生物化学与生物物理进展 *Progress in Biochemistry and Biophysics*, 2006, 33(6): 556~561
www.pibb.ac.cn

PIBB

寡聚核苷酸芯片表达谱系统偏移的校正算法 *

邱浪波 [1,2)**] 王广云 [1,2)] 王正志 [1)]

([1)]国防科学技术大学机电工程与自动化学院, 长沙 410073;
[2)]空军工程大学电讯工程学院, 西安 710077)

摘要 多芯片对比实验中，由于多方面的变异因素，使得芯片间存在明显的系统偏移. 因此，芯片表达谱数据的校正处理是关键的数据预处理步骤. 当前，已经提出了很多校正算法，比如：比例常数校正、非线性校正、分位数校正等. 提出了一种新的校正算法. 在选择的最小秩差异探针集上，进行非线性 *M-A* 校正. 并采用迭代策略减弱基准芯片方法对基准芯片选择的敏感性. 在标准测试集上，同几种已知的方法进行了对比分析.

关键词 高密度寡聚核苷酸芯片，系统偏移，校正，转录产物
学科分类号 Q-332

基因芯片是分子生物学、微电子学和信息学等多学科交叉形成的新型生物技术. 目前，已经广泛应用于分子生物学、生物医学等研究领域，如DNA 测序、基因调控网络和癌症检测等. 高密度寡聚核苷酸芯片在玻璃片上按阵列固定匹配 / 失配(PM/MM)寡核苷酸探针对. 寡核苷酸探针的长度通常为 20～25 bp . 对于每一个待检测的 mRNA，通常用 11～20 对探针来检测. 芯片实验中试验变量对检测到的基因表达密度有很强的影响[1,2]. 为了得到准确的特异性杂交水平，必须对观测数据进行修正. 寡聚核苷酸芯片数据预处理一般包括三个步骤：背景修正(background correction)，系统校正(normalization)，表达值综合(summary). 芯片表达谱数据的有效预处理是芯片数据后续分析的关键步骤[3~10].

大部分实验涉及多芯片，因此，很重要的一点就是排除非生物因素引起的芯片间的变异，使得来自不同芯片的数据具有可比性. 这些非生物因素变异来源主要包括：反转录效率的差异，染色标记或者杂交反应差异，芯片本身的物理特性，实验试剂的效应，实验室环境等等[1,2]. 在对比分析中，有效的系统校正处理能够减少芯片间的系统偏差，减少在芯片处理过程中的技术因素对芯片检测结果的影响，使检测的结果能真实反映生物功能的差别. 因此，系统校正是多芯片对比分析试验中的重要处理环节. 到目前为止，已经提出了很多校正算法，比如：比例常数校正(scaling normalization)，非线性校正(nonlinear normalization)，基于正交变换的对照校正(contrast normalization)，迭代多项式拟和校正(cyclic loess normalization)，分位数校正(quantile normalization)等[3~11]. Affymetrix 公司提供的标准分析软件 MAS 5.0 采用的是基于基准芯片的 Scaling normalization 方法，以基准芯片探针强度均值与目标芯片探针强度均值的比例常数为参数，对目标芯片的每个探针进行比例校正[3]. Scaling normalization 进行的是线性校正，所以不能有效地消除芯片间的非线性偏移. Li 和 Schadt 等[4,8]分别提出了基于非线性拟和的 Nonlinear normalization 方法. 前面几种方法都是基于基准芯片的校正方法，因此，对基准芯片的选择比较敏感[11]. Contrast normalization 对探针杂交强度矩阵进行正交变换，在变换矩阵上，以第一分量为基准分量对各分量进行非线性校正，然后进行反正交变换，得到校正探针杂交强度矩阵[5]. Cyclic loess normalization 方法采用了一个迭代的过程，芯片两两之间进行 *M-A* 非线性局部回归估计校正，直到所有芯片间的校正收敛[6]. Cyclic loess normalization 方法的时间效率是 $o(mn^2)$，当芯片数较多的情况下，时间效率较低. Quantile

*国家自然科学基金资助项目(60471003).
** 通讯联系人. Tel: 0731-4574991, E-mail: qlbogfkd@nudt.edu.cn
收稿日期：2005-12-09，接受日期：2006-01-28

normalization用各芯片上秩相同的探针杂交强度的中位数替换所有秩相同的探针杂交强度，是时间效率最高的方法[11].

