新技术讲座

基因芯片数据分析与处理*

王永煜 张幼怡**
(北京大学第三医院血管医学研究所, 北京 100083)

摘要 基因芯片技术在基因表达分析等应用过程中产生大量的数据, 如何处理和分析这些数据并从中提取出有价值的生物学信息是一个极为重要的问题. 其过程包括原始数据的获取及处理、标准化数据的统计学分析、以及数据的存储和交流等.

关键词 基因芯片, 数据分析, 聚类
学科分类号 Q332

随着生命科学进入“后基因组时代”, 基因组研究的重心也逐渐转向了基因功能的研究. 基因芯片 (microarray) 技术无疑为基因功能研究提供了一种强有力的工具. 高通量特点使其在基因表达分析、疾病诊断和治疗、新药发现等众多领域得到广泛应用[1]. 但是基因芯片的应用过程产生了大量的关系复杂的数据, 如何处理和分析这些数据并从中挖掘出有意义的生物信息已成为限制该技术进一步发展的主要“瓶颈”. 本文就基因芯片数据处理与分析问题作简要介绍.

1 原始数据的获取及处理

1.1 原始数据的获取

提取生物样品的 mRNA 并反转录成荧光素或同位素标记的 cDNA, 在液相中与基因芯片上的探针杂交, 经洗膜后用图像扫描仪捕获芯片上的荧光或同位素信号, 由此获得的图像就是基因芯片的原始数据. 此后还需用图像分析软件从中提取各点的吸光度值、面积和吸光度比等数据并转化成基因表达矩阵 (gene expression matrix), 才能进行进一步的统计学和生物学分析.

1.2 原始数据的处理

当前已开发出许多相关图像处理分析软件. 它们能自动定位并识别芯片上每个杂交点, 通过背景调整或分割技术除去图像上各种形式的噪声, 再定量各点的信号强度比率, 最后决定相应基因的表达变化情况.

1.2.1 背景处理: 图像上各点的吸光度值包含了样品和背景信号, 在提取数据前必须将背景扣除. 一般解决办法是以芯片图像中每个方格 (grid) 内除杂交点以外各像素的吸光度平均值作为背景, 将各点的强度减去这个背景值即可. 然而这种方法并不准确而且会使 1%～5% 的点产生无意义的负值. Brown 等[2]提出利用整个芯片杂交点外的平均吸光度值作为背景的 best-fit 方法, 使该问题得到较好的解决, 并有效地提高了处理数据的质量.

1.2.2 杂交点质量: 由于点样或膜变形等原因目前较多的软件对杂交点的识别定位仍需要人为的干预和调整. 以玻璃等硬质材料为片基的芯片, 其杂交点边缘一般比较清晰易于界定, 但对于膜阵列芯片通常杂交点边缘比较模糊不易识别, 且背景难以确定易造成误差. 为此, Jain 等[3]开发出一个完全自动化的图像处理软件, 从斑点的划格, 定位到计算吸光度比值等都不需人工参与, 并且获得的数据较可靠. 对基因芯片图像分析的各种方法可参阅 Yang 等[4]的综述.

1.2.3 数据的标准化: 其目的是避免基因芯片实验中因系统差异 (systematic variation) 造成芯片间数据比较的困难. 大部分标准化的方法采用调整标准化系数使平均比值 (ratio) 为 1 或平均 ratio 对数值为 0. 最常用的是“看家基因 (house-keeping gene)”法, 它预先选择一组表达水平不变的看家基因, 计算出这组基因平均 ratio 值为 1 时的标准

* 国家重点基础研究发展规划项目基金 (973) (G2000056906) 和国家自然科学基金 (30070872) 资助项目.
** 通讯联系人.
Tel: 010-62092306, E-mail: zhangyy@bjmu.edu.cn
收稿日期: 2002-08-16, 接受日期: 2002-10-16

化系数，然后将其应用于全部的数据以达到标准化的目的[5]. 此外，整体平均值法（global mean normalization）和密度依赖（intensity-dependent）标准化法也很常用[6]. 但至今仍无很理想的标准化方法.

2 标准化数据的统计学分析

原始数据标准化并转化成基因表达矩阵后，通过统计学分析，可从中揭示出一些重要的生物学信息. 目前大致有两类分析方法即差异分析和聚类分析.

2.1 差异分析

主要目的在于筛选出不同条件下表达明显差异的基因. 当比较两个不同生物样本时，可根据 ratio 值来筛选，一般 ratio 值在 0.5~ 2.0 范围内的基因不存在显著表达差异，该范围之外则认为表达发生显著改变. 然而由于不同实验数据变化差别很大，因此根据实验条件不同来调整域值更为合理. 在分析两种生物条件下多个重复样本的数据时，可通过 t 检验来筛选差异基因. 最近，Jin 等[7]用无参数的 Mann-Whitney 方法鉴别差异表达的基因，观察了卡托普利（captopril）对心肌梗塞大鼠心肌组织基因表达水平的影响，结果用定量 PCR 验证，发现基本没有假阳性. 在比较多种生物条件下的芯片数据时，可用 F 检验筛选特异表达的基因. 有时需要鉴别基因的某一特定行为，则可采用假表达谱（pseudoprofile）的方法，例如欲鉴别在肺癌中高表达而在正常肺组织和其他肿瘤组织中低表达的基因，就可先假设具有这样一假表达谱，然后在实际芯片数据中去寻找与其相吻合的基因. 此外，聚类中的监督分析方法同样也适于这种情况下的候选基因的鉴别.

