Hans Journal of Computational Biology 计算生物学, 2011, 1, 22-26
http://dx.doi.org/10.12677/hjcb.2011.12004 Published Online December 2011 (http://www.hanspub.org/journal/hjcb)

Likelihood Ratio-Permutation Test of Differentially Expressed Cancer-Related Genes

Jingzhe Li, Guofen Zhang
Department of Mathematics, Zhejiang University, Hangzhou
Email: zhang-hh@zju.edu.cn
Received: Sep. 18th, 2011; revised: Oct. 23rd, 2011; accepted: Oct. 25th, 2011.

Abstract: The particularity of the differentially expressed cancer-related genes brings new challenges in the field of gene selection. Statisticians have succeeded in proposing new statistics and methods to solve the problem of selecting differentially expressed cancer gene. This article will advance a new method called "Likelihood Ratio-Permutation Test", in order to select differentially expressed genes under the cancer model, after referencing some existing research.

Keywords: DNA Microarray (Gene Chips); Gene Selection; Likelihood Ratio Test; Permutation Test

癌症差异表达基因的似然比－置换检验法

李靖喆，张恒奋
浙江大学数学系，杭州
Email: zhang-hh@zju.edu.cn
收稿日期：2011 年 9 月 18 日；修回日期：2011 年 10 月 23 日；录用日期：2011 年 10 月 25 日

摘　要：癌症的差异表达基因具有其特殊性，这种特殊性给基因选择带来了新的挑战。许多统计学家提出了新的统计量和检验方法，不断地在这一领域取得突破和完善。本文将在借鉴这些已有研究成果的基础上，运用统计学中两种经典的常用方法，提出一种有效的手段——似然比－置换检验法，用以甄选癌症的差异表达基因。

关键词：DNA 微阵列(基因芯片)；基因选择；似然比检验；置换检验

1. 引言

差异表达基因的检测问题一直以来在生物科学领域中占据着重要的地位。近年来，DNA 微阵列技术(microarray)——又称为基因芯片(gene chip)取得了令人瞩目的发展。这项技术可以在同一时间对大量基因表达水平精确、快速的分析并予以记录，多以矩阵形式$\left(x_{ij}\right)_{n\times p}$表示。其中 n 代表样本容量，p 为基因数量，通常有 $p \gg n$ [1]。

鉴于此，人们往往需要从成千上万个基因中筛选出与某一特定疾病相关的基因，从而通过控制这些致病基因，以达到控制疾病的目的[2]。那些在患病者的样本和正常人样本上表达有显著差异的基因，很自然的被认为同这一疾病有着明显的关系，即为致病基因。

传统的单个基因两样本 t 检验被广泛地运用于差异表达基因的检验。然而，Tomlins 等人在对前列腺癌症(Prostate Cancer)的差异基因表达的研究中发现 t 检验并没有取得所预期的效果[3]。于是，他们在对基因芯片进行细致分析之后指出，致病基因并非在所有实验组样本中，而仅仅是在其一个子集上都显示出差异表达的特性。而在此模型下，实验组与对照组的样本均值的差异往往不那么明显。

D. Ghosh 等人在 2008 年提出了上述差异表达的概率分布模型[4]。他们假设致病基因在对照组全体和实验组的一个子集上服从同一分布 F_0，而在实验组的其他样本上服从另一分布 F_1，即：

$$x_1, x_2, \cdots, x_m \sim F_0$$

$$y_1, y_2, \cdots, y_n \sim \pi_0 F_0 + (1-\pi_0) F \tag{1}$$

其中 $0 < \pi_0 < 1$，且 $F_0 \neq F$

令 $y_1 \le y_2 \le \cdots \le y_n$，我们可以考虑对于某个 k，$y_{k+1}, y_{k+2}, \cdots, y_n$ 来自于不同的分布，$1 < k < n$。其中的 $y_{k+1}, y_{k+2}, \cdots, y_n$，即实验组数据中那些差异表达的部分，通常被称作异常值(outlier)。2005 年 Tomlins 等人提出的 COPA 统计量[3]，R. Tibshirani 等在 2007 年提出的 OS 统计量[5]以及 H. Lian 等在 2008 年提出的 MOST 统计量[6]都在一定程度上改进了 t 统计量的不足，并取得了良好的效果。

