1	Colon cancer
GO：0000278：mitotic cell cycle	10/22 = 0.45	10/10 = 1	Adenocarcinoma

4.3 发现置信度较高的组合功能概念

OntoExpress 只能评价单个功能分类单元与实验条件的相关性，而 RuleGO 能够自主地发现与基因差异表达现象高度相关的概念组合. 我们的结果显示，在两套数据集的结果规则中，分别有 62 % 与 82% 的规则包含的功能概念数大于 1. 并且我们还注意到（表 3），某些单个功能概念构成的规则置信度较低，而该功能概念与其他功能概念组合形成的规则置信度就明显提高. 这说明现有的基因功能分类体系用于基因表达谱分析时，其分类单元还不够精细，或有必要同时综合多个分类准则对基因进行分类.

Table 3 Two example rules demonstrating apparent increase in confidence after incorporating more gene function units

Rules with single concept	Confidence	Rules with combinatorial concepts	Confidence
GO：0003723：RNA binding	50/96 = 0.52	GO：0003723：RNA binding， GO：0030529：ribonucleoprotein complex	23/25 = 0.92
GO：0005830：cytosolic ribosome（sensu Eukarya）	8/14 = 0.57	GO：0005830：cytosolic ribosome（sensu Eukarya）， GO：0006412：protein biosynthesis	8/8 = 1

4.4 组合功能概念的规则支持度和置信度较高

按规则中涉及的功能概念数将规则分类，得到 1-规则（前件中包含 1 个功能概念的规则）、2-规则、3-规则和 4-规则四类. 这些都是满足置信度 $conf \geqslant 0.9$ 及动态频数阈值限的规则（表 4），我们观察到支持度较高（$occ > 10$）的规则主要分布在 2-规则和 3-规则中，可见单个概念构成的可靠规则并不多，支持度和置信度都高的规则多是由 2 个或 3 个功能概念组成的. 这在生物学上也易于理解：比如 Zhou[6] 就指出，单纯从生物学过程角度对基因分组还不够，因为位于细胞不同位置的基因产物即使参与同样的生物学过程，它们的性质也是不同的，表达行为也可能很有差异.

Table 4 Distribution of rules with the number of covered genes ≥10 and the confidence ≥0.9 in different rule-classes (classified according to the number of function units incorporated in the rule)

Classification of rules	Number of rules	Mean support	Number of rules with support > 5
Rules with 1 concept	44	2.07	1
Rules with 2 concepts	46	4.65	8
Rules with 3 concepts	20	7.30	8
Rules with 4 concepts	6	6.17	1

5 讨 论

RuleGO 摒弃了中间阶段的基因，使得阳性基因集合和阴性基因集合更“纯”，相互区分更明显，从而寻找与实验条件相关的功能类更准确. 从我们的实验结果来看，这样操作提高了算法对相关功能类的发现能力.

RuleGO 采用关联规则思想，并且结合 GO 的特点改进 Apriori 算法，加入动态最小支持度技术、冗余相关规则排除技术等，对生物学问题更有针对性，更有能力发现与实验条件相关的基因功能类，同时排除冗余规则，大大加快了搜索速度，也使结果简洁明了.

与商业化的 OntoExpress 相比，RuleGO 可以挖掘出 OntoExpress 不能挖掘出的组合功能概念形式. 我们的数值实验表明，可信度和支持度都高的质量好的规则，大都由 2 ~ 3 个功能概念组成，而用 OntoExpress 不能将这部分信息发掘出来. 譬如，与疾病分型有关的一组基因的蛋白质产物，实际是位于细胞位置 A 中生物过程 B 的酶，而位于细胞位置 A 的基因产物或参与生物过程 B 的基因产物都没有整体上的差异表达倾向，在这种情况下，与 OntoExpress 相比 RuleGO 的优越性是明显的.

RuleGO 仍有待改进的空间. 用任何统计模型进行统计检验，都存在Ⅰ型错误（假阳性）和Ⅱ型错误（假阴性），二者相互制约，只有增加实验样本大小才能同时降低两类错误率. 对于 RuleGO，降低阳性集的假阳性和阴性集的假阴性更为重要，为此需要提高检验标准. RuleGO 的核心是关联规则挖掘算法，要求阳性集、阴性集具有一定的样本量，而提高检验标准会使样本量减少，目前的方法是在检验标准和样本含量之间进行折中. 我们将在进一步的工作中对此进行改进：依据基因的 P 值为基因加权，这样既考虑了错误率，又保证了 RuleGO 输入集的样本含量，并使挖掘出的规则更加可信.

