Vol. 28 高等学校化学学报 No. 9
2007 年 9 月 CHEMICAL JOURNAL OF CHINESE UNIVERSITIES 1640 ~ 1644

不相交主成分分析(PCA)和遗传算法(GA)用于差异表达基因的识别

苏振强[1,3], HONG Hui-Xiao[2], TONG Wei-Da[3], PERKINS Roger[2], 邵学广[4], 蔡文生[1,4]

(1. 中国科学技术大学化学系, 合肥 230026; 2. Division of Bioinformatics, Z-Tech at FDA's National Center for Toxicological Research, Jefferson, AR 72079, USA;
3. Center for Toxicoinformatics, National Center for Toxicological Research(NCTR), US Food and Drug Administration(FDA), Jefferson, AR 72079, USA;
4. 南开大学化学系, 天津 300071)

摘要 建立了一种基于不相交主成分分析(Disjoint PCA)和遗传算法(GA)的特征变量选择方法, 并用于从基因表达谱(Gene expression profiles)数据中识别差异表达的基因. 在该方法中, 用不相交主成分分析评估基因组在区分两类不同样品时的区分能力; 用 GA 寻找区分能力最强的基因组; 所识别基因的偶然相关性用统计方法评估. 由于该方法考虑了基因间的协同作用更接近于基因的生物过程, 从而使所识别的基因具有更好的差异表达能力. 将该方法应用于肝细胞癌(HCC)样品的基因芯片数据分析, 结果表明, 所识别的基因具有较强的区分能力, 优于常用的基因芯片显著性分析(Significance analysis of microarrays, SAM)方法.

关键词 基因芯片; 主成分分析(PCA); 遗传算法(GA); 基因芯片显著性分析(SAM); 偶然相关

中图分类号 O652 **文献标识码** A **文章编号** 0251-0790(2007)09-1640-05

作为一种高通量筛选工具, 基因芯片(Microarray)技术的发展为研究基因表达的方法带来了一场革命[1]. 在微芯片上置入寡核苷酸探针或相应于 mRNA 序列的 cDNA, 与细胞内的 cDNA 或 cRNA 进行杂交, 可以纵观细胞整体基因表达的情况. 传统的基因表达方法只能逐一研究单个基因的表达, 不能同时监测多个基因的表达情况, 基因芯片技术则可以同时监测成千上万个基因表达水平的变化, 这种同时监测数万个基因的不同表达水平构成了与某一生理或病理现象相关的基因表达谱(Gene expression profiles).

随着基因芯片技术的深入研究和广泛应用, 分析基因芯片数据向人们提出了新的挑战. 基因芯片数据分析的基本任务之一是从数万个基因中寻找与疾病相关的差异表达基因, 这些基因将成为诊断或愈后相关疾病的基因标签(Gene signature)[2]. 由于基因芯片数据通常只有相对较少的样本(与上万个基因数相比)和相对较小的信噪比, 使得差异表达基因的识别非常困难.

倍数分析法(FC)是最早用于基因芯片数据分析的方法[3], 也是最早用于识别差异表达基因的方法. FC 通过设定基因表达变化倍数的阈值(Cut-off)来识别差异表达的基因, 但在实际应用中如何选择合适的阈值却没有标准可供参考. 由于基因表达的变化程度以及实验中显著表达的基因数都依赖于所研究的生物系统本身以及所采用的实验条件等诸多因素, 只简单地设定 FC 的阈值不足以确定基因的表达是否存在显著差异, 所得到的结果也不可靠. 为了弥补 FC 方法的不足, 许多统计学方法已被用于差异表达基因的识别, 如传统的 *t*-检验方法和基因芯片显著性分析(Significance analysis of microarrays, SAM)方法[4]等.

统计学方法通常假定基因之间是相互独立的, 评价基因是否在统计上有显著差异表达时只独立地

收稿日期: 2007-01-08.
基金项目: 国家自然科学基金(批准号: 20325517 和 20575031)资助.
联系人简介: 蔡文生, 女, 教授, 博士生导师, 主要从事化学信息学研究. E-mail: wscai@nankai.edu.cn;
TONG Wei-Da, 男, 博士, 主要从事毒理信息学及生物信息学研究. E-mail: weida.tong@fda.hhs.gov

考察单个基因，没有考虑基因间的协同作用. 但是，在基因芯片实验中，mRNA 在不同的生物条件下被抽提，这些生物条件与所研究的作用机理在功能上相互关联，很多有协同调控作用或者共享相同生物学途径(Biological pathway)的基因往往协同表达. 因此，在识别差异表达基因时，考虑基因之间的相互依赖关系是非常必要的.