然而，上述方法仍然面临着一些共同的缺陷。其中重要的一点是，上述方法所用到的统计量大都是基于分位数来建立的，而分位数的选择具有相当大的主观性。鉴于我们事先往往并不知道异常值的个数或所占的比例，故很难用一个普遍的标准去代表所有的模型。此外，鉴于分位数和中位数的分布难以获得，这样的方法也不能提供有效的论证来阐明其检验水平 α。

本文中，我们将基于第一部分所述的模型，给出一种以似然比检验为基础的差异基因判别法。

2. 方法

似然比检验是统计学中的一个经典而常见的方法。它常用于区分某一批样本是来源于两个已知分布中的哪一个[7]。

本文最终的目的是要对于单个基因 g，做如下检验：

$$\begin{aligned} &H_0'': \text{基因} g \text{非致病基因} \\ &\leftrightarrow H_1'': \text{基因} g \text{是致病基因} \end{aligned} \tag{2}$$

设基因 g 在 m 个对照组样本和 n 个实验组样本上的基因表达水平分别为 $(x_1, x_2, \cdots, x_m)$ 和 $y_1, y_2, \cdots, y_n$。假设差异表达部分要普遍高于正常值，我们记：

$$(z_1, z_2, \cdots, z_{m+n}) = (x_1, x_2, \cdots, x_m, y_1, \cdots, y_n).$$

于是，上述 H_1'' 成立，当且仅当 $x_1, x_2, \cdots, x_m$ 和 $y_1, y_2, \cdots, y_n$ 服从(1)中的分布；而 H_0'' 成立则等价于 $z_1, z_2, ..., z_{m+n} \sim F_0$。亦即我们只需检验：

$$\begin{aligned} &H_0': z_1, z_2, \cdots, z_{m+n} \sim F_0 \\ &\leftrightarrow H_1': x_1, x_2, \cdots, x_m \sim F_0, y_1, y_2, \cdots, y_n \\ &\sim \pi_0 F_0 + (1-\pi_0) F_1, 0 < \pi_0 < 1, F_0 \neq F_1. \end{aligned} \tag{3}$$

至此，检验问题转化为样本 $z_1, z_2, \cdots, z_{m+n}$ 来自于哪一个分布的问题。为简单起见，我们进一步假设 F_0 和 F_1 均为正态分布，且方差相等。即：

$$F_0 \sim N(\mu_0, \sigma^2), F_1 \sim N(\mu_1, \sigma^2)$$

这样，上述假设检验问题可继续转化为：

$$\begin{aligned} &x_1, x_2, \cdots, x_m \sim N(\mu_0, \sigma^2), \\ &y_1, y_2, \cdots, y_n \sim \pi_0 N(\mu_0, \sigma^2) + (1-\pi_0) N(\mu_1, \sigma^2), \\ &0 < \pi_0 < 1. \\ &H_0: \mu_0 = \mu_1 \leftrightarrow H_1: \mu_0 < \mu_1. \end{aligned}$$

接着，构造如下似然比统计量：

$$\lambda = \frac{\sup_{\theta \in \Theta} \prod_{i=1}^{m+n} f_\Theta(z_i)}{\sup_{\theta \in \Theta_0} \prod_{i=1}^{m+n} f_{\Theta_0}(z_i)} \tag{5}$$

其中，待估参数 $\theta = (\mu_0, \mu_1, \sigma, \pi_0)$；$\Theta = \{\theta\}$ 为全参数空间，$\Theta_0 = \{\theta \mid \mu_0 = \mu_1\}$ 为 H_0 成立时的子参数空间，$f_\Theta(z_i), f_{\Theta_0}(z_i)$ 分别为参数空间 Θ，Θ_0 下 z_i 的概率密度函数，$i = 1, 2, \cdots, m+n$。

