DOI:10.16016/j.1000-5404.2004.07.014
第26卷第7期 第三军医大学学报 Vol. 26, No. 7
594 2004年4月 ACTA ACADEMIAE MEDICINAE MILITARIS TERTIAE Apr. 2004

论著

文章编号:1000-5404(2004)07-0594-04

基因芯片表达数据的标准化策略研究

伍亚舟[1],张彦琦[1],黄明辉[2],杨梦苏[3],曾志雄[3],易 东[1] ([1]第三军医大学预防医学系卫生统计学教研室,重庆 400038;[2]香港城市大学深圳研究院,深圳 518057;[3]香港城市大学基因组科技应用研究中心,香港)

提 要:目的 从统计学、数学和生物信息学相结合的角度,探讨针对不同密度基因芯片(microarray, gene chip)如何进行数据标准化处理与分析,寻找表达上有显著性差别的基因。方法 在Excel和Spss软件中,采用总强度标准化法(total intensity normalization, TIN)和局部加权线性回归标准化法(locally weighted linear regression normalization, LWLRN)对高低密度芯片数据进行分析处理。结果 这些方法能有效地减少系统误差产生的影响,使处理后的数据更加具有可比性和可靠性。结论 本研究提供的分析方法在减少系统误差基础上能较好地找出芯片上有显著性差异表达的基因,不过标准化方法还处于起步和发展的阶段,需进一步完善;但随着非线性技术的不断完善,计算机软硬件的快速发展,芯片海量数据分析处理方面的研究将得到新的突破。

关键词:基因芯片(微阵列);标准化因子;表达比;显著性差异
中图法分类号:R319; R394.2 文献标识码:A

Normalization and analysis of expression data of gene chips

WU Ya-zhou[1], ZHANG Yan-qi[1], HUANG Ming-hui[2], YANG Meng-su[3], ZENG Zhi-xiong[3], YI Dong[1] ([1]Department of Health Statistics, College of Preventive Medicine, Third Military Medical University, Chongqing 400038; [2]Shenzhen Institute, City University of Hong Kong, Shenzhen 518057, China; [3]Applied Research Center for Genomics Technology, City University of Hong Kong, Hong Kong)

Abstract: **Objective** To explore how to conduct normalization and analysis of different density of gene chips based on statistics, maths, and bioinformatics for the purpose of finding significantly differentially expressed genes. **Methods** By the aid of Excel and SPSS softwares, data from high or low density chips were analyzed by total intensity normalization (TIN) and locally weighted linear regression normalization (LWLRN) methods. **Results** These methods could effectively reduce the influence resulted from the system errors, so the processed data were more comparable and reliable. **Conclusion** The methods mentioned in this paper can be helpful for the finding of significantly differentially expressed genes on the basis of decrease of system errors. However, normalization methods, still at the beginning and developing stages, need to be further improved. With the development of non-linear technology, software, and hardware of the computer, great achievements will be made in the studies of the analysis of enormous amount of microarray data.

Key words: gene chip (microarray); normalization factor; expression ratio; significant difference

基因芯片是近年来实验分子生物学的技术突破之一。如何对该技术产生的海量实验数据进行准确合理的分析,已成为有效应用这项技术的主要瓶颈问题,并已成为生物信息学的重要研究内容。

简单地说,基因芯片数据分析就是对芯片上高密度杂交点阵图像经过软件处理,从中提取杂交点的荧光强度信号进行定量分析,找出表达上有显著性差异的基因来构建表达矩阵,通过数据挖掘(data mining),最终整合出杂交点的生物学信息,从而发现基因的表达谱与功能可能存在的某种联系。但是由于芯片制备和点样等过程中随机误差和系统误差的存在,在确定阳性表达基因时,往往有较大的假阳性率[1]。为此,我们必须对提取的原始实验数据进行标准化处理,其目的是减少系统变化所带来的误差,以平衡各个基因表达水平在量纲上的和各个信道之间的差异来达到平衡,从而使处理后的数据具有可比性。本研究将结合笔者在常用的Excel和Spss软件中处理过的实例,给出两种标准化方法处理的过程。