本文提出了一种迭代的鲁棒基准芯片校正方法(iterative robust baseline normalization, IRB normalization). 通过对各芯片进行秩排序，选择一个秩差异最小的探针子集. 利用 Tukey biweight 算法得到一个伪基准芯片，然后基于伪基准片对目标芯片采用 *M-A* 非线性校正. 对上述过程进行迭代，当达到最大迭代数或者探针杂交强度校正前后的差值低于某个阈值时停止. 以 Affymetrix 公司提供的标准检验数据集 HG_U133A Spike-in Dataset 作为测试数据，对几种校正方法在表达值水平上进行了对比分析. 分析结果将在文章的对比分析部分给出.

1　问题描述

在基因芯片表达数据中，大部分基因是非显著性差异表达的，因此，假设非显著性差异表达的探针在不同的芯片上具有相似的杂交强度分布. *M-A* 散点图显示在对数刻度下，同一探针在两个不同芯片中的差异表达和探针杂交强度之间的关系. 因此，*M-A* 散点图能够很好地显示芯片间的系统偏移. 定义：

$$M_{ik}^{j}=\log_2(y_{ij})-\log_2(y_{kj}) \quad (1)$$

$$A_{ik}^{j}=\frac{1}{2}(\log_2(y_{ij})+\log_2(y_{kj})) \quad (2)$$

其中，y_{ij} 定义为探针 j 在芯片 i 上的杂交强度值，M_{ik}^{j} 定义为探针 j 在芯片 i 与芯片 k 上杂交强度的对数差，描述探针在不同芯片上杂交强度的比例差异，A_{ik}^{j} 定义为探针 j 在芯片 i 与芯片 k 上杂交强度的对数均值. 图 1 显示了几个未处理芯片 PM 探针两两比较的 *M-A* 散点图.

理想状态下，散点云团应该以 M=0 为对称轴成彗星状分布. 从图 1 可以看出，芯片间存在明显的系统偏移，为了得到芯片间有意义的对比分析结果，必须对芯片数据进行校正，去除芯片间的系统偏移.

