第40卷 第5期
2010年9月
吉林大学学报(工学版)
Journal of Jilin University (Engineering and Technology Edition)
Vol. 40 No. 5
Sept. 2010

基于元分析的差异表达基因识别

刘桂霞,田 原,郑 明,赖丽娜,臧雪柏,周春光

(吉林大学 计算机科学与技术学院,长春 130012)

摘 要:针对传统差异表达基因分析方法使用单一数据集,不能处理异质性特性数据集、分析结果偏差大的问题,提出了联合驱动排除的概念,设计实现一种使用多研究数据集的元分析算法。应用公共数据库GEO中GDS2490和GDS2491数据集对设计的算法与微芯片显著性分析方法进行了对比。实验结果表明:设计的算法可以对数据集实行联合驱动排除,与微芯片显著性分析方法相比,可以有效分析异质性特性的数据集,准确找到差异表达的基因,验证了算法的有效性。该方法为差异表达基因识别提供了新思路。

关键词:人工智能;元分析;联合驱动排除;芯片显著性分析;异质性

中图分类号:TP18 **文献标志码**:A **文章编号**:1671-5497(2010)05-1308-05

Identification of differentially expressed genes based on meta-analysis

LIU Gui-xia, TIAN Yuan, ZHENG Ming, LAI Li-na, ZANG Xue-bai, ZHOU Chun-guang

(*College of Computer Science and Technology, Jilin University, Changchun 130012, China*)

Abstract: Traditional methods of differentially expressed genes analysis used only one single dataset of a study, so they couldn't handle the heterogeneity between studies and lead to inconsistency of analysis results. To overcome the above shortcoming, we proposed the concept of integration-driven exclusion, designed and implemented an algorithm to use several datasets of different studies. In this algorithm we presented a meta-analysis tool to identify differentially expressed genes. Using datasets GDS2490 and GDS2491 from GEO, we compared this algorithm with the method of significance analysis of microarrays. Experiment results show that the designed algorithm can identify differentially expressed genes accurately by integration-driven excluding, handle heterogeneity between studies effectively. This algorithm provides a new approach to identify differentially expressed genes.

Key words: artificial intelligence; meta-analysis; integration-driven exclusion; significance analysis of microarrays; heterogeneity

传统的差异表达基因识别[1-2]技术有微芯片显著性分析SAM(Significance analysis of microarrays)等方法,它们是对单一的研究数据集分析基因的差异表达,难以应用在基因数量巨大而样本数量很小的情况[3],因为它们不能有效分析具有异质性特性的研究数据集,分析结果偏差

收稿日期:2009-09-24.

基金项目:国家自然科学基金项目(60973092,60673099,60873146);"863"国家高技术研究发展计划项目(2009AA02Z307).

作者简介:刘桂霞(1963-),女,教授.研究方向:计算智能和生物信息学.E-mail:liugx@jlu.edu.cn

通信作者:臧雪柏(1963-),女,高级工程师.研究方向:网格计算.E-mail:xbzang@yahoo.com.cn

较大。本文设计实现了基于元分析基因差异表达识别的算法,其优点为:①采用标准的统计量效能,可以广泛地用于各种研究数据集,各个独立的不同形式研究统一于效能,使得对于多个研究分析成为可能;②元分析中,计算效能有固定模型(Fixed-effects model, FEM)和随机模型(Random-effects model,REM)两种模型,针对不同的问题,可以有不同的选择;③能有效分析具有异质性特性的研究数据集。

1　原理和算法思想

元分析以估计整体效果为目的,总结与综合各种研究。Mehra 等[4]最早证明了使用元分析程序分析芯片的可行性,自此元分析成为癌症临床数据分析的标准统计方法[5-6],一些组织的研究者也把元分析应用于分析基因表达芯片[7-8]。

效能 d_i 是元分析的核心,它是元分析的标准统计量[9]。在元分析中,效能主要用来比较实验组与对照组基因表达值的差别,计算式为

$$d_i = (\overline{x_t} - \overline{x_c})/s_p \tag{1}$$

式中:$\overline{x_t}$ 和 $\overline{x_c}$ 分别为研究 i 中实验组和对照组基因表达值均值;s_p 为实验组和对照组基因表达值的联合标准差。