1 理论及方法

1.1 散点图(scatter plot)

经过图像分析软件(Gene Pix Pro4.0)处理后得到*.gpr文

基金项目:国家自然科学基金资助项目(60371034)
Supported by the National Natural Science Foundation of China (60371034)
作者简介:伍亚舟(1977-),男,湖北省红安县人,硕士研究生,主要从事统计数据处理、生物信息学方面的研究。电话:(023)68752343
通信作者:易 东,E-mail:yd-house@hotmail.com,电话:(023)68752344
收稿日期:2003-09-01;修回日期:2004-02-06

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

件, 在 Excel 软件中打开即可得到芯片荧光强度信号经过转化后的原始数据, 经过数据整理, 提取出我们需要进行分析的相关数据列。

在做数据分析之前, 我们首先作两种红绿信号强度之间的散点图, 具体分析为: ①作散点图(Y 轴--F635median-B635, X 轴--F532median-B532)或取对数得到相应的散点图, 如图 1; ②在散点图上, 看散点的分布是否在 y= x 这条直线上: 若点几乎分布在一条直线上, 可以不必进行标准化处理; 否则, 需要进行标准化处理。

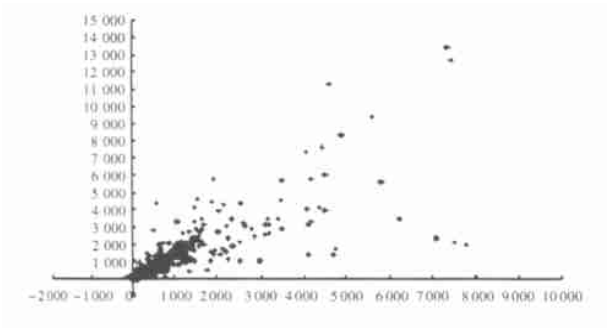


图 1　cy5/cy3(红/绿)两种荧光信号强度之间的散点图

Fig 1　The scatter diagram of the fluorescent intensities of Cy5/Cy3 (red/green)

1.2　表达比(Ratio)

表达比是反映两种信号强度关系之间的最基本比较[2]。大部分微阵列实验研究的两个问题是: 基于表达模式上与生物学相关两个样本之间的关系; 寻找有差异表达基因最为简单的方法。

假如在一张芯片上有 N 个元素点(即点样点, 每个点对应于一个序列), 如果用 R_i 和 G_i 分别表示红绿两种信号的荧光强度值, 且我们将两种红绿信号荧光强度的比(或 cy5/cy3 之比)记为 $Ratio_i$, 则有

$$Ratio_i = \frac{R_i}{G_i} \tag{1}$$

(这里 i= 1, 2, …, N, 下面的 i 表示相同意义), 若取对数有

$$log_2(Ratio_i) = log_2 \frac{R_i}{G_i} \tag{2}$$

这里 R_i 和 G_i 可为信号的中位数(Median)和均值(Mean)等强度值。

通常情况下, 我们认为表达比 $Ratio_i$ 大于 2(或取对数后大于 1)的基因表达是上调的, 表达比 $Ratio_i$ 小于 0. 5(或取对数后小于-1)的基因表达是下调的。虽然表达比 $Ratio_i$ 为表达水平改变提供了一个比较直观的衡量标准, 但是这种认定标准是比较粗糙的, 并没有坚实的理论依据作为基础, 因此会出现较高的假阳性率; 另一方面, 有一些生物学意义上比较重要的基因, 其表达比变化不大, 单用表达比分析就很难发现它们。为此, 有时并不能直接根据原始表达比对芯片的表达数据进行有效可靠的分析。

1.3　标准化方法

1.3.1　总强度标准化方法(total intensity normalization, TIN)