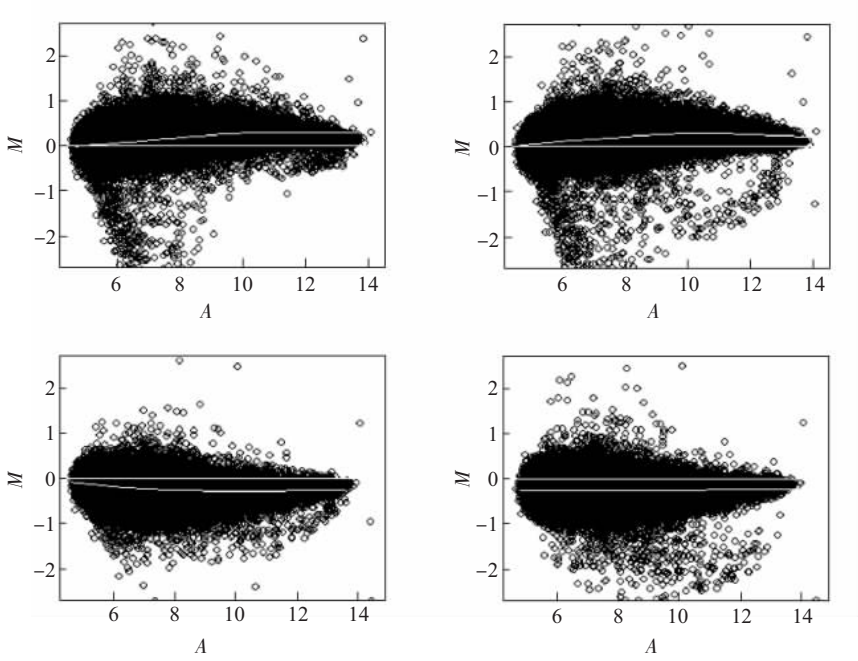

Fig. 1　4 pairwise *M* vs. *A* plots using HG_U133A Spiked-in data for unadjusted data

2　方法描述

校正是在对芯片进行背景修正后消除系统偏移的步骤. 许多校正方法忽略了一个事实，在确定校正曲线的过程中，不应该包含具有显著性差异表达的基因或者探针，这些显著性差异表达的基因或者探针有可能对拟和校正曲线造成污染. 从而，又最终影响显著性差异表达基因的检测. 因此，拟和校正曲线的一个关键环节是选择非显著性差异表达基因或者探针. Schadt 等[4]提出了不变子集(noinvariant)的思想，通过一个迭代的过程选择目标芯片和基准芯片之间秩不变的探针集. 然后，在

选择的不变探针集上拟和校正曲线. 但是，Schadt 的不变探针集选择方法仅在某个单一的目标芯片和基准芯片之间选择. 一个更合理的考虑应该是在全部芯片上选择稳定的探针. IRB normalization 方法采用的策略是，对每个芯片上的探针进行排序，可以得到同一个探针在不同芯片中杂交强度的秩向量，计算秩的最大差异. 对于具有较小秩差异的探针，认为是稳定的. 由于探针数量一般都很大，所以，为了节约计算代价，可认为，取前 m 个最小秩差异探针就能够描述芯片的系统偏移. 通常，m 取 4 000～5 000 个. 通过选择最小秩差异探针子集，得到一个子探针杂交强度矩阵，然后，计算子探针杂交强度矩阵的伪基准芯片.

基准芯片的选择对芯片校正结果影响非常大. 一种较为合理的方法是计算同一探针在不同芯片中杂交强度的均值或者中值，作为在基准芯片上该探针的杂交强度.

考虑到芯片数据的高噪声特点，采用了 Tukey biweight 算法来估计同一探针在所有芯片中杂交强度的中值. Tukey biweight estimator 是 m-estimator 的一种形式，这是一种稳定的鲁棒估计方法，对于奇异点有很好的鲁棒性[12]. Tukey biweight estimator 定义的标准化函数如下：

$$u_{ij}=\frac{\log_2(y_{ij})-Me_j}{cS_j+\varepsilon} \tag{3}$$

其中，y_{ij} 表示探针 j 在芯片 i 上的测量杂交强度，Me_j 定义为探针 j 在所有芯片上杂交强度的对数中值，S_j 定义为探针 j 相对于其对数中值 Me_j 的绝对残差的中值，c 是调整常数(一般在 4 到 12 之间)，ε 是一个极小数，避免零除问题.

相应的权重函数为：

$$w(u)=\begin{cases}0 & |u|>1\\ (1-u)^2 & |u|\leqslant 1\end{cases} \tag{4}$$

最后计算探针杂交强度的鲁棒估计中值

$$\log_2(y_j^B)=\frac{\sum_{i=1}^{N} w(u_{ij})\log_2(y_{ij})}{\sum_{i=1}^{N} w(u_{ij})} \tag{5}$$

将 y_j^B 定义为探针 j 在基准芯片上的杂交强度. 由 Tukey biweight estimator 计算每个探针在所有芯片中测量杂交强度的中值，从而，得到一个伪基准芯片.