本文提出了一种不相交主成分分析(Disjoint PCA)[5,6]和遗传算法(GA)相结合的特征变量选择方法，并用于从基因表达谱数据中识别差异表达的基因. 不相交 PCA 用于评估基因组在区分两类不同样品时的区分能力；GA 用于寻找区分能力最强的基因组；识别基因的偶然相关性采用一种统计方法进行评估.

1　原理与算法

1.1　不相交主成分分析

不相交 PCA 是指用两个独立的 PCA 分别提取两类不同样品的特征信息，然后再根据样品在两个主成分空间的投影进行分析. 对于差异表达基因的筛选问题，即对每一类样品分别进行 PCA 分析，然后根据所选择的基因数据在两类主成分空间投影的残差对其差异表达进行评价. 在不相交 PCA 分析中，两类样品之间互不相干，有利于分别提取两类样品的特征信息.

假设将两类不同样品的矩阵用 $\boldsymbol{X}_q(n_q \times p)$ 表示，q 为类别序号，对于两类样品的情况 $q=1$ 或 2，n_q 为两类样品的数目，p 为样品的特征总数(即基因芯片数据中的基因数目)，则两类样品的 PCA 模型可以通过下式得到：

$$\boldsymbol{X}_q^{\mathrm{c}} = \boldsymbol{X}_q - \mathrm{repmat}(\boldsymbol{m}_q, n_q, 1) \tag{1}$$

$$\boldsymbol{X}_q^{\mathrm{c}} = \boldsymbol{T}_q\boldsymbol{P}_q + \boldsymbol{E}_q \tag{2}$$

式中，$\boldsymbol{X}_q^{\mathrm{c}}$ 为 $\boldsymbol{X}_q$ 的中心化矩阵，$\boldsymbol{m}_q$ 为第 q 类样品的均值向量，$\mathrm{repmat}(\boldsymbol{m}_q, n_q, 1)$ 表示把向量 $\boldsymbol{m}_q$ 作为模块按 n_q 行 1 列重复平铺成一个矩阵，最终，矩阵的行数为 n_q，列数为向量 $\boldsymbol{m}_q$ 的长度，$\boldsymbol{T}_q$，$\boldsymbol{P}_q$ 和 $\boldsymbol{E}_q$ 分别为第 q 类样品主成分分解的得分矩阵、载荷矩阵及残余矩阵(保留 k_q 个主成分). 主成分空间的维数 k_q 根据 Malinowski 因子指示器函数(IND)[7]随主成分数的变化确定.

根据不相交 PCA 的原理，第 r 类中第 i 个样品在第 q 类样品主成分空间中投影的残差可由下式获得：

$$e_{qr,i} = (\boldsymbol{x}_{r,i} - \boldsymbol{m}_q)(\boldsymbol{I} - \boldsymbol{P}_q\boldsymbol{P}_q^{\mathrm{T}}) \tag{3}$$

式中，$\boldsymbol{x}_{r,i}$ 为第 r 类中第 i 个样品的基因表达数据，$\boldsymbol{m}_q$ 和 $\boldsymbol{P}_q$ 的含义同上，T 表示转置.

对于基因芯片数据中的每一个基因，其区分能力可表示为[8]

$$DP_j = (e_{qr,j}^{\mathrm{T}}e_{qr,j} + e_{rq,j}^{\mathrm{T}}e_{rq,j})/(e_{qq,j}^{\mathrm{T}}e_{qq,j} + e_{rr,j}^{\mathrm{T}}e_{rr,j}) \tag{4}$$

式中，q 和 r 分别为 1 或 2，表示样品的类别，$e_{qr,j}$ 表示第 q 类样品在第 r 类样品主成分空间投影残差的第 j 列.