首先给出一些假设条件: 对于需要比较的两个样本, 分别以红绿两种染料标记染色(用 cy5/cy3 表示), 在生物特性相似情况下, 假设参与杂交的起始 RNA 数量是相等的, 并且假定在每个点样点里, 有成千上万的单个 RNA 分子, 每个分子的分子量大约是相同的, 在每个点样点里分子的数量也是相同的; 另外, 我们假设芯片上的元素点表示对基因的一个任意取样(sampling)。在这些假设基础上, 我们认为: 来自于每个样本大约相等的被标记分子的数量应该杂交这些阵列元素点, 从而对于每个样本, 所有元素点的杂交强度总和也应该是相等的。

根据这一思想, 可以得到一个标准化因子(或归一化因子), 用 NF 表示, 即

$$NF = \frac{\sum_{i=1}^{N} R_i}{\sum_{i=1}^{N} G_i} \tag{3}$$

这里 R_i 和 G_i 分别表示红(cy5)和绿(cy3)的信号强度值, 可为基因点信号的中位数(Median)或均值(Mean)等强度值。这时对于红绿单个信号强度值, 利用上述标准化因子, 可以表示为

$$G_i' = NF * G_i$$
$$R_i' = R_i \tag{4}$$

那么经过标准化处理后的表达比这里记为

$$Ratio_i' = \frac{R_i'}{G_i'} = \frac{1}{NF} \times \frac{R_i}{G_i} \tag{5}$$

校正后的所有 $Ratio_i'$ 的平均值为 1, 上述方程等同于从原始表达比 $Ratio_i$ 取以 2 为底对数的值中减去一个常量(即标准化因子取以 2 为底对数后的值), 即为

$$log_2(Ratio_i') = log_2(\frac{R_i'}{G_i'}) = log_2(\frac{R_i'}{G_i}) - log_2(NF) \tag{6}$$

1.3.2　局部加权线性回归标准化方法(locally weighted linear regression normalization, LWLRN)

上述的总强度标准化方法, 在数据中可能会出现的某些系统性误差时, 得到的结果会产生一定的偏差。如发现 $log_2(Ratio_i)$ 值[即 $log_2(R_i/G_i)$] 对于强度有一定系统依赖性(systematic dependence)[3,4], 特别是对于低强度[$log_2(R_i/G_i)$为 0 或小于 0] 的信号点, 这里的信号值接近背景值, 相对误差增加。对此, 我们给出局部加权线性回归标准化方法 LWLRN[1,5], 该方法可以较好地改善 $log_2(Ratio_i)$ 值对信号强度依赖所产生的影响。

LWLRN 基于以下 3 个假设: ①大部分基因并不是有差异变化的; ②上调与下调基因的数目相差不是特别大; ③前两个假设在信号范围内成立。为此这里我们最简单的设想方法是: 对于芯片上元素点的表达测量值 $log_2(Ratio_i)$ 是 $log_{10}(R_i * G_i)$ 信号强度的一个函数, 它们之间的散点图关系应该呈现一定的线性关系。这里我们给出了一张高密度芯片标准化前的散点图, 如图 2。

局部线性回归标准化方法发现 R-I 图中存在有某些系统偏差, 其线性关系也不是特别好, 通过建立一个局部线性回归方程可校正这些偏差。对于每个点来说, 该线性回归作为 $log_{10}(R_i * G_i)$ 的一个函数、并从经验观测值 $log_2(Ratio_i)$ 中减去计算好的最适合的所有 $log_2(Ratio_i')$ 的平均值。

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

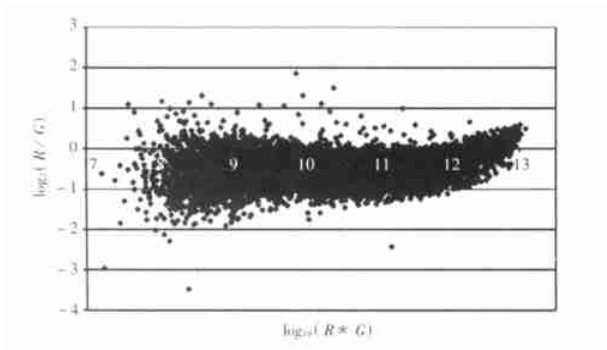


图 2 标准化处理前 $\log_2(Ratio_i)$ 与 $\log_{10}(R_i * G_i)$ 之间的散点关系图(R-I 图)