式(1)是基于 t 检验的元分析。

元分析基于如下假设:每个独立的研究数据集共享一个共同的效能值 μ,但每个研究的观测效能值不一样,主要受选样误差 ε_i 和不同研究间的差异 δ_i 的影响。元分析的假设使用层次模型,表示为

$$d_i = \theta_i + \varepsilon_i,\ \varepsilon_i \sim N(0, s_i^2) \tag{2}$$

$$\theta_i = \mu + \delta_i,\ \delta_i \sim N(0, \tau_i^2) \tag{3}$$

式中:τ_i^2 为研究间的方差,代表不同研究间的差异;s_i^2 为研究中的方差,代表选样误差。

$$s_i^2 = (n_t^{-1} + n_c^{-1}) + d'^2_i[2(n_t + n_c)]^{-1} \tag{4}$$

$$d'_i = 3d_i/[4(n-2)-1] \tag{5}$$

式中:n 为研究 i 中样本总个数;n_t、n_c 分别为研究 i 中实验组样本个数与对比组样本个数。

基于以上假设,提出两种模型:固定模型只考虑同一个研究,不同样本间的差异(即 $s_i^2 \neq 0$),而忽略各个研究间的差异(即 $\tau_i^2 = 0$);随机模型是同时考虑到两种差异的存在(即 $s_i^2 \neq 0, \tau_i^2 \neq 0$)。选择固定模型还是随机模型依赖统计量 Q,其计算公式为

$$Q = \sum w_i(d_i - \hat{\mu}) \tag{6}$$

$$w_i = s_i^{-2} \tag{7}$$

$$\hat{\mu} = (\sum w_i d_i)/\sum w_i \tag{8}$$

当统计量 Q 符合卡方分布时,选用固定模型,反之使用随机模型,卡方的自由度为 $n-1$。使用 Q-Q 图显示统计量 Q 是否符合卡方分布,当统计量 Q 符合卡方分布时,基因点分布在图像的平分线上,反之与平分线分离。

合并效能 d_i,表示为 $\hat{\mu}(\tau_i^2)$,公式为

$$\hat{\mu}(\tau_i^2) = \frac{\sum (s_i^2 + \tau_i^2)^{-1} d_i}{\sum (s_i^2 + \tau_i^2)^{-1}} \tag{9}$$

合并方差 s_i^2,表示为 $Var[\hat{\mu}(\tau_i^2)]$,公式为

$$Var[\hat{\mu}(\tau_i^2)] = \frac{1}{\sum (s_i^2 + \tau_i^2)^{-1}} \tag{10}$$

$$\tau_i^2 = \max\left[0, \frac{Q-(n-1)}{S_1 - (S_2/S_1)}\right] \tag{11}$$

$$S_r = \sum w_i^r \tag{12}$$

判断差异表达基因使用统计量 Z,Z 为 $\hat{\mu}(\tau_i^2)$ 与它的标准误差的比值。定义一个阈值 Z_{th},且阈值大于零。当 $Z \geqslant Z_{th}$ 时,基因为上调基因;$Z \leqslant -Z_{th}$ 时,基因为下调基因[10]。

2　算法实现过程及流程图

本文研究的算法可以处理来源于任何基因数据库的数据,对数据的格式要求如下:

(1)形成文本文档。

(2)文本格式中列与列之间用 tab 隔开,第一列为基因 ID,第二列为基因名字,余下列为基因不同条件下的表达值。

(3)最后一行只写 end。

(4)把要分析的各个研究中相同的基因的数据存入数据库 MySql 中,供分析时使用。

算法实现过程描述如下:

(1)数据由数据库抽取,录入 Class gene (gene 类,用于存放基因名称、基因实验组与对比组样本个数、基因表达值、Q 和 Z 等信息)中,计算 Q 值、固定模型下的 Z_f 值和随机模型下的 Z_r 值。

(2)各个基因形成基因链 LinkedList geneslist。

(3)由 geneslist 显示各个基因的名字及基因的不同研究的不同条件下的表达值,并且生成一

个符合卡方分布的随机数列,数列大小与基因个数相同,把数列与 Q 值数列比较,在 Q-Q 中显示出来。根据 Q-Q 图判断 Q 值是否符合卡方分布,以此确定使用固定模型还是随机模型。

(4)选择模型,输入阈值 Z_{th} 和随机扰动的次数 B。本算例中选择阈值为 1.7,随机扰动次数为 300。

(5)根据随机扰动的次数进行随机扰动,得到期望值 Z_e。由观察值 Z_o(固定模型下的 Z_f 值或随机模型下的 Z_r 值)、期望值 Z_e 和阈值 Z_{th} 形成图像在页面中显示。