用 s_{qr}^2 表示第 r 类样本在第 q 类主成分空间投影的平均残余方差，即

$$s_{qr}^2 = \frac{1}{(n_r - k_q - 1)(p - k_q)}\sum_{i=1}^{n_r} e_{qr,i}e_{qr,i}^{\mathrm{T}} \tag{5}$$

式中，n_r，k_q 和 p 的含义同上，$e_{qr,i}$ 表示 r 类样本在 q 类主成分空间投影残差的第 i 行，则式(6)中行列式的值 $|\boldsymbol{A}|$ 可表示对两类不同样品的区分能力.

$$|\boldsymbol{A}| = \begin{vmatrix} s_{rr}^2 & s_{rq}^2 \\ s_{qr}^2 & s_{qq}^2 \end{vmatrix} = s_{rr}^2 \times s_{qq}^2 - s_{qr}^2 \times s_{rq}^2 \tag{6}$$

从式(6)可以看出，$|\boldsymbol{A}|$ 表示每类样品在自身主成分空间投影的平均残余方差与在另一类主成分空间投影的平均残余方差的差异. 其值越小，说明两类样品在各自主成分空间投影的残余越小，而在另一类主成分空间投影的残余越大，即不相交 PCA 提取的两类样品特征的差别越大，也表明 $|\boldsymbol{A}|$ 所对应的基因组对两类样品的区分能力越强.

1.2 差异表达基因的识别

差异表达基因识别的目的在于从两类样品基因芯片的实验数据中寻找表达差异较大的基因. 采用上述区分能力的判据, 本文建立了一种基于不相交主成分分析和遗传算法的差异表达基因识别算法, 其识别过程可描述如下:

(1) 先从全体样品中随机选取部分(80%)样品, 并把基因按区分能力(DP)降序排列, 再用 GA 从前 100 个基因中搜寻基因组, 并计算相应的行列式值$|\boldsymbol{A}|$. GA 的染色体长度为 100, 每个位对应于一个基因, 数值为"1"表示对应的基因被选取, "0"则表示不被选取. 该过程共重复执行 2000 次, 通过每次所选取的基因及$|\boldsymbol{A}|$确定显著差异表达的基因.

(2) 将样品的类别随机扰乱, 重复上述选取基因的过程. 此过程称为随机检验, 用于分析所选取基因的偶然相关性. 随机检验共执行 4000 次, 通过每次所选取的基因及$|\boldsymbol{A}|$评估基因的偶然相关性.

(3) 对以上两个过程中选取的基因分别计算选取频率f_i和逆向积累概率p_f, 以随机检验的结果为参照评估偶然相关性, 并确定显著差异表达的基因. f_i和p_f分别定义如下:

$$f_i = N_i/\mathrm{Loop} \tag{7}$$

$$p_f = N_{>f}/N_{>0} \tag{8}$$

式中, N_i是基因 i 被选取的次数, Loop 为选取过程的循环次数, $N_{>f}$和$N_{>0}$分别为选取频率大于 f 和 0 的基因数. f_i表示基因 i 的相对选取次数, p_f表示选取频率大于 f 的基因所占的比例. 式(7)和式(8)的计算结果与基因选取过程的循环次数 Loop 有关, 随着循环次数的增大, 所得结果趋于稳定. 本文分别采用了 2000 和 4000 次. 显然, 在多次基因选取过程中, 某个基因被选取的次数越多, 偶然相关的可能性越小. 而随机检验选取的基因, 无论选取频率有多大, 都是偶然相关的结果. 因此, 随机检验的逆向积累概率 1% 所对应的选取频率$f_{1\%}$, 表示在显著性水平为 0.01 时可能发生偶然相关的选取频率. 对于从实际基因芯片数据中选取的基因, 只要选取频率大于$f_{1\%}$, 就可认定是显著差异表达的基因.

2 基因芯片数据

为了考察本文所提出的差异表达基因识别方法, 采用了 Chen 等[9]报道的一组基因芯片实验数据. 这组数据由 82 个肝细胞癌(HCC)和 74 个非癌的肝组织样品组成, 每个样品的基因表达谱由 3964 个 cDNA 所代表的 3180 个基因组成.