Fig 2 The scatter diagram of $\log_2(Ratio_i)$ *vs* $\log_{10}(R_i * G_i)$ before normalization (R-I)

我们先作如下的假设

$$X_i = log_{10}(R_i * G_i)$$
$$Y_i = log_2(R_i / G_i) \qquad (7)$$

根据所有的点(X_i, Y_i)我们可以得到一条回归直线所对应的方程 $Y(X_i) = a X_i + b$, 从而根据这个函数方程逐点校正 $\log_2(Ratio_i')$值, 即为

$$log_2(ratio_i') = log_2(ratio_i) - y(x_i) = log_2(ratio_i) - log_2(2^{y(x_i)}) \qquad (8)$$

或等同于下面的方程

$$log_2(Ratio_i') = log_2(Ratio_i * \frac{1}{2^{y(x_i)}}) = log_2(\frac{R_i}{G_i} * \frac{1}{2^{y(x_i)}}) \qquad (9)$$

这里的 $2^{y(x_i)}$即相当于标准化因子 NF。同其他标准化方法一样, 我们对两种信号强度方程作如下转换

$$G_i' = 2^{y(x_i)} * G_i$$
$$R_i' = R_i \qquad (10)$$

应用该方法后得到的 $\log_2(Ratio_i')$ 与 $\log_{10}(R_i * G_i)$ 之间的散点关系图如图3。比较图2 和图 3, 我们发现经过用回归校正后的散点图形有更好的线性关系。

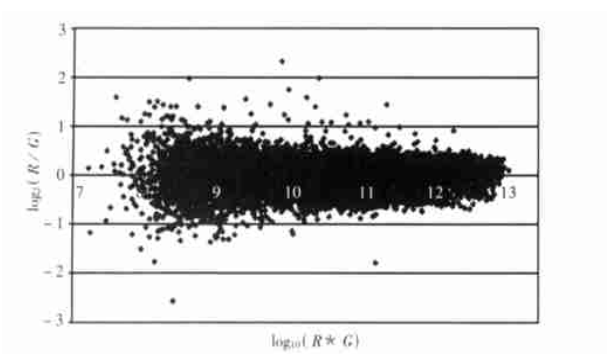


图 3 经标准化处理后 $\log_2(ratio_i)$ 与 $\log_{10}(R_i * G_i)$ 之间的散点关系图(R-I 图)

Fig 3 The scatter diagram of $\log_2(ratioi)$ *vs* $\log_{10}(R_i * G_i)$ after normalization (R-I)

2 实例分析

2.1 实例 1: 1 152 点密度基因芯片的分析

我们采用图像分析软件 Gene Pix Pro4.0 和数据处理软件 Excel 与 Spss 对来自 NIA(National Institute of Aging)中的小鼠 Clone set 里得到的 1 152 个元素点密度基因芯片, 采用 TIN 进行标准化处理。这里对两种荧光信号(635 代表红, 532 代表绿)的强度采用中位数(median)进行处理, F 和 B 分别表示前景和背景的荧光强度, 标准化因子计算如下式:

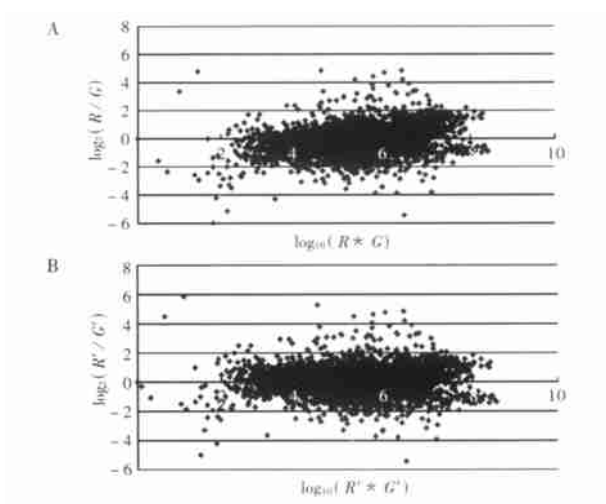
$$NF = \frac{\sum_{i=1}^{N}(F635_i Median - B635_i)}{\sum_{i=1}^{N}(F532_i Median - B532_i)}$$

经过标准化后我们发现, 某些基因表达比 Ratio 在前后发生了较明显的变化: 1 152 点中, Ratio> 2 的基因从标准化前的 58 个变为标准化后的 51 个, 而 Ratio< 0.5 的基因从标准化前的 276 个变为标准化后的 292; 而且经过与原图像比较发现标准化后的数据有更好的一致性, 能更好的反映基因的表达比。

2.2 实例 2: 7 680 点密度基因芯片的分析

我们对 mouse7500 II 型芯片(7 680 个元素点)采用 LWLRN 进行标准化处理。对两种信号真实强度的计算同 TIN。

经过标准化后我们发现, R-I 散点图在标准化前后有了较明显的改善, 标准化后其线形趋势更好些, 如图 4A、4B; 发现表达比 Ratio 在标准化前后也发生了较明显的变化: 7 680 点中, Ratio> 2 的基因从标准化前的 327 个变为标准化后的 235 个, 而 Ratio< 0.5 的基因从标准化前的 159 个变为标准化后的 240 个, 而且经过与原图像比较发现标准化后的数据有更好的一致性, 能更好的反映表达比。



A: Before normalization; B: After normalization

图 4 R-I 散点图分布情况

Fig 4 The R-I scatter plot

3 结果

本研究提出的两种数据标准化方法 TIN 和 LWLRN 可以有

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

效地减少系统变化产生的误差, 能比较满意的找出基因芯片中在表达上有显著性差别的基因, 处理后数据也更加具有可比性和可靠性, 主要表现在: ① 对于芯片图像上信号本来表达比较强的点, 经过 TIN 和 LWLRN 处理的数据更能反映每个点的真实表达水平, Ratio> 2.0 的点呈现红色, Ratio< 0.5 的点呈现绿色。②如有些强度本来比较弱的点, 图像上显示的亮度不特别明显, 如图 5, 但其表达比的改变比较大, 经过 LWLRN 标准化后我们会选择出来, 以避免漏选。③对于那些表达比变化并不是特别大, 但其信号强度又比较强的点(图像上显示黄色), 经过处理后其表达水平也发生了明显变化, 对这样的点可以避免错选。