当 $\theta \in \Theta$ 时，有如下的：

$$f_\Theta(z_i) = \begin{cases} \dfrac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\dfrac{(z_i-\mu_0)^2}{2\sigma^2}\right\}, i = 1, \cdots, m \\ \pi_0 \cdot \dfrac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\dfrac{(z_i-\mu_0)^2}{2\sigma^2}\right\} + \\ \qquad (1-\pi_0) \cdot \dfrac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\dfrac{(z_i-\mu_1)^2}{2\sigma^2}\right\}, \\ i = m+1, \cdots, m+n \end{cases} \tag{6}$$

对(6)式中的各待估参数 $\mu_0, \mu_1, \sigma, \pi_0$ 分别求其极大似然估计(*MLE*)，代入上式即可得到 $\theta \in \Theta$ 时的极大似然函数值。但是这些极大似然估计难以得到显式表达，因此接下来需要对其作出一个近似的估计。

这里不妨用异常值(outlier)个数的估计来代替对 π_0 的估计。这时，令 $y_1 \le y_2 \le \cdots \le y_n$，当其中异常值的个数为 $n-k$ 时，有：

$$f_\Theta(z_i) = \begin{cases} \dfrac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\dfrac{(z_i-\mu_0)^2}{2\sigma^2}\right\}, \\ i = 1, \cdots, m+k \\ \dfrac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\dfrac{(z_i-\mu_1)^2}{2\sigma^2}\right\}, \\ i = m+k+1, \cdots, m+n \end{cases} \tag{7}$$

其中，$1<k<n-1$。进而对于新的待估参数 $\theta^*=(\mu_0,\mu_1,\sigma,k)$，有：

$$\sup_{\theta^*\in\Theta}\prod_{i=1}^{m+n}f_{\Theta}(z_i)=\sup_{k}\prod_{i=1}^{m+k}\frac{1}{\sqrt{2\pi}\hat{\sigma}}\exp\left\{-\frac{(z_i-\hat{\mu}_0)^2}{2\hat{\sigma}^2}\right\}\cdot\prod_{i=m+k+1}^{m+n}\frac{1}{\sqrt{2\pi}\hat{\sigma}}\exp\left\{-\frac{(z_i-\hat{\mu}_1)^2}{2\hat{\sigma}^2}\right\} \quad (8)$$

其中，$\hat{\mu}_0,\hat{\mu}_1,\hat{\sigma}$ 分别为 μ_0,μ_1,σ 在 Θ 中的极大似然估计。经过简单计算得：

$$\begin{aligned}&\hat{\mu}_0=\frac{1}{m+k}\sum_{i=1}^{m+k}z_i \qquad \hat{\mu}_1=\frac{1}{n-k}\sum_{i=m+k+1}^{m+n}z_i\\&\hat{\sigma}^2=\frac{1}{m+n-2}\left[\sum_{i=1}^{m+k}(z_i-\hat{\mu}_0)^2+\sum_{i=m+k+1}^{m+n}(z_i-\hat{\mu}_1)^2\right],\\&\hat{\sigma}=\sqrt{\hat{\sigma}^2}\end{aligned} \quad (9)$$

当 $\theta^*\in\Theta_0$，即 $\mu_0=\mu_1=\mu$ 时，有：

$$f_{\Theta_0}(z_i)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{(z_i-\mu)^2}{2\sigma^2}\right\},\quad i=1,2,\cdots,m+n \quad (10)$$

这时易得：

$$\sup_{\theta^*\in\Theta_0}\prod_{i=1}^{m+n}f_{\Theta_0}(z_i)=\prod_{i=1}^{m+n}\frac{1}{\sqrt{2\pi}\hat{\sigma}_{\Theta_0}}\exp\left\{-\frac{(z_i-\hat{\mu}_{\Theta_0})^2}{2\hat{\sigma}_{\Theta_0}{}^2}\right\} \quad (11)$$

类似上面的推导，有：

$$\begin{aligned}&\hat{\mu}_{\Theta_0}=\frac{1}{m+n}\sum_{i=1}^{m+n}z_i\\&\hat{\sigma}^2_{\Theta_0}=\frac{1}{m+n-1}\left[\sum_{i=1}^{m+n}(z_i-\hat{\mu}_{\Theta_0})^2\right],\\&\hat{\sigma}_{\Theta_0}=\sqrt{\hat{\sigma}^2_{\Theta_0}}\end{aligned} \quad (12)$$