3 结果与讨论

3.1 识别方法的性能

图 1 是根据式(7)和式(8)的计算结果绘制的逆向积累概率图, 横坐标代表基因的选取频率, 纵坐

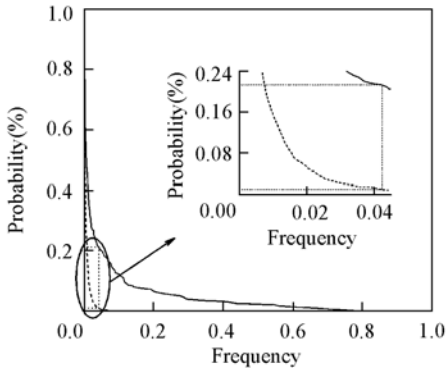

Fig. 1 Reverse cumulative probability *versus* the picking frequency of genes

The solid curve is generated from real data, and the dashed curve from randomized data. The dotted lines are used to determine the number of differentially expressed genes.

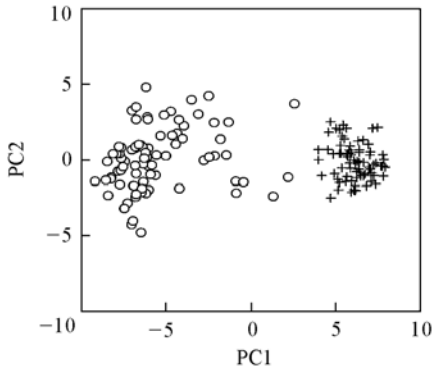

Fig. 2 PCA scores plot of the samples based on the identified 58 differentially expressed genes

"○" Represents the samples with HCC;

"+" represents samples of normal tissue.

标代表逆向积累概率 p_f；虚线来自随机检验，虚线上的每一点表示了相应选取频率下基因发生偶然相关的概率；实线来自实际基因芯片数据的选取结果，实线上的点则表示了相应选取频率的基因在所有选取的基因中所占的比例；点状直线用于确定显著性水平为 0.01 时的差异表达基因. 根据图 1，在显著性水平为 0.01 时，共得到 58 个显著差异表达的基因，表明 58 个识别基因的偶然相关的概率均小于 1%.

为了进一步考察 58 个识别基因的区分能力，对所有样品的 58 个识别基因的表达谱作了主成分分析，取前两个主成分作图，结果如图 2 所示. 从图 2 可以看出，第一主成分即可清晰地把癌症和非癌症两种组织样品完全分开. 由此表明，所提方法识别的显著差异表达的基因可以有效地区分两种不同组织样品. 由于在识别基因的过程中，考虑了基因之间的协同作用，与基因在生物过程中发挥调控作用的实际过程相符，有利于找到真正与疾病相关的基因组.

3.2　与 SAM 方法的比较

SAM(Significance Analysis of Microarrays)方法是美国 Stanford 大学 Tusher 等[4]提出的差异表达基因识别算法(相应软件可以从网站 http://www-stat. stanford. edu/ ~ tibs/SAM/下载). SAM 方法用统计量——相对差异(Relative difference) d_i(类似于 t-检验的统计量 T)来衡量每个基因 i 的表达水平与某种响应量(如不同的癌症类别)之间的相关程度，并采用多次随机扰动基因表达谱数据的方法来判断这种相关关系的显著性. 本文采用 SAM 方法对上述基因芯片数据进行了分析. 采用假阳性率(False Discovery Rate, FDR)[4]小于 0.05 为选择阈值(FDR < 0.05 是基因芯片数据分析中常用的阈值)，共得到 2830 个显著差异表达的基因.

首先，考察了 SAM 方法识别的基因中差异表达最大的 58 个基因与本文方法所识别的 58 个基因之间的重叠情况，发现两种方法识别的大多数基因保持了一致，共有 37 个相同的基因，表明本文方法所识别基因的合理性. 为了比较 SAM 方法识别的基因对两种不同组织样品的区分能力，对 SAM 方法识别的全部基因和差异表达最大的 58 个基因的表达谱作了主成分分析，并分别取前两个主成分作图，结果如图 3 所示.

