PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2007, 34(6): 625~633
www.pibb.ac.cn

芯片数据标准化方法比较研究 *

谈效俊 [1)] 张永新 [3)] 钱敏平 [1,2)] 张幼怡 [3)**] 邓明华 [1,2)**]

([1)]北京大学理论生物学中心，北京 100871；[2)]北京大学数学学院，北京 100871；
[3)]北京大学第三医院血管医学研究所，分子心血管学教育部重点实验室，北京 100083)

摘要 在基因芯片实验中，基因表达水平之间的相关性在推断基因间相互关系时起到非常重要的作用. 未经标准化处理的芯片数据基因之间往往都呈现出很强的相关性，这些高相关性一部分是由基因表达水平变化引起的，而另外一部分是由系统偏差引起的. 对芯片数据进行标准化处理的目的之一是消除系统偏差引起的高相关性，同时保留由真正生物学原因引起的基因表达水平高相关性. 虽然目前对标准化方法已经有了不少比较研究，但还较少有人研究标准化方法对基因之间相关系数的影响，以及哪种方法最有利于恢复基因之间的相关性结构. 通过对基因表达水平数据的模拟，具体比较了几种常用标准化方法的效果，从而给出最有利于恢复基因之间相关性结构的那种标准化方法.

关键词 基因芯片，标准化，相关系数
学科分类号 Q612

基因芯片自问世以来就受到生物学界的广泛关注. 它能高通量地同时测量生物体中成千上万个基因的表达水平，为生物学家研究疾病机理，发现致病基因，预测基因功能和推断信号转导途径等提供了有力的武器[1]. 与此同时，对于生物学和统计学工作者而言，如何分析和解释基因芯片技术所产生的大量数据也形成了一个挑战[2].

现今的基因芯片数据处理步骤大体上可分为 4 个主要阶段：图像分析和数据提取，数据标准化，挑选差异表达基因，以及后续的芯片数据挖掘，如信号转导途径分析等[3]. 其中数据标准化这一步在生物芯片数据分析中承前启后，对分析的最终结果有着巨大影响[4]，因此引起了芯片数据分析工作者的广泛兴趣.

芯片数据标准化算法的核心思想是寻找一些在各张芯片上保持恒定不变的量，芯片上的其他值都根据这些量进行相应调整，以使不同芯片的测量数据能够进行比较. 比如使用“看家基因”进行芯片数据标准化时，研究人员假设“看家基因”在不同芯片上的表达水平保持恒定不变，其他各个基因的测量表达水平都根据“看家基因”的表达水平进行相应的调整. 而另外一些标准化方法则是假设在生物实验中，生物体内改变表达水平的基因只占全基因组基因非常小的一部分. 这样，进行芯片数据标准化时比照的标准不再是部分“看家基因”，而是全基因组中所有的基因. 近年来有研究发现，“看家基因”的表达水平并非如所假定的那样总是保持恒定的表达水平[5]，因此研究人员越来越倾向于使用基于第二种假设的标准化方法.

对于 Affymetrix 公司生产的寡核苷酸单色芯片，目前有很多基于第二种假设的标准化方法. 这些标准化方法之中，最为有名的要数 Yang 等提出的 loess 方法[6,7]，Bolstad 等[7]提出的 quantiles 方法，Workman 等[8]提出的 qspline 方法以及 Li 等[9]提出的 invariantset 方法. 另外还有一种直观的方法也很常用，即将每张芯片上的表达数据做一个变换，使其均值为 0，标准差为 1，我们这里称该方法为 mean 方法.

一直以来，人们都试图对这些方法进行理论分析，并设计出种种标准来比较各种方法的优劣. 比如，如果假设只有很少数基因在不同的条件下会改

*国家自然科学基金资助项目(30570425, 30400552), 国家重点基础研究发展规划资助项目(2003CB715903, 2006CB503806), 微软亚洲研究院开放项目部分资助.
** 通讯联系人. Tel: 010-62767562
E-mail: zhangyy@bjmu.edu.cn;dengmh@pku.edu.cn
收稿日期：2006-12-12，接受日期：2007-02-28

变表达水平，大部分基因的表达水平实际上是恒定不变的，那么经过适当的标准化之后，不同的芯片所得数据应该具有相近的分布. 于是标准化之后的数据分布是否相似可以作为一个评价标准化方法的标准[7]. 另外，研究人员还提出了其他一些衡量标准，如单个基因表达水平的方差[7,10]，不同芯片数据之间的相似度(用欧式距离来衡量)[11]等等.

近来，也有一些研究小组使用实验的方法来研究不同标准化方法的效果[12~14]. 在这些实验中，研究者可以事先确定芯片上一些点的表达水平，人工制造一些“差异表达基因”. 芯片数据标准化之后，通过比较挑选出来的差异表达基因的准确程度，可以比较各种标准化方法的优劣. 总体看来，用实验方法来比较各种标准化方法更加有说服力，但是它也受到一些客观条件，如经费，人力等等的限制，因此这类实验一般规模比较小，有时甚至只有 3 张或 6 张芯片，这么小的样本量所得的结果往往统计显著性不高，结论也可能需要进一步讨论.

在基因芯片实验中，挑选差异表达基因往往并非唯一目的，通过聚类分析等方法来推断基因之间的相互关系也是很常见的[15,16]. 在这类研究中，一个基本的问题是如何确定基因表达水平之间的相关系数. 在未标准化的芯片数据中，基因之间相关系数往往都很高，这其中必然有很大一部分的高相关性并非是由生物学原因引起的[17,18]. 在采取标转化方法消除这些高相关性的时候，人们关心该标准化方法能否最大程度地恢复基因之间的相关性结构.