将(7)~(12)代入(5)，可计算出似然比统计量 λ 的值。

接下来，我们对实验组和对照组的共 $m+n$ 个数据重新排序，得到 $(z_1,z_2,\cdots,z_{m+n})$ 的一组置换样本 $(z_1',z_2',\cdots,z_{m+n}')$，采用上节中的方法计算出相应的 λ_1。利用数学软件，这样的置换我们可以进行 N 次，并将得出的 N 个 λ 值记作 $\lambda_1,\lambda_2,\cdots,\lambda_N$。利用这些 $\lambda_1,\lambda_2,\cdots,\lambda_N$ 给出在 H_0 成立时 λ 在 H_0 下的经验分布函数，以模拟 λ 在 H_0 下的近似分布。对于给定的检验水平 α，记 $\lambda_1,\lambda_2,\cdots,\lambda_N$ 中大于 λ 的值的个数为 M。当 $M<N\alpha$ 时，表明 λ 在该近似分布中位于上方 α 比例的范围内。这是便根据假设检验的原则拒绝原假设 H_0，接受 H_1，认为所检验的基因 g 是致病基因。

3. 模拟和实例

首先，我们利用数学软件进行模拟实验。假设 $m=n=20$，模拟下列 5 种形式的样本：

(a) $x_1,x_2,\cdots,x_{20}$ $\quad y_1,y_2,\cdots,y_{20}\sim N(0,1)$

(b) $x_1,x_2,\cdots,x_{20}$ $\quad y_1,y_2,\cdots,y_5\sim N(0,1)$
$y_6,y_7,\cdots,y_{20}\sim N(2,1)$

(c) $x_1,x_2,\cdots,x_{20}$ $\quad y_1,y_2,\cdots,y_{10}\sim N(0,1)$
$y_{11},y_{12},\cdots,y_{20}\sim N(2,1)$

(d) $x_1,x_2,\cdots,x_{20}$ $\quad y_1,y_2,\cdots,y_{15}\sim N(0,1)$
$y_{16},y_{17},\cdots,y_{20}\sim N(2,1)$

(e) $x_1,x_2,\cdots,x_{20}$ $\quad y_1,y_2,\cdots,y_{15}\sim N(0,1)$
$y_{16},y_{17},\cdots,y_{20}\sim N(4,1)$

每组数据产生后，首先计算 λ，然后进行 $N=1000$ 次置换，得出 $\lambda_1,\lambda_2,\cdots,\lambda_{1000}$。其中，大于 λ 的值的个数记作 M。以上 5 组数据各进行 100 次模拟，得到 $M_1,M_2,\cdots,M_{100}$。表 1 给出这 100 个 M 的频数统计 f_M 和频率 p_M。

从表 1 中我们可以看出：在实验(a)中，拒绝 H_0 的概率约为 5%。对于非致病基因的检测结果是比较合理的。在实验(b,c)中，当 $F_0=N(0,1),F_1=N(2,1)$ 且异常值的个数不少于一半时，M 的值在多数情况下小于 50，“似然比—置换检验法”较为有效地判别出差异表达基因。而对于实验(d)来说，当 $F_0=N(0,1)$, $F_1=N(2,1)$而异常值数目仅为 5 个时，只有约一半的 M 小于 50。可是从相对于(d)拉大 F_0 和 F_1 均值之间差距(方差相对不变)的实验(e)来看，检验又基本恢复了有效性。造成这一现象的原因可能有如下两点：1)模拟数据的固有误差。这里 F_0 和 F_1 两个正态分布的均值，在实验(a~d)中差别相对较小，而在(e)中相对较大；2)置换检验本身的误差。当原先实验组数据 $(y_1,y_2,\cdots,y_n)$ 中有 k 个 $(y_{n-k+1},y_{n-k+2},\cdots,y_n)$ 属于异常值时，如果在一次置换中这 k 个数据的绝大多数甚至全部恰好被分配到实验组，这一次得出的新似然比 λ_j 会显著偏大. 当 $k=5$ 时，异常值 $(y_{16},y_{17},\cdots,y_{20})$ 均被分配到新的实验组 $(y_1',y_2',\cdots,y_n')$ 中的概率约为 2.4%，是一个不容忽视的概率值。综合以上两个原因，便不难解释为何实验(d)有超过一半的模拟造成了 $M\geq 50$ 的结果。