2.2 聚类（clustering）分析

根据统计分析原理，将具有相同统计行为的基因进行归类，从而发现生物学行为相似或相关的一组基因，常采用监督（supervised）分析和非监督（unsupervised）分析的策略. 监督分析是根据已知的参考向量（vector）对基因进行分类，通过建立分类标准，将未知基因“安排”进已知基因的分类中，以此来预测新基因的功能. 非监督分析没有已知参考向量，只是将相同表达行为的基因或样品归为一类，在此基础上寻找相关基因，分析基因的功能. 它们均可实现大量表达数据的简化.

2.2.1 非监督分析：统计学上通过计算相似距离（similarity distances）来比较数据，常用相关系数或欧氏距离表示. 已发展了多种算法应用于芯片数据的聚类分析，包括层次式聚类（hierarchical clustering）、自组织作图（self-organizing maps）、K-means 聚类等. 最常见的是层次式聚类，它计算各数据点间的距离后把相近距离的聚为一组，再计算各组间的距离使之合并成更大的组，不断重复这一过程直到最后聚成一组以树状结构重新安排的数据. 其结果直观而且能以树状结构分支的长短来评价基因的相似性.

近年来已有一些采用聚类分析方法研究细胞功能或疾病特征的报道. 例如在对酵母基因调节机制及信号转导通路等的研究中[8]，通过聚类揭示了各种条件下许多功能相关或共表达的基因. 此外，聚类分析对肿瘤的分类、诊断、疗效和预后的评价同样具有重要的实际应用价值.

2.2.2 监督分析：其目标之一是根据已知基因分组并建立每组的分类标准（classifiers），然后将功能未知的基因安排到各组中，通过比较分析了解未知基因和已知基因之间的关系. 这些分析可以通过监督机器学习技术（supervised machine-learning techniques）进行，包括 SVM（support vector machines）、logistic 回归、神经网络和 LDA（linear discriminant analysis）等方法. Brown 等[9]用几种方法研究酵母 6 类功能基因，发现 SVM 能很好地识别和预测基因的功能.

3 数据的存储及交流

基因芯片数据分析领域尚处初期发展阶段，对数据的储存、分析和结果的交流还缺乏一种广泛接受的方法. 目前已有人号召建立基因芯片数据的标准中心数据库，如 NCBI 提出的 Gene Expression Omnibus 方案和 EBI 的 ArrayExpress 方案，包括数据的提交、存储、标准化和分析软件等内容. 值得一提的是 Brazma 等[10]提出了描述和交流基因芯片数据的统一标准，即 MIAME 方案：用最少的信息来描述基因芯片实验数据，最终目的是建立一个标准格式来记录和交流基因芯片数据.

最近杜克大学还召开了有关基因芯片数据分析方法评价会议[11]，以及每年召开的太平洋地区生物计算会议（http://psb.stanford.edu/）等，对基因芯片数据分析的发展都起了重要的推动作用.

参 考 文 献

1 Debouck C, Goodfellow P N. DNA microarray in drug discovery

and development. Nat Genet, 1999, **21** (1 suppl): 48~ 50
2 Brown C S, Goodwin P C, Sorger P K. Image metrics in the statistical analysis of DNA microarray data. Proc Natl Acad Sci USA, 2001, **98** (16): 8944~ 8949
3 Jain A N, Tokuyasu T A, Snijders A M, *et al*. Fully automatic quantification of microarray image data. Genome Res, 2002, **12** (2): 325~ 332
4 Yang Y H, Buckley M J, Speed T P, *et al*. Analysis of cDNA microarray images. Brief Bioinform, 2001, **2** (4): 341~ 349
5 Bilban M, Buehler L K, Head S, *et al*. Normalizing DNA microarray data. Curr Issues Mol Biol, 2002, **4** (2): 57~ 64
6 Sherlock G. Analysis of large-scale gene expression data. Briefings in Bioinformatics, 2001, **2** (4): 350~ 362
7 Jin H, Yang R, Awad T A, *et al*. Effects of early ACE inhibition on cardiac gene expression following acute myocardial infarction. Circulation, 2001, **103** (5): 736~ 742
8 Roberts C J, Nelson B, Marton M J, *et al*. Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles. Science, 2000, **287** (5454): 873~ 880
9 Brown MP, Grundy W N, Lin D, *et al*. Knowledge-based analysis of microarray gene expression data by using support vector machines. Proc Natl Acad Sci USA, 2000, **97** (1): 262~ 267
10 Brazma A, Hingamp P, Quackenbush J, *et al*. Minimum information about a microarray experiment (MIAME) -toward standards for microarray data. Nat Genet, 2001, **29** (4): 365~ 371
11 Johnson K F, Lin S M. Critical assessment of microarray data analysis: the 2001 challenge, Bioinformatics, 2001, **17** (9): 857 ~ 858

Microarray Data Analysis and Processing*

WANG Yong-Yu, ZHANG You-Yi**

(Institute of Vascular Medicine, Peking University Third Hospital, Beijing 100083, China)

Abstract Application of microarray technology had arised huge volumes of complex data. It is important how to handle and analysis of these data and to extract some valuable information from them. The acquisition and processing of microarray data, the statistical analysis of normalized data, and the storage and communication of data were briefly discussed.

Key words microarray, data analysis, clustering

* This work was supported by grants from The Special Funds for Major State Basic Research Development Program of People's Republic of China (G2000056906) and The National Natural Sciences Foundation of China (30070872).

** Corresponding author. Tel: 86-10-62092306, E-mail: zhangyy@bjmu.edu.cn

Received: August 16, 2002 Accepted: October 16, 2002