从 *M-A* 散点图，可以看出芯片之间的偏移随杂交强度呈现一定的非线性变化. 因此，采用局部多项式回归估计(local polynomial regression fitting, LOESS)来拟和校正曲线. 选择一个目标芯片，提取前面所确定的最小秩差异探针子集，在目标芯片与基准芯片之间计算探针的 M 和 A，利用 LOESS 方法拟和 M 随平均杂交强度 A 变化的校准曲线. 然后利用此曲线对目标芯片进行校正.

对于上述两个步骤，采用了一个迭代的过程. 定义收敛准则为：当用于拟和的最小秩差异探针集的最大修正量小于某个阈值时，校正过程收敛. 为了约束算法的时限，定义最大迭代数.

下面给出算法流程：

1) 对经过背景修正后的探针数据进行对数变换 $Y\leftarrow\log_2 Y$；

2) while(大于阈值或者未达到最大迭代数)

(1) 选择最小秩差异探针集；

(2) 利用 Tukey biweight 求解伪基准芯片；

(3) for $i<=N$ do (注:N 为芯片数量，P 为芯片上的探针数量, m 为最小秩差异探针数量)

a. for $j<=m$ do

计算 $M_{iB}^j=y_{ij}-y_{Bj}$ 和 $A_{iB}^j=\frac{1}{2}(y_{ij}+y_{Bj})$

end

b. 基于 loess 方法, 在 A 上拟和 M, 得到校正曲线 $\hat{f}(A)$；

c. for $k<=P$ do

$y_{ij}=y_{ij}-\hat{f}(A_{iB}^j)$

end

end

end

3)对校正数据进行反对数变换 $Y\leftarrow 2^Y$

3 实验分析

3.1 实验数据

Affymetrix HG_U133A Spike-in Dataset 是 Affymetrix 公司提供的标准检验数据，主要用于对芯片表达值算法的验证和比较. Affymetrix HG_U133A Spike-in Dataset 包含 42 个 HG_U133A 芯片. 这个实验是在人类细胞的背景下，将 42 个校验转录产物分成 14 组，分别以 14 个不同杂交浓度设计进行试验，每个浓度设计重复 3 次. 42 个校验转录产物的试验浓度采用 Latin Square 设计. 采用的浓度值为: 0、0.125、0.25、0.5、1、2、4、8、16、32、64、128、256、512 pmol/L.

因此，可以通过比较芯片转录产物的检测表达水平与期望表达水平之间的差异，来评估各种校正

方法的性能. 为了能够在表达值水平对校正算法进行评估，所有校正算法都在 RMA(Robust Multi-array Analysis)统一的处理框架下对芯片数据进行处理，最终得到一个转录产物表达值数据集. 首先对芯片数据进行背景修正，然后利用各种校正算法在探针层次对数据进行校正处理，最后，采用 median polish 方法对探针集进行综合得到基因表达值(在 bioconductor 的 affy 程序包中包含了文中提到的所有方法的 R 源代码，网址：http://www.bioconductor.org/). 对重复的芯片，通过求均值综合为一个芯片，可以得到 14 个表达值芯片. 由图 2 显示的预处理后的芯片表达值两两比较 *M-A* 散点图，可以直观地看到，经过处理后，芯片的系统偏移已经明显减小.

