生 物 工 程 学 报
journals.im.ac.cn
cjb@im.ac.cn

*Chin J Biotech* 2008, September 25; 24(9): 1643-1648
Chinese Journal of Biotechnology ISSN 1000-3061


研究简报

# 特征代谢通路上的基因表达相关性及共调控表达模式

华琳, 郑卫英, 刘红, 林慧, 高磊

首都医科大学生物信息学实验室与数学教研室, 北京 100069

**摘　要:** 利用随机森林-通路分析法, 通过袋外样本 OOB 的分类错误率筛选特征代谢通路, 在特征通路上作基因表达相关性研究并对通路上的基因采用 MAP(Mining attribute profile)算法挖掘不同实验条件下基因的共调控表达模式, 对共调控表达模式进行聚类。分析结果显示同一特征代谢通路上的基因表达倾向相似, 有 2 条特征代谢通路存在共表达模式, 其中一条通路含 108 个表达模式,对这些模式进行聚类, 其最低聚类的相似系数仍高达 0.623, 说明同一特征代谢通路上的基因共表达模式在不同实验条件下仍具有高度的相似性。对以通路作为基因模块进行复杂疾病的研究具有借鉴意义。

**关键词:** 基因表达, 通路, 模式, 关联

# Relativity of Gene Expression and Co-regulated Gene Patterns in Feature KEGG Pathways

Lin Hua, Weiying Zheng, Hong Liu, Hui Lin, and Lei Gao

*Department of Bioinformatics, Capital University of Medicine, Beijing* 100069, *China*

**Abstract:** We revealed the feature pathways by computing the classification error rates of out-of-bag (OOB) by random forests combined with pathway analysis. At each feature pathway, the relativity of gene expression was studied and the co-regulated gene patterns under different experiment conditions were analyzed by MAP (Mining attribute profile) algorithm. The discovered patterns were also clustered by the average-linkage hierarchical clustering technique. The results showed that the expression of genes at the same pathway was similar. The co-regulated patterns were found in two feature pathways of which one contained 108 patterns and the other contained 1 pattern. The results of clusters showed that the smallest Pearson coefficient of the clusters was more than 0.623, indicating that the co-regulated patterns in different experiment conditions were more similar at the same KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway. The methods can provide biological insight into the study of microarray data.

**Keywords**: gene expression, pathway, pattern, association

微阵列技术(Microarray)对于从分子水平上解释复杂多样的生物学过程有着重要的作用。根据高通量基因表达谱数据, 采用数据挖掘技术识别复杂疾病相关的特征基因与功能, 对于研究疾病机理、预测疾病类型有重要的意义。基因的表达行为不是孤立的, 功能相关基因在表达上倾向于高度相关,已经有许多研究人员采用不同定义的基因模块对复杂疾病进行了有意义的研究。有报道位于同一代谢通路上的基因表达倾向于高度相关[1−3], 由此,我们以代谢通路作为功能基因模块进行研究。我们利用 KEGG 通路信息和基因表达谱数据, 提出基于通路的随机森林 – 通路分析法[4], 通过计算每个通路袋

**Received**: November 27, 2007; **Accepted**: February 27, 2008
**Corresponding author:** Liu Hong. Tel: +86-10-83911553; E-mail: liuhong68@sina.com

外样本数据的分类错误率来发现特征代谢通路，对特征代谢通路上的基因表达作相关性研究。

另一方面，近年来，关联模式发现(APD)方法[5]常被用来发现共调控基因模式。对比有监督和无监督的算法，它们不仅能够在全部基因集中发现共调控的基因，还可以在不同的实验样本或实验条件内发现共调控的基因。因此，每个基因和每个实验条件都可能出现在一个以上的模式中。传统 APD 方法通过表达模式获得关联规则，因此算法多集中在挖掘基因表达模式上。这样挖掘出的表达模式常常集中在模式中的基因都上调或者都下调，而不能发现这样的模式，即模式中一个基因的上调可能会使其他基因下调。我们采用对 APD 方法的改进，即 MAP (Mining attribute profile)[6]算法来挖掘共调控基因表达模式。这种方法可以同时挖掘基因表达变化一致或相反的模式。我们将挖掘出的基因共表达模式进行聚类，研究同一代谢通路上基因共表达模式的相似度。

