DOI:10.16016/j.1000-5404.2004.03.015
第26卷第3期 第三军医大学学报 Vol. 26, No. 3
224 2004年2月 ACTA ACADEMIAE MEDICINAE MILITARIS TERTIAE Feb. 2004

论著

文章编号:1000-5404(2004)03-0224-03

一种基因芯片数据的标准化处理策略及软件实现

易 东[1],杨梦苏[2],李辉智[3],黄明辉[2],王文昌[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]香港城市大学基因组科技应用研究中心,香港;[3]西南政法大学刑侦学院电子技术教研室,重庆 400031)

提 要:目的 建立基因芯片数据处理的标准化方法,研制一套实用、可操作性强的软件Gene-Norm。方法 采用criteria将杂交点分为good和bad两类。同时,在bad杂交点中寻找turn-on-off点。最后,用平均ratio进行标准化处理。结果 将软件Gene-Norm应用于高密度芯片数据的处理,该芯片有10 080个杂交点,每个杂交点有42个记录进行处理,得到了较好的处理结果。讨论 本软件适用于大型数据文件的处理,且提供了许多参数以供选择,因此适用性和可操作性强。

关键词:芯片;标准化;软件
中图法分类号:R319;R394 文献标识码:A

Normalization strategies for microarray data

YI Dong[1], YANG Meng-su[2], LI Hui-zhi[3], HUANG Ming-hui[2], WANG Wen-chang[1] ([1]Department of Health Statistics, Third Military Medical University, Chongqing 400038; [2]Applied Research Center for Genomics Technology, City University of Hong Kong; [3]Department of Electronic Technology, College of Criminal Investigation, Southwest China University of Politics and Law, Chongqing 400031, China)

Abstract: Objective To design a set of software Gene-Norm for the data processing of genechips by normalization methods. Methods Hybridization spots were divided into two groups: "good" and "bad". Turn-on-off spots were extracted from the "bad" spots. Normalization processing of data was performed by using average ratio. Results Application of Gene-Norm in the processing of high-density microarray data (10 080 spots, 42 records of each spot) resulted in satisfactory processing results. Conclusion The software of genechip can be used to process large amount of data. Some parameters can be chosen manually, so the software is of simple operation and wide application.

Key words: microarray; normalization; software

基因芯片的数据分析简单来说就是对微阵列(microarray)高密度杂交点阵图像处理,并从中提取杂交点的荧光强度信号进行定量分析,通过有效数据的筛选和相关基因表达谱的聚类,最终整合杂交点的生物学信息,发现基因的表达谱与功能可能存在的联系[1~4]。因此,如何才能准确地提取出芯片数据(即ratio值),减少噪声的干扰,是基因芯片数据分析的关键所在,该处理过程被称为芯片数据标准化(normalization)。本研究深入研究了噪声的产生过程,建立了normalization处理的基本策略,设置了许多可供选择的质量控制参数,并介绍了作者研制完成的标准化(Gene-Norm)处理软件的研制。

1 芯片数据获取的基本原理

1.1 基因芯片上各克隆荧光信号的意义

当芯片经过杂交反应并由Cy3/Cy5染色后,就采用激光共聚焦扫描原理,对杂交点(spots)进行荧光信号采集。这时,严格配对的杂交分子,其热力学稳定性较高,荧光强;不完全杂交的双键分子热力学稳定性低,荧光信号弱(不到前者的1/35~1/5),不杂交的无荧光。当不同位点信号被激光共焦显微镜,或落射荧光显微镜等扫描后,输入计算机转化为数字信号后,就可得到Cy3和Cy5两个扫描图像文件,见图1。

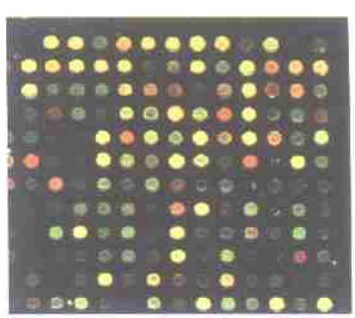


图1 基因扫描(杂交点)图

Fig 1 The scanned image of gene spots

1.2 图像获取

当扫描完成后,就同时导入同一区域两个通道(channel)扫描的图像文件;将两个channel扫描的图像用不同的颜色显示并重叠。图1中Cy3用绿色表示,Cy5用红色表示,黄色表示两者的表达强度相同,黑色代表没有信号。在图像的获取过程中会产生许多噪声,主要原因包括:芯片制作的质量和实验过程。