参 考 文 献

1 Ramsay G. DNA chips: state-of-the art. Nature Biotechnology, 1998. **16**（1）: 40 ~ 44

2 Butte A J, Ye J, Neiderfellner G. Determining significant fold differences in gene expression analysis. Pacific Symposium on Biocomputing, Hawaii, 2001

3 Mariani T J, Budhraja V, Mecham B H, *et al*. A variable fold change threshold determines significance for expression microarrays. FASEB J, 2003, **17**（2）: 321 ~ 323

4 Smyth G K, Yang Y H, Speed T. Statistical issues in cDNA microarray data analysis. Meth Mol Biol, 2003, **224**: 111 ~ 136

5 Draghici S, Khatri P, Martins R, *et al*. Global functional profiling of gene expression. Genomics, 2003, **81**（2）: 98 ~ 104

6 Zhou X, Kao M C. Transitive functional annotation by shortest-path analysis of gene expression data. Proc Natl Acad Sci USA, 2002, **99**（20）: 12783 ~ 12788

7 Ashburner M, Ball C. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet, 2000, **25**（1）: 25 ~ 29

8 Rison S C, Hodgman T C, Thornton J M. Comparison of functional annotation schemes for genomes. Funct Integr Genomics, 2000, **1**（1）: 56 ~ 69

9 Lennon G G, Auffray C. The I. M. A. G. E. Consortium: An integrated molecular analysis of genomes and their expression. Genomics, 1996, **33**（1）: 151 ~ 152

10 Benson D, Karsch M. GenBank. Nucleic Acids Research, 2003, **31**（1）: 23 ~ 27

11 Wheeler D L, Church D M, Federhen S, *et al*. Database Resources of the National Center for Biotechnology. Nucl Acids Res, 2003, **31**（1）: 28 ~ 33

12 Pruitt K D, Katz K S. Introducing RefSeq and LocusLink: curated human genome resources at the NCBI. Trends in Genetics, 2000, **16**（1）: 44 ~ 47

13 Havidsten T R, Komorowski J. Predicting gene function from gene expressions and ontologies. Pac Symp Biocomput, 2001, **6**: 299 ~ 310

14 Han J, Kambr M. Data Mining: Concepts and Techniques. San Francisco: Morgan Kaufmann Publishers. 2000. 230 ~ 236

15 Alon U, Barkai N, Notterman D A. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Cell Biol, 1999, **96**（12）: 6745 ~ 6750

16 Badea L. Functional Discrimination of Gene Expression Patterns In Terms of The Gene Ontology. Pacific Symposium on Biocomputing, Hawaii, 2003

17 Ross D T, Scherf U. Systematic variation in gene expression patterns in human cancer cell lines. Nat Genet, 2000, **24**（3）: 227 ~ 235

18 Chu E, Cogliati T, Copur S, *et al*, Identification of *in vivo* target RNA sequences bound by thymidylate synthase. Nucleic Acids Research, 1996, **24**（16）: 3222 ~ 3228

19 Klingler H, Hemmerle C, Bannwart F, *et al*, Expression of the hMSH6 mismatch-repair protein in colon cancer and HeLa cells. Swiss Med Wkly, 2002, **132**（5 ~ 6）: 57 ~ 63

Mining Association Rules of GO Function Classes and Gene Expression Difference*

TU Kang[1], YU Hui[1], GUO Zheng[1,2]**, LI Xia[1]

([1] *Department of Medical Mathematics and Biomedical Engineering, Harbin Medical University, Harbin* 150086, China;

[2] *Harbin Gene-Tech Biochip Development Inc, LTD, Harbin* 150086, China)

Abstract To adapt to the hierarchical structural property of Gene Ontology, the standard Apriori algorithm is modified into a novel algorithm, RuleGO, which mines association rules of GO function classes and gene expression difference. The inputs of RuleGO are one set of differential expressed genes and another set of non-differential expressed genes, and the outputs of RuleGO are association rules linking GO function combinations to gene differential expression. Rules mined by RuleGO may guide insights into gene expression difference at the functional level, towards the clarification of the process of pathological changes or the mechanism of medicine. Both RuleGO and OntoExpress are applied to the datasets of colon cancer and adenocarcinoma, and RuleGO turned out to be more powerful to mine relevant function rules than OntoExpress. The experimental results also reveal that rules with both high significance and high support mostly involve more than one gene function classes, suggesting that considering the combination of multiple gene function classes may be more resonable in gene expression analysis than taking into account only a single gene function class.

Key words gene expression profile, gene expression difference, characteristic functional classes, association rule, Gene Ontology

---

* This work was supported by grants from The National Natural Science Foundation of China (39970397, 30170515), State 863 High Technology R&D Project of China (2003AA2Z2051, 2002AA2Z2052), The Natural Science Foundation of Heilongjiang (GB03C602-4, F01777), The Natural Science Foundation of Harbin (2003AA3CS113) and The 211 Project of The Tenth Five-year Plan of Harbin Medical University.

** Corresponding author. Tel: 86-451-86615933, E-mail: markgz@0451.vip.com

Received: January 17, 2004 Accepted: March 30, 2004