## 1 材料与方法

### 1.1 数据集

我们选择数据集 Canine dataset[7]，此数据集是研究药物是否会对狗冠状动脉产生损伤。数据集中包含 29 只狗，药物实验条件为 29 个。实验结果分为有损伤和无损伤。其中有损伤狗为 12 只，无损伤狗为 17 只。Microarray 数据部分中包含 12473 个基因的表达值，通路数据部分含 129 条 KEGG 通路和 312 条 BioCarta 通路，共 441 条通路。我们仅选择 129 条 KEGG 通路(相当于 129 个基因集)，每条通路的基因为 1 到 151 个。总基因数为 2883 个(除去各通路中的重复基因，实际基因数为 1442 个)。数据集描述见表 1。

**表 1 数据集描述**
**Table 1 Data description**

| Index | Description |
|---|---|
| Number of conditions | 29 |
| Number of genes | 1442 |
| Log ratios | 2 |
| $\delta$ | ± 1 |

### 1.2 随机森林算法筛选特征代谢通路

对每条 KEGG 通路，我们应用 bootstrap 法有放回地随机抽取新的自助样本并由此形成分类树，每次未被抽到的样本组成袋外数据(Out of bag, OOB)。我们采用随机森林分类法求出 OOB 的分类错误率。我们对全部 129 条 KEGG 通路分别计算了 OOB 分类错误率，结果为 3.45%、6.90%、10.34%、17.24%、24.72%和 35.69%等。分类错误率越低，说明通路的分类效果越好。因此，我们考虑选择分类错误率 <10.34%作为特征代谢通路。

### 1.3 特征代谢通路的基因表达相关性研究

我们将筛选出来的特征 KEGG 通路,按无损伤和有损伤分别计算每个通路上两两基因的 Pearson 相关系数[8]：即

$$\rho_{XY}=\frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2\sum_{i=1}^{n}(Y_i-\bar{Y})^2}}$$

其中，$X_i$ 和 $Y_i$ 分别表示基因 $X$, $Y$ 在第 $i$ 个样本中的表达水平。在每个通路上分别计算有损伤和无损伤组的 Person 相关系数的均值。为了检测两两基因的相似性测度[9]是否具有统计学意义，我们采用随机重排方法得到基因相似性测度的分布，给定显著性水平$\alpha$=0.05，可确定临界值 $r_0$，若 $r>r_0$(即 $P$<0.05)，则说明相关系数有统计学意义。

### 1.4 挖掘共调控基因表达模式

#### 1.4.1 *数据预处理*

对 29 个实验条件下的基因表达值取以 2 为底的对数，并采用变换 $e_i=\frac{X_i-\bar{X}}{S}$ 进行标准化处理。其中 $X_i$ 为基因 $i$ 的对数表达值，$\bar{X}$，$S$ 分别为对数值的均数和标准差。

#### 1.4.2 *共调控基因表达模式*

设 $e_{i,u}$ 表示第 $i$ 个基因 $g_i$ 在第 $u$ 个实验条件 $c_u$ 下的标准化后的对数表达值，设$\delta$为用户指定的界值，则令 $\begin{cases}e_{i,u}\geq\delta, & g_i\text{上调}\\ e_{i,u}<\delta, & g_i\text{下调}\end{cases}$，若对于基因 $g_i$, $g_j$, 满足 $e_{i,u}\geq\delta$ 且 $e_{j,u}\geq\delta$或 $e_{i,u}\leq-\delta$且 $e_{j,u}\leq-\delta$, 这时我们称这 2 个基因表达一致，反之，若 $e_{i,u}\geq\delta$且 $e_{j,u}\leq-\delta$或 $e_{i,u}\leq-\delta$且 $e_{j,u}\geq\delta$, 则称这 2 个基因表达相反。我们设 $U$ 表示实验条件集合，I, J$\subseteq G$是2个在实验条件集合 $U$ 下表达相反的 2 个基因集，在 I, J 内中的基因表达一致。此时,我们称{I(–)J}为一个基因共调控模式。如图 1，上面四条线表示基因 A、B、C、D 在 8 个实验条件下的表达水平，下面 4 条线表示基因 E、F、

G、H 在 8 个实验条件下的表达水平,我们将基因共调控模式表示为: {A, B, C, D (-) E, F, G, H}.