Table 1. Results of the simulations
表 1. 模拟实验的结果

组别	频数(频率)			Min$M$	Max$M$	Med$M$	$\overline{M}$
(a)	$0 \le M < 10$	$10 \le M < 50$	$M \ge 50$	4	985	440	473.75
	1(1%)	6(6%)	93(93%)				
(b)	$0 \le M < 10$	$10 \le M < 50$	$M \ge 50$	0	18	0	1.91
	93(93%)	7(7%)	0(0%)				
(c)	$0 \le M < 10$	$10 \le M < 50$	$M \ge 50$	0	134	3	11.38
	67(67%)	29(29%)	4(4%)				
(d)	$0 \le M < 10$	$10 \le M < 50$	$M \ge 50$	0	419	57.5	87.28
	10(10%)	38(38%)	52(52%)				
(e)	$0 \le M < 10$	$10 \le M < 50$	$M \ge 50$	3	89	21	21.86
	18(18%)	77(77%)	5(5%)				

注：(a)~(e)各进行 100 次数据模拟，每次模拟进行 1000 组置换并反馈 M. Min$M$、Max$M$、Med$M$、 $\overline{M}$ 分别代表 M 的最小值、最大值，中位数和均值。

接下来，我们将似然比—置换检验法运用于一个具体的基因芯片上以检验其效果。这里采用的是 West 等人在 2001 年给出的 hu6800 基因芯片数据[8]。该芯片包含了 49 个乳腺癌样本上的 7129 个基因表达信息。对于这 49 个样本的淋巴结的观测得知，其中 24 个样本有阳性反应的淋巴结(LN+)，而另外 25 个则全部为阴性(LN–)。我们的目的则是从 7129 个基因中选择造成(LN+)的关键基因。具体做法如下：

1) 首先对于每一单个基因 j 计算出似然比 λ_j 的值，所有基因均做 200 次置换，结果记为 $\lambda'_{j1}, \lambda'_{j2}, \cdots, \lambda'_{j,200}, j = 1,2,\cdots,7129$。记其中大于 λ_j 的个数为 M'_j。首先选择那些 $M'_j < 20$ 的基因 j，这样的 j 共 324 个，记作 $j_k, k = 1,2,\cdots,324$。

2) 对第一步中初选的基因 $j_k, k = 1,2,\cdots,324$ 重新做 800 次置换，类似的得出 $M''_{j_k}, k = 1,2,\cdots,324$。令 $M_{j_k} = M'_{j_k} + M''_{j_k}$，作为全部 1000 次置换检验的结果中大于 λ_{j_k} 的个数。$M_{j_k} < 50$ 的基因被选择。这样的基因共 166 个，其中满足 $M_{j_k} < 10$ 的基因 34 个。

表 2 列举了部分被其他各种方法检验出，并被生物学上证实确实属于差异表达的基因，意在将本方法与其他方法做出比较。此外，表 3 给出的基因，现存的方法均没有检出的报告，但却在基因描述中，抑或是有文献指出，该基因同癌症有一定的联系。它们是否确实属于致病基因，仍有待生物学上的进一步证实。

Table 2. Result-contrasts between the method above and others
表 2. 本文方法与已知其他方法检验结果的对比

基因名称	参考排名	M
ATM	*ORT*/7,*LRS*/10	*5
THRA	*ORT*/17	243
*SMARCA*4	*ORT*/18	*6
TRADD	*ORT*/19	*9
*IL*6	*COPA*/17,*OS*/5	435
*LCN*2	*COPA*/21	141
*AGTR*1	*OS*/14	84
*CASC*3	*OS*/16,*LRS*/25	375
*GABRG*2	*LRS*/6	*35
CHGB	*LRS*/11	226
MGLL	*LRS*/13	*47
*SOD*2	*t*/24	*20