作者简介:易 东(1964—),男,重庆市人,博士,教授,主要从事生物信息学方面的研究,发表论文30余篇。电话:(023)68752344

收稿日期:2002-09-20;修回日期:2002-12-10

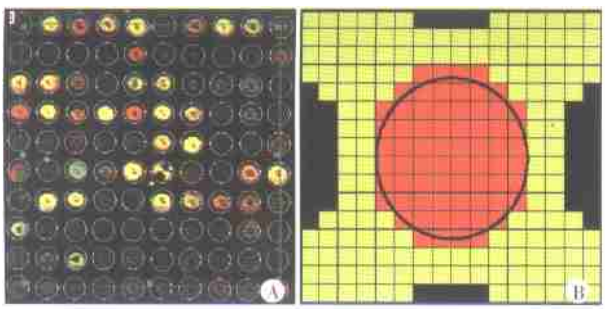
?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

其中实验过程是噪声的重要来源, 如杂交的不均衡性、染色的不均、冲洗过程的污染等等。因此, 标准化方法应围绕这几个方面进行。

1.3　图像的网格标定(griding)及数据的获取

将图像的点样克隆信号圈定杂交点, 并将其内部区域作为待分析的信号(图 2), 同时在圈定的各克隆周围再按照预设的值圈定一定范围的区域, 将该区域内的信号作为背景噪音; 计算机分析每个克隆扣除背景噪音后的信号强度, 并按照不同的要求给出两个 channel 信号强度的相对比例。最后以文本形式输出表达强度数据。

在这个过程中才能如何准确地圈定杂交点 s 是信号提取的关键(图 2B), 解决该问题要用到计算机图像的自动识别技术。但由于受到噪声的影响, 所以目前仍然是研究的难点。



A: Griding image of gene spots; B: Refinement of gene spots by circling

图 2　扫描图的 **griding** 和 **circling spots** 示意图

Fig 2　Diagram of griding and circling spots on scanned image

2　芯片数据标准化的基本策略

这里, 我们主要围绕自行研制的软件 Genechip-Norm, 用图 3 的形式简单阐述。

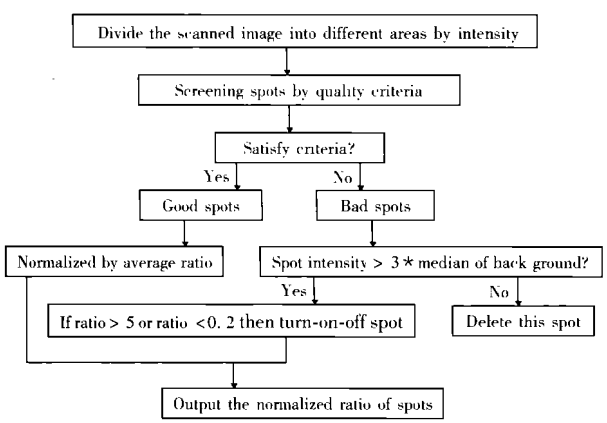


图 3　标准化基本策略

Fig 3　Basic strategies for normalization

3　标准化方法的软件实现

本软件 Gene-Norm 是用 PowerBuilder6.0 编制完成的, 适用于对大型数据库的分析。该软件的主界面如图 4 所示。

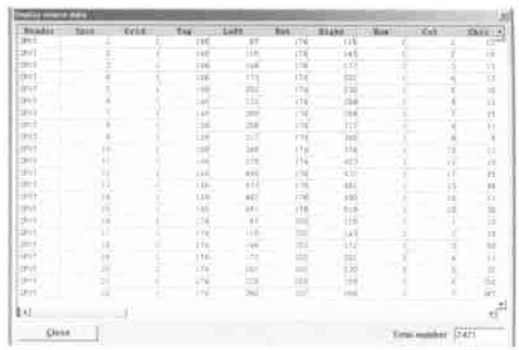


图 4　Gene-Norm 软件的主界面

Fig 4　The main interface of Gene-Norm

3.1　数据的输入格式

本软件适用于处理文本格式的数据文件, 因为许多激光共聚焦扫描的最后输出结果均为. txt 的文本文件。该文档包括: 基因描述、杂交点的定位参数、杂交点的荧光强度参数、杂交点的背景强度参数、杂交点与杂交点之间、杂交点与背景之间各种相关参数描述。图 5 界面右下角的 Total number 为读入的基因总数目。