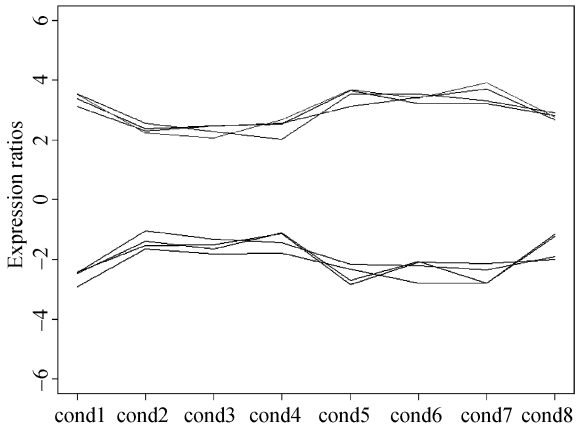


**图 1 共调控基因表达模式**
**Fig. 1 A sample expression pattern that can be discovered by APD methods**

#### 1.4.3 *挖掘基因共调控表达模式算法*

我们采用 Gyenesei[10,11]等提出的 MAP(Mining attribute profiles)算法。第一步创建基因表达模式 GP(Gene pattern)树。GP 树通过顺序读取实验条件下的基因表达数据, 并把每个实验条件定位在一个 GP 树通路上。若从树上的第一个基因开始, 在 2 个或更多的实验条件下, 基因表达模式相同, 则只定位一个 GP 通路。这样可以缩小树的结构。这样的挖掘算法可以节省计算机运行时间。第二步是采用自上而下的递归策略从构建好的 GP 树上挖掘基因共调控表达模式。通过不断分解 GP 树形成互不关联的子树, 直到最后的子树仅有一个单支,从这些单支上便可以列举出所有的共调控基因表达模式。

#### 1.4.4 *对共调控基因表达模式聚类*

由于 Microarray 数据中固有的噪声,使得发现的很多共调控基因表达模式是冗余的。因此,我们对这些模式采用聚类的方法,设 $P_i=\{P_i^L-P_i^R\}$ 和 $P_j=\{P_j^L-P_j^R\}$ 表示 2 个基因模式, $P_i^L, P_i^R, P_j^L, P_j^R$ 表示 4 个基因集。我们定义 $P_i$ 和 $P_j$ 的相似度度量为[3]:

$$s(P_i,P_j)=\frac{\max\{P_i\Delta P_j, p_i\nabla P_j\}}{P_i^L\cup P_i^R\cup P_j^L\cup P_j^R}$$

其中:

$P_i\Delta P_j=(P_i^L\cap P_j^L)+(P_i^R\cap P_j^R), P_i\nabla P_j=(P_i^L\cap P_j^R)+(P_j^L\cap P_i^R)$ 括号内的表达式表示基因集交集或并集的基因数目。我们以 s 值做为相似度量,采用层次聚类法进行聚类,并分析聚类后的结果。

## 2 数值实验结果

### 2.1 特征代谢通路的筛选

我们选择 $n$=5000 棵树构建随机森林,对 129 个 KEGG 通路分别采用随机森林算法计算袋外样本 OOB(Out-of-bag)分类错误率并从中筛选出 OOB 分类错误率≤10.34%的 11 条重要通路,通路列表如下:

**表 2 OOB 分类错误率≤10.34%的 11 条 KEGG 通路**
**Table 2 11 pathways ranked by OOB error rates of ≤10.34%**

| Pathway | OOB error rates (%) | Number of genes |
|---|---|---|
| Aminoacyl-tRNA biosynthesis | 3.45 | 19 |
| Leukocyte adhesion | 3.45 | 59 |
| Circadian rhythm | 6.90 | 9 |
| Parkinson's disease | 6.90 | 15 |
| One carbon pool by folate | 10.34 | 11 |
| Phenylalanine_ tyro | 10.34 | 12 |
| Arginine and proline | 10.34 | 46 |
| Glycine and serine | 10.34 | 34 |
| N-Glycans biosynthesis | 10.34 | 32 |
| Aminosugars metabolism | 10.34 | 19 |
| Wnt signaling pathway | 10.34 | 68 |