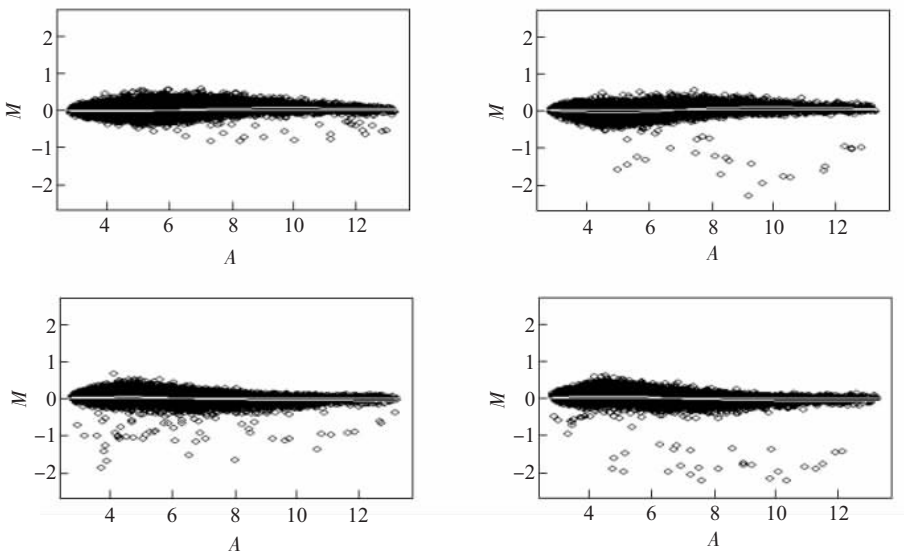

Fig. 2 4 pairwise *M* vs. *A* plots using HG_U133A Spiked-in data by IRB normalization

3.2 非校验转录产物比较分析

在经过 RMA 处理和重复芯片综合后，得到的 14 个不同浓度设计的表达值芯片，总共有 96 种不同浓度设计的两两比较. 对每个两两比较，计算所有非校验转录产物的对数比例差异表达 *M*. 表 1 给出了各种方法处理后，非校验转录产物全部两两比较的 *M* 的分位数级差均值(internal quantile range, *IQR*). 所有芯片都是以人细胞为背景，所以，认为所有非校验转录产物在所有芯片中应该是非显著性差异表达的. 两两比较的对数比例差异表达 *M* 的 *IQR* 越小，芯片间的变异就越小.

Table 1 *IQR* of fold-change estimates for non-differential probesets

Method	*IQR*
None	0.1319
Scaling	0.1300
Nonlinear	0.0916
Contrasts	0.1044
Quantile	0.0887
Cyclic loess	0.0919
IRB	0.0891

从表 1 可以看出，同非校正数据相比，各种校正方法都减少了芯片间的变异. 本文方法得到了仅次于 Quantile normalization 方法的 *IQR* 最小值. 其他几种方法也得到了比较好的结果. Scaling normalization 方法的效果相对较差，这也说明芯片之间的系统偏移呈现一定的非线性.

在 *M-A* 散点图中，对 *M* 拟和 LOESS 曲线，可以观察处理后的总体偏移情况. 图 3 显示不同方法处理后，所有两两比较的 *M* 绝对值的平均 LOESS 曲线. 理想状态下，LOESS 曲线应该和 $M=0$ 重合. 可以看到，Scaling normalization 方法明显很差，特别是在低探针杂交强度分布区. 同时发现，由于在高杂交强度段的样本量相对较小，各种方法均呈现上扬趋势. IRB normalization 方法在大部分区间的性能明显优于其他方法. 在高杂交强度末端，其偏离 $M=0$ 的程度要略高于其中几种方法. 可能是在高杂交强度末端存在较多的过饱和杂交强度探针，在最小秩差异子集的选择中未能入选，从而使得在高杂交强度末端的曲线拟和中样本量偏少. 过饱和杂交强度探针在后续的分析中一般要被过滤掉.