(6)数据分析完毕。结束条件为基因链中每一个基因都进行 B 次随机扰动。

算法实现流程图如图 1 所示。

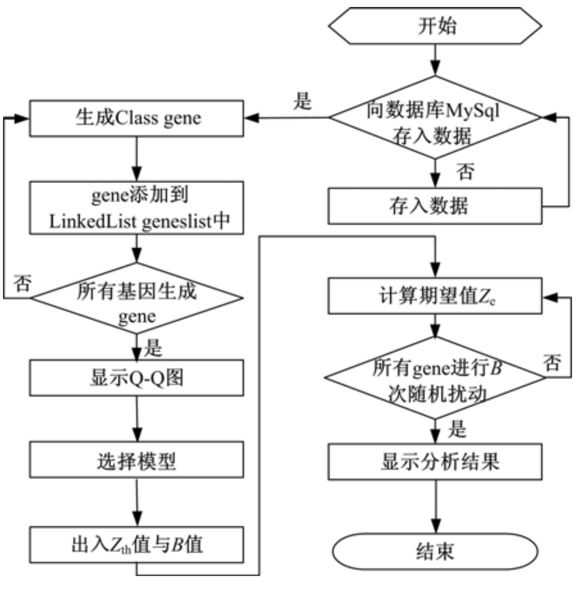


图 1 算法流程图

Fig. 1 Flow chart of the algorithm

3 实验与分析

本文定义综合驱动排除(Integration-driven exclusion,IDE)的概念。IDE 表示那些没有被元分析发现的、但对其使用的数据集单独分析却被发现的差异表达基因。考虑一种极端情况:两个研究的数据集 DS1 和 DS2,一个基因 g 在数据集 DS1 中是上调基因,在数据集 DS2 中却是下调基因,且 $d_{g1}=-d_{g2}$,其他的统计量都相同。经过元分析 $z_g=0$,则 g 不是差异表达基因,即基因 g 是综合驱动排除的。本文作如下假设:①产生 IDEs 的原因是不同研究数据集的差异;②保守地认为 IDEs 是伪差异表达基因;③单独分析数据集,把伪差异表达基因当成差异表达基因的情况是一种研究数据集异质性的表现。在进行差异表达基因识别时,常常发现的差异表达基因数目庞大,但是其中经过证实的差异表达基因只是其中很小的一部分,即存在大量的伪表达基因。其中 IDEs 是伪表达基因的一种,排除了它就可以更准确地发现真正差异表达基因,从而提高算法的准确性。

SAM 算法是美国 Stanford 大学 Tusher 等[11]提出的基因差异表达识别算法。用统计量来衡量每个基因的表达水平与某种响应量之间的相关程度,并采用多次随机扰动的方法判断相关程度的显著性。本文使用肺部大气管上皮细胞基因表达值作为实验数据,用于对比 SAM 方法与元分析在处理具有异质性特性的研究数据集方面的能力。

肺部癌症肿瘤显示神经内分泌特性,长期吸烟者神经分泌细胞数量增加。本文分析的数据是肺部大气管上皮细胞基因在吸烟者与非吸烟者中不同的表达[12]。数据来源于 Gene expression omnibus,分别是数据集 GDS2490 和数据集 GDS2491。算例选择从基因编号 1007_s_at 到 204000_s_at 的基因形成文档 90-5-6.txt 与 91-4-5.txt。两个文档中相同的基因有 2106 个。文档 90-5-6.txt 为实验 1 的数据集 DS1,文档 91-4-5.txt 为实验 2 的数据集 DS2。本文用探针名作为与其相对应基因的名称。

通过 Q 值数列与卡方数列比较的 Q-Q 图如图 2 所示。可以看到,Q 值数列不服从卡方分布,且差距较大,应该选择随机模型合并效能及其方差。当取阈值为 1.7($p<0.001$, p 为 P-value)、随机扰动次数为 300 时,上调基因 2 个,下调基因