从表 2 中可以看出 OOB 分类错误率为 3.45%有 2 条通路, OOB 分类错误率为 6.90%有 2 条通路, 分类错误率为 10.34%为 7 条通路。我们可以直接从列表中作生物学的解释。Circadian rhythm 通路反应的是心血管循环系统功能,它显然与动脉损伤相关。其中基因 CSNK1A1 既在通路 Circadian rhythm 上,同时也在心血管系统中与组织缺氧和 p53 活动相关的通路[12]上, 显然也与动脉损伤相关。基因 PRKCA2 在信号通路上, 有证据证明[13]基因 PRKCA 和蛋白质激酶及活动的 RhoA 能够形成一种复合体, 会对动脉炎症产生影响。Aminoacyl-tRNA biosynthesis 通路和 One carbon pool by folate 通路看起来和动脉病理炎症或反应不直接相关, 但是这些通路上的基因具有一定的生物学意义。如通路 Aminoacyl-tRNA biosynthesis 上的基因 GARS 是自体免疫疾病的自体抗体靶标[14]。通路 One carbon pool by folate 上的基因 MTTHFD2,有证据表明当遇到化学物质时, 它在动脉血管上皮细胞中是上调的[15]。

为了验证随机森林法筛选特征代谢通路的有效

性, 我们另选了 2 个数据集, Breast cancer 数据集(49 个乳腺癌病人, 一期 6 人, 二期 16 人, 三期 27 人, 22215 个基因表达值)和 Gender 数据集(32 个个体, 男 17 人, 女 15 人, 22 283 个基因表达值)[7]。2 个数据集通路数据部分均含 129 条 KEGG 通路和 312 条 BioCarta 通路, 共 441 条通路。对乳腺癌数据集, OOB 分类错误率排在前 5 位的通路与乳腺癌相关, 而对于 Gender 数据集, OOB 分类错误率排在前两位的通路均与性别相关。

### 2.2 特征代谢通路的基因表达相关性研究

我们采用随机重排方法得到基因相似性测度的分布,给定显著性水平$\alpha$=0.05, 可确定临值 $r_0$, 数值计算结果得到$r_0$=0.254。除了通路 One carbon pool by folate 的无损伤基因的相关系数无统计学意义, 其他通路所有的相关系数均有统计学意义, 这说明同一代谢通路中的基因在表达上倾向于相似。我们将其余 10 条通路(10.34% 6 条, 6.9% 2 条和 3.45% 2 条)的每条通路无损伤和有损伤的 Person 相关系数进行比较(见图 2)。其中 1~6 条通路的分类错误率为 10.34%, 7、8 条通路的分类错误率为 6.9%,9、10 条通路的分类错误率为 3.45%。从图 2 中发现: 2、3、4、6、7、10 通路有损伤的相关系数要低于无损伤相关系数, 说明这 6 条通路中, 无损伤组的大多数基因相对有损伤显示为高表达, 而 1、5、8、9 通路有损伤的相关系数高于无损伤相关系数, 说明这 4 条通路中有损伤组的大多数基因相对无损伤显示为高表达。