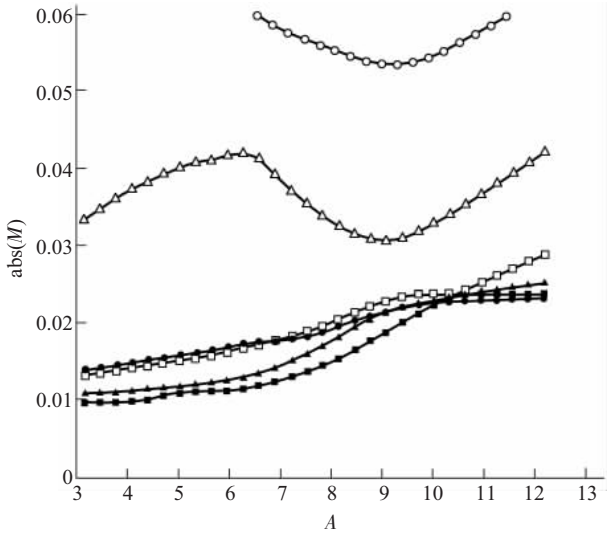

Fig. 3 Comparing the ability of methods to reduce pairwise differences between arrays by using average absolute distance from loess smoother to *x* axis in pairwise *M* vs. *A* plots using spike-in dataset.

○—○: Scaling; △—△: Contrast; □—□: Cyclic; ▲—▲: Quantile; ●—●: Nonlinear; ■—■: IRB

3.3 校验转录产物表达值分析

在数据中，已知 42 个探针集在不同芯片中的浓度. 利用这些已知的浓度信息，作为一个真实的标准，评估各种校正算法对表达值估计的影响. 拟和下面的线性模型

$$\log_2(E) = \beta_0 + \beta_1 \log_2(c) + \varepsilon \qquad (6)$$

其中，c 定义为表达产物的浓度，E 定义为表达产物的测量表达值. 理想的结果是斜率 β 应尽可能靠近 1. 斜率越靠近 1，校正后的偏移就越小. 由拟和模型得到的测量表达值估计与测量表达值的 R^2 度量越靠近 1，表明测量表达值与浓度之间的变异就越小. 表 2 显示了不同校正方法处理后，校验探针集测量表达值与已知的探针集浓度之间线性拟和模型的斜率和 R^2 估计值. 本文方法得到了最大的拟和斜率 β 和最小的拟和变异估计 R^2.

Table 2 Regression slope and R^2 estimates for spike-in transcripts

Method	Slope	R^2
None	0.679	0.971
Scaling	0.679	0.971
Nonlinear	0.678	0.971
Contrasts	0.680	0.970
Quantile	0.678	0.971
Cyclic loess	0.678	0.970
IRM	0.682	0.971

3.4 显著性差异表达转录产物的检测

检测显著性差异表达转录产物是芯片数据分析的一个重要方面. 在 HG_U133A 数据中，同一校验转录产物的浓度在不同芯片间成倍差分布，因此，认为其在不同芯片间的比例差异表达是显著的. 由于在实际的检测处理中，还可能涉及一些其他方面的处理，比如：异常点检测，基因过滤等，这超出了本文的讨论范围. 文章只是分析了在两两对比中，*M* 绝对值最大的前 *K* 个探针集中检测到校验转录产物的数量. 表 3 列出了不同校正算法处理后,平均检测到的校验转录产物的数量.

Table 3 Average numbers of spiked-in transcripts detected for dataset adjusted by several methods

Method	Number of detected transcripts (K=50)
None	30.186
Scaling	30.340
Nonlinear	30.362
Contrasts	30.319
Quantile	30.431
Cyclic loess	30.450
IRB	30.489

显然，检测到的校验转录产物数量越大，校正方法对差异表达基因的污染就越小. 各种方法都明显提高了差异表达检测精度. 本文方法得到了最大的检测数量. 同时，可以看到几乎对于每种方法，仍然有差不多 10 个转录产物没有检测到，主要是因为低浓度产物受噪声的影响特别强. 从试验的浓度设计可以看出，在任意两个芯片的对比分析中，总存在一定数量校验转录产物的低浓度与低浓度的差异比较. 因此，这部分低浓度与低浓度的差异比较很难直接检测到. 实际上，弱信号的检测也是当前表达谱芯片处理的一个难点.