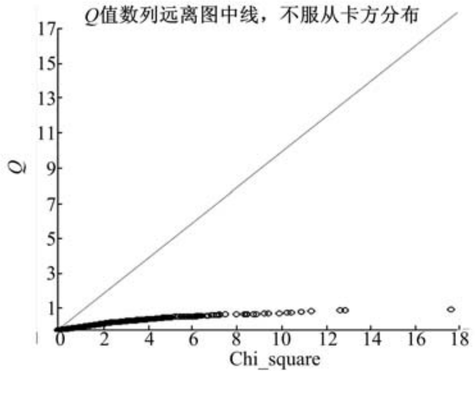


图 2 Q-Q 图

Fig. 2 Quantile-quantile plots

5 个，分析差异表达基因的算法采用假发现率(False discovery rate，FDR)来描述分析得到伪差异表达基因的概率，FDR 是评价识别基因差异表达精度的标准。本实验 FDR＝29％，如图 3 所示。差异表达基因共 7 个，如表 1 所示，但是其中有 2 个是伪差异表达基因。FDR 表达式为

$$\mathrm{FDR} = \frac{(1/B)\sum_{b}\sum_{(i)} \mathrm{I}(\mid Z_{(i)}^{*b} \mid \geqslant Z_{\mathrm{th}})}{\sum_{(i)} \mathrm{I}(\mid Z_{(i)} \mid \geqslant Z_{\mathrm{th}})} \tag{13}$$

式中：$Z_{(i)}^{*b}$ 为第 b 次随机扰动中第 i 个基因的 Z 值；函数 I 在比较为真时等于 1，在为假时等于 0。P-value 计算公式为

$$p = (1/GB)\sum_{b}\sum_{(i)} \mathrm{I}(\mid Z_{(i)}^{*b} \mid \geqslant Z_{\mathrm{th}}) \tag{14}$$

式中：G 为基因的总数[8]。

本文定义 FDN 为比较实验统一的比较条件，FDN 的表达式为

$$\mathrm{FDN} = \sum_{b}\sum_{(i)} \mathrm{I}(\mid Z_{(i)}^{*b} \mid \geqslant Z_{\mathrm{th}}) \tag{15}$$

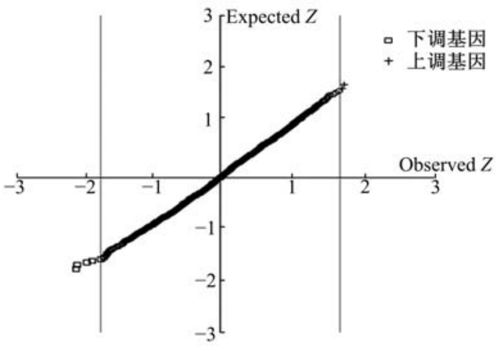


图 3　差异表达基因的发现

Fig. 3　Identification of differentially expressed genes

表 1　差异表达基因

Table 1　Differentially expressed genes

基因名称	$\hat{\mu}(\tau_i^2)$	$Var[\hat{\mu}(\tau_i^2)]$	Z	上/下调
SOX9	0.92	0.56	1.73	上
PLK2	0.92	0.55	1.75	上
GPX2	−1.62	1.26	−2.04	下
TALDO1	−1.84	1.63	−2.03	下
NQO1	−1.21	0.81	−1.90	下
ABCC1	−1.06	0.68	−1.82	下
FTH1	−0.83	0.48	−1.70	下

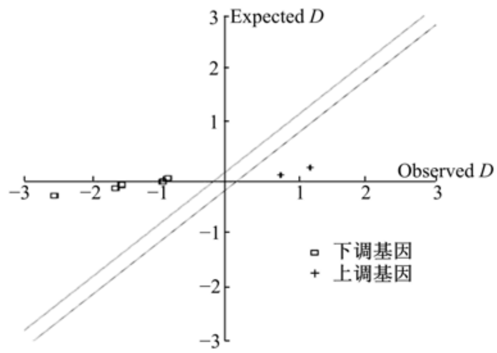
从数据库中找出实验中 7 个差异表达基因的数据，用 GDS2490 中的数据形成 90.txt 文本，用 GDS2491 中的数据形成 91.txt 文本。然后，将 90.txt 与 91.txt 输入进数据库，作为元分析使用的研究 3 的数据集 DS3 和研究 4 的数据集 DS4。阈值为 1.7，分析发现，7 个基因中上调基因 2 个，下调基因 3 个，不差异表达的基因 2 个，即 IDEs，如表 2 所示。FDR＝0.13％，即 FDN＝195。