注：这些结果是对特定的基因(表中第 1 列)分别做 1000 次置换得到的结果，带*号表示似然比—置换检验法成功检测出该基因为致病基因(*ATM* 等 6 个)。其中 *ATM*、*SMARCA*4、*TRADD* 共 3 个基因满足。

Table 3. Some may-be-expressed genes by the method above
表 3. 本文所检验出的可能属于致病的基因

基因 ID	基因名称	基因描述	M	备注
Hs.587979	*PTOV*1	prostate tumor overexpressed 1	2	P. Benedit[9]
Hs.194143	*BRCA*1	breast cancer 1, early onset	42	
Hs.654445	*TNFSF*8	tumor necrosis factor (ligand) superfamily, member 8	3	
Hs.487062	*IGF*2*R*	insulin-like growth factor 2 receptor	4	Y. Oka[10]

4. 结果与讨论

本文在借鉴和改进多种现有方法的基础上，从考虑两样本不同分布的角度入手，提出了用于癌症差异表达基因判别的似然比—置换检验法，浅显易懂，操作方便，并取得了良好的效果。似然比—置换检验法较 *COPA*、*ORT* 等方法具有的优势在于，基于分布函数考虑的似然比统计量更能有效和充分地代表数据本身具有的特征，信息损失相对比较少。我们知道，*COPA*、*ORT* 等方法仅仅是利用中位数和上方 r 分位数(一般取 0.25)分别代表显著基因的属性。以模拟实验(d~e)为例，上方 0.25 分位数仍是从 $N(0, 1)$中取到的样本值，这使得差异表达部分的信息大量丢失。而基于分布的统计量则有效的避免了这种信息丢失。对于特定的临床试验，可以根据需要确定检验水平。然而，我们最终反馈的值 M 并不是一个确定的结果而是随机的，当异常值个数偏少的时候，检验错误的可能性也将增加。由于似然比统计量中用到了异常值部分的方差的估计，故当异常值的个数为 1 时本方法并不适用。

对于特定的基因来说，我们不知道置换反应所得到的那些大于 M 的 λ_j 的值，到底有多少比例是因为置换检验本身所造成的。这并不是仅仅从 M 的大小上可以判断出来的，我们需要知道异常值部分所占的比例进而对其进行估算。对于决定异常值个数的指标 k，是否可以将其同 M 综合考虑进行筛选，如果考虑进来会不会对结果有进一步的改善，诸如中位数、均值等数字指标是否对我们有用，都是今后值得进一步探讨的问题。

参考文献 (References)

[1] R. Graham. DNA chips: State-of-the art. Nature Biotechnology, 1998, 16: 40-44.

[2] J. D. Storey, R. Tibrashini. Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences of the United States of America, 2003, 100(1): 9440-9445.

[3] S. A. Tomlins, et al. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. Science, 2005, 310(5748): 644-648.

[4] D. Ghosh, A. M. Chinnaiyan. Genomic outlier profile analysis: mixture models, null hypotheses, and nonparametric estimation. Biostatistics, 2009, 10: 60-69.

[5] R. Tibshirani, T. Hastie. Outlier sums for differential gene expression analysis. Biostatistics, 2007, 8: 2-8.

[6] H. Lian. MOST: Detecting cancer differential gene expression. Biostatistics, 2008, 9: 411-418.

[7] 茆诗松, 王静龙, 濮晓龙. 高等数理统计(第二版)[M]. 北京: 高等教育出版社, 2006.

[8] M. West, et al. Predicting the clinical status of human breast cancer by using gene expression profiles. Proceedings of the National Academy of Sciences of the United States of America, 2001, 98: 11462-11467.

[9] P. Benedit, et al. PTOV1, a novel protein overexpressed in prostate cancer containing a new class of protein homology blocks. Nature Publishing Group, 2001, 20: 1455-1464.

[10] Y. Oka, et al. M6P/IGF2R tumor suppressor gene mutated in hepatocellular carcinomas in Japan. Hepatology, 2002, 35(5): 1153-1163.