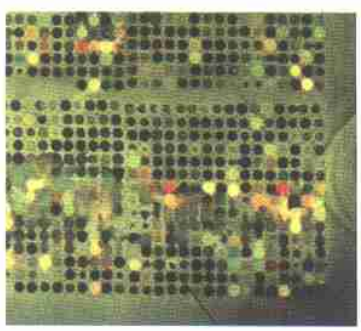


图 5　芯片图像上的信号不强点分布情况

Fig 5　The distribution of poor spots on chip image

4　讨论

目前标准化方法并没有一个完全统一的标准[6,7]。大多数标准化算法(包括 TIN 和 LWLRN), 既可以应用于全局(全部的数据), 也可以应用于局部(部分的数据)的标准化。总强度标准化(TIN)方法在低密度(如几十到几百个, 一般 1 000 个元素点左右附近)且信号强度值较强芯片的数据分析中, 应用得比较多, 处理起来比较方便, 而且效果也不错。局部标准化也有其优点: 它能很好的校正阵列上系统的偏差, 包括阵列上不一致的元素点, 表面的变化和与阵列点杂交上差别等。然而, 基因表达数据分析方法仍处于起步和探索阶段, 很多方法其应用还不是很完善, 如 TIN 方法在低密度芯片方面应用较好, 但对于高密度的芯片, 尤其对于那些信号强度低的点来说, 此法找出的差异基因存在些许不足(如漏找或错找); 不过 LWLRN 方法比较好的解决了这个问题, 经过处理后的表达比可以更好的反映基因的表达水平, 也获得了令人比较满意的结果。当然, 由于背景信号太强、点样等方面引起的系统误差, 本文提出的方法对其处理还不能完全消除, 为此有些方法还在进一步的完善发展中。

参考文献:

[1] Quackebush J. Microarray data normalization and transformation[J]. Nat Genet, 2002, 32(Suppl): 496—501.

[2] Chen Y, Dougherty E R, & Bittner M L. Ratio-based decisions and the quantitative analysis of cDNA microarray images[J]. J Biomed, 1997, 364—374.

[3] Yang Y H, Dudoots Luu P, *et al*. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation[J]. Nucleic Acids Res, 2002, 30(4): e15.

[4] Yang I V, Chen E, Hasseman J P, *et al*. Within the fold: assessing differential expression measures and reproducibility in microarray assays[J]. Genome Biol, 2002, 3(11): research0062.

[5] Cleveland W S. Robust locally weighted regression and smoothing scatterplots[J]. J Amer Stat Assoc, 1979, 74: 829—836.

[6] 易　东, 杨慧芬, 等. 一种基因芯片数据的标准化处理策略及软件实现[J]. 第三军医大学学报, 2004, 26(3): 224—226.

[7] 易　东, 黄梦芬, 李辉智, 等. 基因表达聚类结果的信息熵评价方法[J]. 第三军医大学学报, 2004, 26(4): 317—319.

(编辑　张大春)

(上接 593 页)

病, 多见于育龄妇女, 极少数发生在绝经后。绒癌恶性程度高, 早期就可发生血行转移至全身, 引起组织和器官出血坏死。育龄妇女的绒癌多数继发于葡萄胎及正常分娩后, 肿瘤常位于子宫肌层, 也可突向宫腔或穿破浆膜, 大小在 0.5～5 cm; 中晚期肿瘤可广泛侵入子宫肌层并破坏血管, 造成恶性大出血。绝经后的绒癌罕见, 一般发生在绝经后 2～4 年, 有多次分娩史[2], 临床以异常阴道流血为主要症状, 容易误诊为子宫内膜癌、子宫肉瘤等。本例患者发病距末次妊娠 21 年, 病灶巨大, 且无子宫外转移征象均增加了诊断的难度; 及时纠正诊断的关键在于诊刮和血 HCG 的测定。由于绒癌的组织病理特点决定了诊刮就可能导致急性大量出血, 甚至危及生命, 故术前必需做好针对大出血的急救准备; 从临床治疗看, EMA-CO 方案治疗绝经后绒癌效果欠佳的主要原因是肿瘤体积过大; 故对于这种肿瘤巨大的绝经后绒癌应当在化疗 1～2 个疗程后及时采取手术, 以缩短治疗时间。此外, 介入治疗也是一种良好的选择[3]。

关键词: 绒毛膜癌; 绝经

中图法分类号: R737.3　　文献标识码: B

参考文献:

[1] 宋鸿钊, 董淑英, 杨秀玉. 近代滋养细胞肿瘤的研究进展[J]. 中华妇产科杂志, 1993, 28(8): 476—478.

[2] 李　菲, 李　力, 张洁清. 绝经后绒癌 16 例临床分析[J]. 广西医科大学学报, 1994, 01: 125—127.

[3] 杨秀玉, 杨　宁, 宋鸿钊, 等. 超选择性动脉插管持续灌注化疗治疗绒癌耐药性患者的分析[J]. 中华妇产科杂志, 1996, 31(4): 199—202.

(编辑　陈聪连)

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net