表 2　表达基因

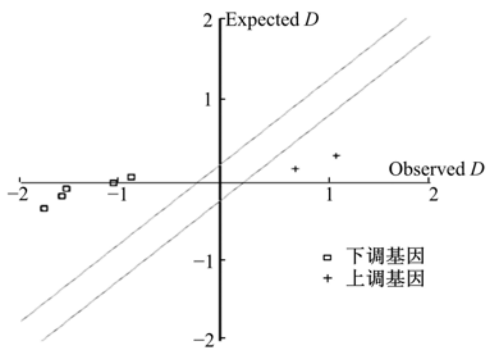
Table 2　Expressed genes

基因名称	$\hat{\mu}(\tau_i^2)$	$Var[\hat{\mu}(\tau_i^2)]$	Z	上/下调
SOX9	0.92	0.56	1.73	上
PLK2	0.92	0.54	1.75	上
GPX2	−1.62	1.26	−2.04	下
TALDO1	−1.84	1.63	−2.03	下
NQO1	−1.21	0.81	−1.90	下
ABCC1	−0.47	0.36	−1.11	不
FTH1	−0.45	0.290	−0.11	不

用 SAM 方法分析数据集 DS3，在 FDN＝195 时，全部 7 个基因都为差异表达基因，上调基因 2 个，下调基因 5 个，如图 4(a)所示。用 SAM 方法分析数据集 DS4，在 FDN＝195 时，7 个基因均为差异表达基因，上调基因 2 个，下调基因 5 个，如图 4(b)所示。



(a)



(b)

图 4　差异表达基因的发现

Fig. 4　Identification of differentially Expressed genes

对比结果表明,IDEs 真实存在,SAM 方法由于使用单一研究数据集分析差异表达基因,不能发现 IDEs,难以有效分析异质性研究数据集。元分析由于使用多个不同的研究数据集分析差异表达基因,可以有效地分析异质性研究数据集,发现使用单一数据集难以发现的 IDEs,从而更准确地分析差异表达基因。

4 结束语

本文提出了基于元分析算法的实现,该算法综合多个独立研究的实验数据,计算效能及其联合方差,并且选用合理的模型合并效能及其方差,通过与阈值的比较找出差异表达的基因。与传统的 SAM 方法比较,元分析算法更加准确、高效,同时能够有效地分析具有异质性特性的研究数据集,为研究差异表达基因识别提供了新的手段。

参考文献:

[1] Bernthaler A, Mühlberger I, Fechete R,et al. A dependency graph approach for the analysis of differential gene expression profiles[J]. Mol Biosyst,2009,5(12):1720-1731.

[2] Pounds Stan, Rai Shesh N. Assumption adequacy averaging as a concept for developing more robust methods for differential gene expression analysis [J]. Computational Statistics and Data Analysis ,2009,53:1604-1612.

[3] Wu Xiao-yan,Wu Zhen-yu,Li Kang. Identification of differential gene expression for microarray data using recursive random forest[J]. Chin Med J ,2008,121(24):2492-2496.

[4] Mehra R, Varambally S, Ding L,et al. Identification of GATA3 as a breast cancer prognostic marker by global gene expression meta-analysis[J]. Cancer Res,2005,65:11259-11264.

[5] Egger M, Smith G D. Meta-analysis. Potentials and promise[J]. Bmj, 1997, 315:1371-1374.

[6] Rhodes D R, Yu J, Shanker K,et al. Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression[C]//Proceedings of the National Academy of Sciences of the United States of America, USA,2004.

[7] Schneider J, Ruschhaupt M, Buness A,et al. Identification and meta-analysis of a small gene expression signature for the diagnosis of estrogen receptor status in invasive ductal breast cancer[J]. Int J Cancer,2006,119(12):2974-2979.

[8] Rhodes D R, Barrette T R, Rubin M A,et al. Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer[J]. Cancer Res, 2002,62:4427-4433.

[9] Jung Kyoon Choi, Ungsik Yu, Sangsoo Kim,et al. Combining multiple microarray studies and modeling interstudy variation[J]. Bioinformatics,2003,19:84-89.

[10] Smith D D, Sætrom P, Snφve Jr,et al. Meta-analysis of breast cancer microarray studies in conjunction with conserved cis-elements suggest patterns for coordinate regulation[C]//Proceedings of the National Academy of Sciences of the United States of America,USA,2001.

[11] Tusher V G, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response[J]. Proc Natl Acad Sci U S A , 2001,98:5116-5121.

[12] Carolan B J, Heguy A, Harvey B G,et al. Up-regulation of expression of the ubiquitin carboxyl-terminal hydrolase L1 gene in human airway epithelium of cigarette smokers[J]. Cancer Res, 2006,66(22):10729-10740.