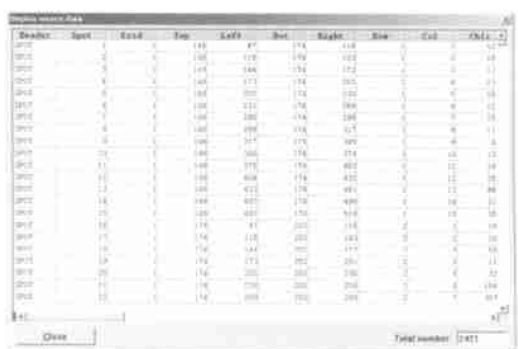


图 5　数据文件格式

Fig 5　Format of data file

3.2　标准化处理各个参数的意义

此步处理是整个标准化处理的关键。在进行标准化之前, 必须确定以下几个参数(图 6)。这里, 我们给出这些参数(criteria)的具体意义: ①M-STEPS: 它决定了扫描图像区域划分的大小, 此值越大划分的区域越大, 默认值为 300, 表示 300 个点为一个区域进行划分; ②CORR: 表示 Cy3 和 Cy5 两幅图像中同一位置杂交点强度分布的相关一致性, 0< CORR< 1 越接近与 1, 表明一致性越好; ③CH1GB1、CH2GB2: Cy3 和 Cy5 图像中杂交点内强度大于背景的像素占该杂交点所有像素的百分比, 该值越接近于 100%, 质量越好; ④FLAG: 人为观察图像后表定的点, FLAG= 1 表示删除该点, FLAG= 0 表示该点参与标准化; ⑤CH1KSP, CH2KSP: Komogorov-Smimov 似然统计量, 描述杂交点的强度分布与背景强度分布的相似的显著性; ⑥Up-threshold, Down-threshold: 分别表示上调和下调的界值; ⑦GTB: 在 bad 杂交点中杂交点强度大于背景强度中位数的倍数。当然, 这些参数也可以采用默认值(default)。

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

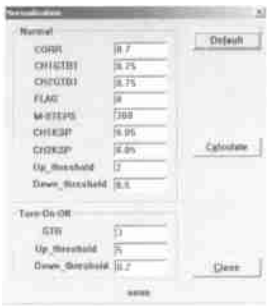


图 6 标准化处理中几个重要的参数

Fig 6 Some important parameters in normalization processing

3.3 标准化处理结果的表达

在标准化处理后结果表达中,我们保留了原始数据文件的所有数据。同时,给出了标准化后的 ratio,如图 7 中的 Nrat。图



图 7 结果表达

Fig 7 Results of normalization processing

中的 Mark 标定了哪些点是 good 杂交点(Mark= 1),哪些点是 turn-on-off 杂交点(Mark= 2)。同时,我们可以使用 Find 命令在结果中查找感兴趣的基因。

4 讨论

本研究简单介绍了 Gene-Norm 软件的处理策略,介绍了各个参数的意义。该软件使用了 Powerbuolder6.0,因此,可以应用于对大型数据文件的处理,如我们曾将该软件应用于高密度芯片的处理,该芯片有 10 080 个杂交点,每个杂交点有 42 个记录进行处理,得到了较好的处理结果。本软件设置了许多参数可供选择,这些参数的确定主要依据芯片制作的质量和实验的质量控制标准,当然如果对标准化处理的原理和过程不熟悉,也可以使用默认参数。本软件目前仍有许多地方可以开发完善,如可以嵌入多种标准化计算方法,可以将结果文件与图像文件联接,在查询基因时直接将该杂交点显示出来。

参考文献:

[1] Collins F S, Patrinos A, Jordan E, *et al*. New goals for the U.S. Human Genome Project: 1998—2003[J]. Science, 1998, 282(5 389): 682—689.

[2] Brazma A, Vilo J. Gene expression data analysis[J]. FEBS Lett, 2000, 480(1): 17—24.

[3] 易 东,杨梦苏,等. 基于伪 F 统计量的模糊聚类方法在基因表达数据分析中的应用[J]. 中国卫生统计, 2002, 19(3): 146—151.

[4] Geschwind D H. Sharing gene expression data: an array of options[J]. Nat Rev Neurosci, 2001, 65(2): 435—438.

(编辑 张大春)

关于计量单位的使用

在医学领域中几乎时时处处都离不开计量单位。国务院于 1984-02-27 公布了《关于在我国统一实行法定计量单位的命令》和《中华人民共和国法定计量单位》。1993-12-27 国家技术监督局发布了《量和单位》国家强制标准,全国必须统一采用的计量单位。

我国的法定计量单位包括:①国际单位制的基本单位,如长度(m)、质量(kg)、时间(s)等;②国际单位制的辅助单位;③国际单位制中具有专门名称的导出单位,如频率(Hz)、压力(Pa)、能量(J);④国家选定的非国际单位制单位,如时间(min、h、d)、转速(r/min);⑤由以上单位构成的组合形式单位,如速度(m/s)、浓度(mol/L);⑥由词头和以上单位所构成的倍数和分数单位,如 μm、kPa、ml 等。在医学写作时必须严格执行国务院的命令和有关的国家标准。

各种单位除在无数值的叙述性文字中和科普读物中可使用中文名称或中文符号外,均应使用符号。单位符号都是正体拉丁和希腊字母,无复数形式,符号后不附加圆点(正常语法名字尾标点符号除外)。来源于人名的单位,例如:Pa、Gy、W 等和 10^6 以上的词头,例如:M、G、T 等的首字母大写以外均为小写,“升”的符号可用大写(L)也可用小写(l),一般在单独使用不加词头时用大写。数值与单位符号之间、单位符号与相邻的其他符号(例如:化学元素符号)之间应留半个阿伯数字空隙,词头与其所组合的单位符号之间不需空隙。词头不能单独使用,也不能重叠使用(例如:“mμm”应为“nm”)。

本刊编辑部

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