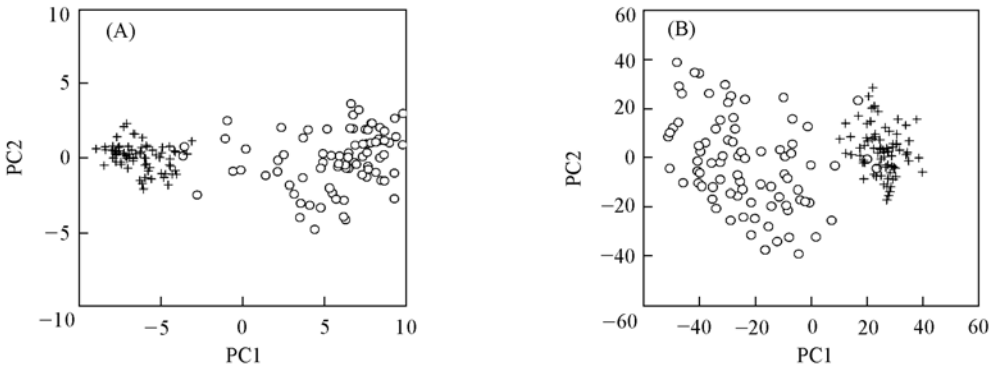

Fig. 3　PCA scores plots of samples based on top 58(A) and 2830 significant genes(B) identified with SAM

由图 3 可以看出，无论是差异表达最大的 58 个基因[图 3(A)]，还是全部显著差异表达的基因[图 3(B)]都不能将两类不同组织样品完全分开. 因此，与 SAM 方法识别的基因相比，本文方法识别的基因具有更强的区分能力. 此外，比较图 3 中的两个基因组可以看出，图 3(B)所示基因组的区分能力更差. 由此表明，SAM 方法识别的显著差异表达的基因中存在大量非疾病相关的基因. 这些基因的存在，影响了所选择基因的区分能力.

参　考　文　献

[1] Allison D. B., Cui X., Page G. P., *et al*.. Nature Reviews Genetics[J], 2006, **7**(1): 55—65

[2] Guan Z., Zhao H.. Bioinformatics[J], 2005, **21**(4): 529—536

[3] Black M. A., Doerge R. W.. Bioinformatics[J], 2002, **18**(12): 1609—1616

[4] Tusher V. G., Tibshirani R., Chu G.. PNAS USA[J], 2001, **98**(9): 5116—5121
[5] Bicciato S., Luchini A., Di Bello C.. Minerva Biotecnologica[J], 2002, **14**(3/4): 281—290
[6] Bicciato S., Luchini A., Di Bello C.. Bioinformatics[J], 2003, **19**(5): 571—578
[7] Malinowski E. R.. Analytical Chemistry[J], 1977, **49**(4): 612—617
[8] Wold S.. Pattern Recognition[J], 1976, **8**: 127—139
[9] Chen X., Cheung S. T., So S., *et al*.. Molecular Biology of the Cell[J], 2002, **13**(6): 1929—1939

Identification of Differentially Expressed Genes Using Disjoint Principal Component Analysis Coupled with Genetic Algorithm

SU Zhen-Qiang[1,3], HONG Hui-Xiao[2], TONG Wei-Da[3*], PERKINS Roger[2],
SHAO Xue-Guang[4], CAI Wen-Sheng[1,4*]

(1. *Department of Chemistry, University of Science and Technology of China, Hefei* 230026, *China*;
2. *Division of Bioinformatics, Z-Tech at FDA's National Center for Toxicological Research, Jefferson, AR* 72079, *USA*;
3. *Center for Toxicoinformatics, National Center for Toxicological Research (NCTR), US Food and Drug Administration (FDA), Jefferson, AR* 72079, *USA*;
4. *Department of Chemistry, Nankai University, Tianjin* 300071, *China*)

Abstract A new method for the feature selection using disjoint principal component analysis (PCA) coupled with genetic algorithm (GA) was proposed and was used to identify differentially expressed genes based on microarray gene expression profiles. The discriminatory power of combination of genes is assessed with using disjoint PCA, the combinatorial optimization problem of genes is solved by using GA, and the chance correlation of genes is assessed by a statistic method. Due to considering the cooperation between genes which is a way to approximate the synergistic regulation by genes during the biological processes, the genes identified by our method are capable of powerful ability to express the differences. This method has been applied to analyze the gene microarray data of hepatocellular caricinoma (HCC). It is found that the genes identified by the proposed method has more discriminatory power in distinguishing two-class samples than those identified by SAM (significance analysis of microarrays), which is very popular in the analysis of microarray data.

Keywords Microarray; Principal component analysis (PCA); Genetic algorithm (GA); Significance analysis of microarrays (SAM); Chance correlation

(Ed.: A, G)