4 结　论

在多芯片对比试验中，由于多种非生物变异因素使得芯片之间存在系统偏移. 系统偏移对于多芯片分析中差异表达基因的筛选，疾病组织样本的分类诊断，以及调控网络模型的构建等等后续分析都有很强的影响. 为了得到有生物学意义的分析结果，必须对原始芯片数据进行校正，消除这些非生物因素引起的系统偏移. 文章提出了一种新的校正算法. 并在标准测试数据上，与几种广泛采用的寡聚核苷

酸芯片校正算法进行了对比分析. 在多个方面的对比分析中，本文方法均显示了更好的性能. 本文方法在基因芯片研究中广泛采用的开放式软件 Bioconductor 上实现，具有较强的实用性. 为寡聚核苷酸芯片数据校正处理提供了一种新的实用方法.

参 考 文 献

1 Hartemink A, Gifford D, Jaakkola T, *et al*. Maximum likelihood estimation of optimal scaling factors for expression array normalization. International Symposium on Biomedical Optics. In: Bittner M, eds. Microarrays: Optical Technologies and Informatics. California USA: SPIE, 2001, **4266**: 132～140

2 Rocke D M, Durbin B. A Model for measurement error for gene expression arrays. J Comput Biol, 2001, **8** (6): 557～569

3 Affymetrix Microarray Suite User Guide, Version 5.http://www.affymetrix.com/support/technical/manuals.affx.

4 Schadt E, Li C, Eliss B, *et al*. Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data. J Cell Biochem, 2002, **84** (S37): 120～125

5 Magnus Å strand. Contrast normalization of oligonucleotide arrays. J Comput Biol, 2003, **10** (1): 95～102

6 Dudoit S, Yang Y H., Callow M J, *et al*. Statistical methods for identifying genes with differential expression in replicated cDNA microarray experiments. Stat Sin, 2002, **12** (1): 111～139

7 Li C, Wong W H. Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. Genome Biol, 2001, **2** (8): 1～11

8 Li C, Wong W H. Model-based analysis of oligonucleotides arrays: expression index computation and outlier detection. Proc Natl Acad Sci USA, 2001, **98** (1): 31～36

9 Irizarry R A, Bolstad B M, Collin F, *et al*. Summary of affymetrix GeneChip probe level data. Nucleic Acids Res, 2003, **31**(4): e15

10 Irizarry R A, Hobbs B, Collin F, *et al*. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics, 2003, **4** (2): 249～264

11 Bolstad B M, Irizarry R A, Astrand M, *et al*. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics, 2003, **19** (2): 185～193

12 Hubbell E, Liu W M, Mei R. Robust estimators for expression analysis. Bioinformatics, 2002, **18** (12): 1585～1592

A Robust Method to Normalize System Bias for High-density Oligonucleotide Array Gene Expression Profile*

QIU Lang-Bo[1,2)**], WANG Guang-Yun[1,2)], WANG Zheng-Zhi[1)]

([1)] *College of Mechatronics Engineering and Automation, National University of Defense Technology, Changsha* 410073, China;

[2)] *Telecommunication Engineering Institute, Air Force Engineering University, Xi'an* 710077, China)

Abstract In multiarray experiments, there is some system bias, which contaminated by experimental factors such as spot location (often referred to as a print-tip effect), arrays, dyes, and various interactions of these effects. For comparable each other, it is necessary to normalize the raw expression profile data. Normalization is the key step in low level processing. In fact, many normalization methods have been developed, i.e. Scaling normalization, Nonlinear normalization, Quantile normalization and so on. New baseline normalization is presented. First, select the subset of probes, which have the min rank range. Second, do nonlinear normalization on robust baseline. Iterative strategy weakens the sensitivity of the baseline method to select baseline. With the standard test dataset, compare it with other methods. The results show that the novel method has better performances than others in several ways.

Key words high-density oligonucleotide array, system bias, normalization, transcript

---

*This work was supported by a grant from The National Natural Sciences Foundation of China (60471003).

**Corresponding author . Tel: 86-731-4574991, E-mail: qlbogfkd@nudt.edu.cn

Received: December 9, 2005 Accepted: January 28, 2006