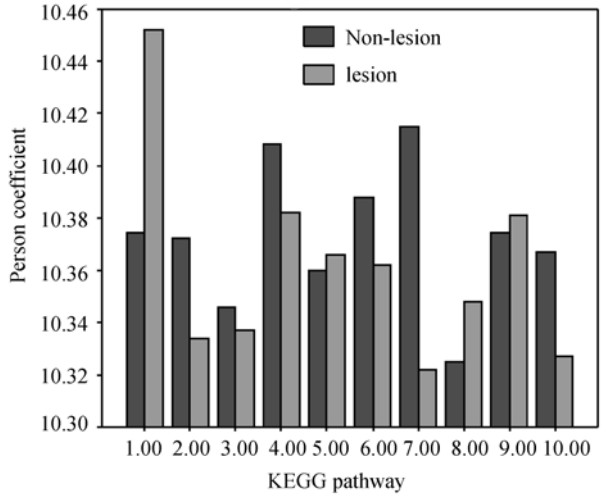


**图 2 10 条重要通路 Person 相关系数比较**

**Fig. 2 Pearson coefficient in 10 feature pathways of two groups(lesion and non-lesion)**

### 2.3 挖掘共调控基因表达模式

我们应用 MAP(Mining attribute profiles)算法挖掘基因共调控表达模式。我们设定每个模式的最小实验条件数为 10, 最小调控基因数为 10。我们的目的是发现在至少 10 个生物实验条件下, 最少有 10 个共调控基因的基因表达模式。结果显示,仅在通路 Arginine and proline 和通路 Glycine and serine 上挖掘出共调控基因表达模式。其中在通路 Arginine and proline 上发现了 108 个共调控基因表达模式, 在 Glycine and serine 上仅发现一个共调控基因表达模式。见表 3。表 3 中列出共调控基因表达模式的挖掘结果。图 3 给出 Arginine and proline 通路最大的共调控基因表达模式: 10 个实验条件, 14 个基因, 这里的基因用基因探针号表示。其中横轴表示实验条件,纵轴表示标准化后基因的对数表达值, 每个颜色的曲线表示每个基因在不同实验条件下的对数表达值变化。其共调控基因表达模式为:

{GL-Cf-483, GL-Cf-1694, GL-Cf-1847, GL-Cf-3839, GL-Cf-6385, GL-Cf-9199, GL-Cf-1356(-)GL-Cf-698, GL-Cf-1978, GL-Cf-2648, GL-Cf-3198, GL-Cf-7804, GL-Cf-8850, GL-Cf-10255}

图 4 给出 Glycine and serine 通路的共调控基因表达模式。即:

{GL-Cf-483, GL-Cf-633, GL-Cf-1575, GL-Cf-2958, GL-Cf-3100, GL-Cf-6385, GL-Cf-8428, GL-Cf-12575 (-)GL-Cf-4154, GL-Cf-7804}

### 2.4 对共调控基因表达模式聚类

我们以 s 值作为相似度量,对在通路 Arginine and proline 上挖掘的 108 个共调控基因表达模式采用层次聚类法进行聚类, 双向聚类结果见图 5, 横向表示对基因聚类, 纵向表示对共调控基因表达模式聚类。其中参与共调控表达模式的基因为 19 个基因。表 4 给出相似系数在 0.9 以上的共调控基因表达模式的聚类结果。我们发现共调控表达模式的结点中, 聚类的相似系数最低的仍高达 0.623。可见同一代谢通路上的基因在不同实验条件下的共调控表达模式也十分相似。

## 3 讨论

根据高通量基因表达谱数据, 采用数据挖掘技术识别与复杂疾病相关的特征基因或特征基因集合, 对于研究疾病机理, 预测疾病类型有着重要的意义。

**表 3 MAP 算法挖掘的共调控基因模式结果**
**Table 3 Results of co-regulated gene patterns discovered by the MAP methods**

| Pathway | Index | Results |
|---|---|---|
| Arginine and proline | Number of co-regulated patterns | 108 |
| | Number of genes of largest pattern | 14 |
| | Average Pearson's correlation | 0.32 |
| Glycine and serine | Number of co-regulated patterns | 1 |
| | Number of genes of the pattern | 10 |
| | Average Pearson's correlation | 0.38 |

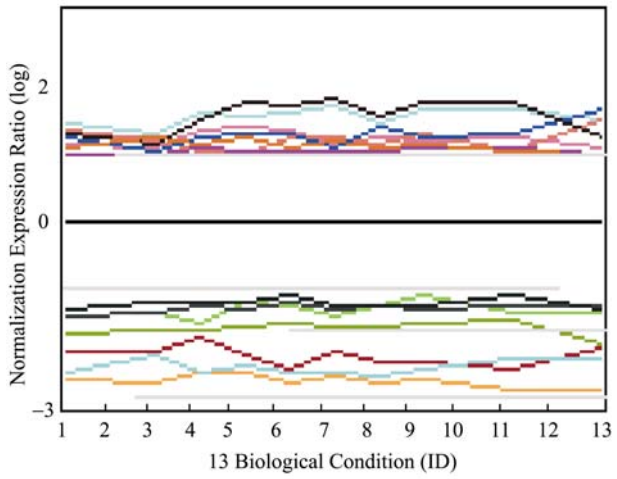


**图 3 Arginine and proline 通路最大共调控基因数的共调控模式**
**Fig. 3 Co-regulated pattern with most genes discovered by MAP methods in Arginine and praline pathway**

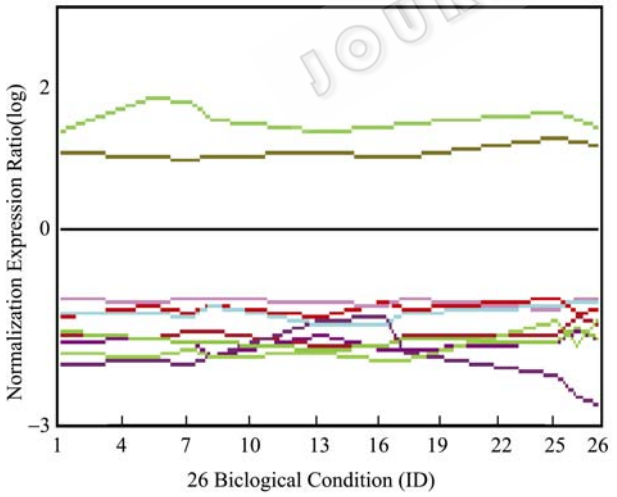


**图 4 Glycine and serine 通路基因共调控模式**
**Fig. 4 Co-regulated pattern discovered by MAP methods in Glycine and serine pathway**

而基因的表达行为并不是孤立的，位于同一代谢通路上的基因在表达上倾向于高度相关。因此,我们以通路为单元，采用随机森林算法筛选出特征代谢通路，再对特征代谢通路上的基因进行相关性研究和共调控基因表达模式的研究。它充分地利用了 Microarry 的数据信息和通路信息。重要的是，它可以通过先发现特征代谢通路，再在特征代谢通路基础上进行研究，这样可以直接从生物学上解释基因的功能。

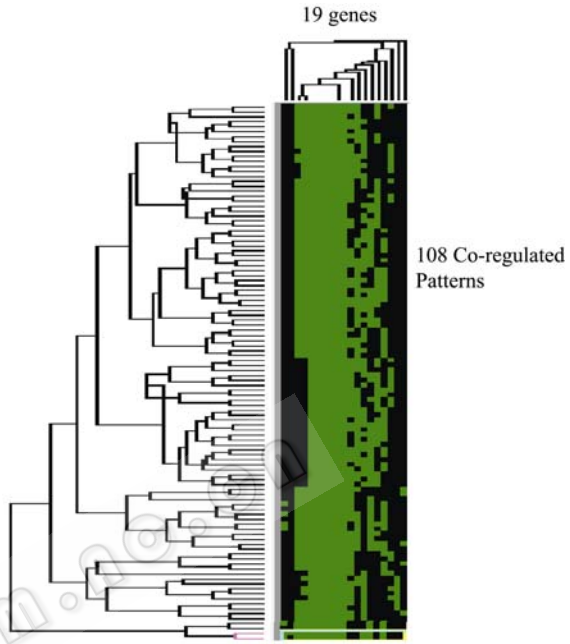


**图 5 基因共调控表达模式聚类图**
**Fig. 5 Clustering of the 108 patterns discovered by MAP methods**

**表 4 共调控基因表达模式聚类结果(相似系数>0.9)**
**Table 4 Cluster results of the 108 patterns discovered by MAP methods (similarity coefficient>0.9)**

| Similarity coefficient | Number of co-regulated patterns |
|---|---|
| 0.929 | 2 |
| 0.923 | 6 |
| 0.916 | 14 |
| 0.909 | 23 |

另一方面，我们采用传统 APD 算法的改进算法，即 MAP 算法，它可以更深入的挖掘 Microarray 数据。此算法可以同时挖掘在不同实验条件下的基因共调控表达模式并能够发现传统的 APD算法所不能发现的隐藏调控模式，而这些隐藏的调控模式可能具有重要的生物学功能。同时, APD 算法丢失的一些重要信息也能够被 MAP 算法发现。因此, MAP 算法对于分析 Microarray 数据，挖掘基因谱的表达模式具有很好的应用前景。

我们为了消除 Microarray 数据的噪声，采用层

次聚类将挖掘的共调控基因表达模式进行聚类，证实在特征代谢通路上共调控基因表达模式具有很大的相似性。

本实验使用的是 KEGG 通路数据库，它的数据库每天都在不断更新，因此可能与真实的生物学过程有一定出入。另外，基因并不是独立的发挥功能，我们还将进一步研究这些基因如何互作而导致所观察的表型，这将是更深入的生物学课题。

## REFERENCES


[1] Curtis RK, Oresic M, Vidal-Puig A. Pathways to the analysis of microarray data. *Trends Biotechnol*, 2005, **23**: 429–435.

[2] Harris MA,Clark J, Ireland A, *et al*. The gene ontology (GO) database and informatics resource. *Nucl Acids Res*, 2004, **32**: D258–D261.

[3] Kanehisa M, Goto S, Kawashima S, *et al*. The KEGG resource for deciphering the genome. *Nucl Acids Res*, 2004, **32**: D277–D280.

[4] Pang H, Lin AP, Holford M, *et al*. Pathway analysis using random forests classification and regression. *Bioinformatics*, 2006, **22**: 2028–2036.

[5] Prelic A. Discovering frequent closed itemsets for association rules. *Lec Notes Comp Sci*, 1999, **1540**: 398–416.

[6] Attila Gyenesei, Ulrich Wagner, Simon Barkow-Oesterreicher, *et al*. Mining co-regulated gene profiles for the detection of functional associations in gene expression data. *Bioinformatics*, 2007, **23**: 1927–1935.

[7] http://bioinformatics.med.yale.edu/pathway-analysis/rf.htm.

[8] Guo Z, Li X, Rao SQ. Ananlysis Method of Medical Information. Harbin: Harbin Press, 2001.
郭政，李霞，饶绍奇. 医学信息分析方法. 哈尔滨：哈尔滨出版社, 2001.

[9] Heyer, LJ, Kruglyak, S, Yooseph, S. Exploring expression data: identification and analysis of co-expressed genes. *Genome Res*, 1999, **9**: 1106–1115.

[10] Gyenesei A. Frequent pattern discovery without binarization:mining attribute profiles. *PKDD* 2006. *Lect Notes Artif Intell*, 2006, **4213**: 528–535.

[11] Georgii E, Lother Richter, Ulrich Ruckert, *et al*. Analyzing microarray data using quantitative association rules. *Bioinformatics*, 2005, **21**: 123–129.

[12] Appella E, Anderson CW. Post-translational modifications and activation of p53 by genotoxic stresses. *Eur J Biochem*, 2001, **268**: 2764–2772.

[13] Bolick DT, Orra W, Whetzel A, *et al*. 12/15-lipoxygenase regulates intercellular adhesion molecule-1 expression and monocyte adhesion to endothelium through activation of RhoA and nuclear factor-kB. *Arterioscler Thromb Vasc Biol*, 2005, **25**: 2301.

[14] Maglott D, Ostell J, Pruite KD, *et al*. Entrez gene: gene-centered information at NCBI. *Nucl Acids Res*, 2005, **33**: D54–D58.

[15] Sato N, Koichi K, Kentaro S, *et al*. Changes of gene expression by lysophosphatidylcholine in vascular endothelial cells: 12 up-regulated distinct genes including 5 cell growth-related, 3 thrombosis-related, and 4 others. *J Biochem*, 1998, **123**: 119–126.