Ploner等[18]提出了一个判断芯片数据预处理方法是否能够重建基因之间相关性结构的标准，他们认为如果一组芯片实验覆盖了全基因组的大部分基因，那么从中随机抽取的一对，其相关系数平均而言应该为 0，他们使用这个标准比较了几种芯片数据预处理方法. 虽然如此，Qiu 等[17]担心标准化处理会破坏由生物学原因引起的基因表达水平的高相关性，因此可能会影响到对基因之间相互关系的推断.

本文观察了标准化处理对基因表达水平之间相关性的影响(包括由生物学原因引起的高相关性和由于系统偏差引起的高相关性)，并提出了一种基于相关系数的评价标准，即认为能够使标准化之后的相关系数最接近于真实相关系数的标准化方法是最好的方法. 通过对实验数据的模拟，我们对比了各种标准化方法的表现，由此对前文所列举的 5 种芯片数据标准化方法进行了比较.

1 方法和数据

1.1 芯片数据与数据预处理

本文使用的是 SJCRH 数据库[19]提供的 Affymetrix 芯片数据. 整个数据库共包含 9 组，共 335 张芯片，其中每组芯片的数量从 11 张到 88 张不等(表 1)，每张芯片上包含的基因数目为 12 558 个. 我们使用 RMA 方法[20]从原始数据文件(cel 文件)中提取基因表达水平.

Table 1 Nine groups of chips in SJCRH database

Groups of chips	BCR	E2A	Hyperdip	Hypodip	MLL	Normal	Pseudodip	TALL	TEL
Number of chips	16	27	88	11	21	19	29	45	79

这些芯片数据中，E2A 组芯片由于规模适中，被用来估计模拟实验中的各种参数. 另外，各组数据还被用来与模拟数据对照以观察模拟结果是否合理.

1.2 整体框架

为了研究各种标准化方法对基因之间相关系数的影响，需要比较标准化之后基因间的相关系数与真实相关系数. 实验数据中基因间的真实相关系数是无法得知的，我们通过模拟的手段来解决这个问题.

在每次模拟实验中，我们生成 12 758 个基因在 27 个时间点下的表达水平. 其中 12 558 个基因之间的理论相关系数为 0，另外 200 个基因彼此之间的理论相关系数为 r. 对这样一组模拟数据，我们分别采用 loess, invariantset, qspline, mean 和 quantiles 等标准化方法处理，然后计算这 200 个基因之间的相关系数与 r 的差别，分别记做 d_{lo}，d_{in}，d_{qs}，d_{me}，d_{qu} (图 1). 较好的标准化方法会产生较小的 d 值，这样，通过比较 d 值，便可以对标准化方法进行比较和评价.

为了保证比较结果的可靠性，我们进行了多次实验：分别取 r 为约 0，0.1，…，0.9 等 10 个值，在每个 r 下进行 20 次模拟实验. 这样，实验次数总计 200 次. 最后，对这 200 次的实验结果进行总结以比较各种标准化方法.

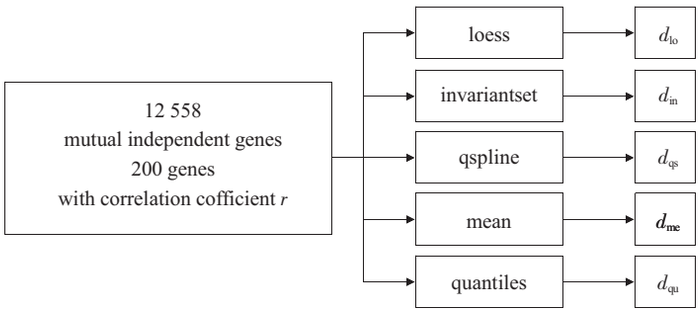

Fig. 1　Comparison of different normalization methods using simulated data

1.3　芯片数据模拟

设第 j (j = 1，2，…，M)张芯片对第 i (i = 1，2，…，N)个基因表达水平的测量值为 $\hat{g}_{ij}$. 一般认为，$\hat{g}_{ij}$ 不仅反映基因的真实表达水平，还受到系统偏差和随机误差的影响，即

$$\hat{g}_{ij} = g_{ij} + s_{ij} + e_{ij} \quad (1)$$

其中 g_{ij} 为第 i 个基因在第 j 张芯片上的真实表达水平，而 s_{ij} 和 e_{ij} 则分别反映了系统偏差和随机误差的影响.

如果系统偏差为 0，则(1)式即为 $\hat{g}_{ij} = g_{ij} + e_{ij}$. 则由概率论的知识，如果 g_{ij} 与 e_{ij} 相互独立，则可以预测 2 个基因表达水平测量值之间的相关系数

$$r = \sigma_g^2/(\sigma_g^2+\sigma_e^2) \quad (2)$$

其中 σ_g^2是基因真实表达水平的方差，σ_e^2是测量随机误差的方差[21]. 本文称相关系数 r 为理论相关系数. 如果存在系统偏差，则基因之间的理论相关系数就会被扭曲而偏离 r.

我们按照(1)式和(2)式，分别生成 12 558 个表达水平线性无关基因和 200 个表达水平线性相关基因的模拟表达水平.

1.3.1　表达水平线性无关基因的模拟.

取 SJCRH 数据库中的 E2A 组作为参照数据，我们按照如下方法来估计式(1)右边的 3 组参数，并用这些参数生成与 E2A 组芯片实验同样规模的模拟数据(同样的基因数目和芯片数目)，并且模拟数据中的基因是两两不相关的. 估算 3 组参数步骤如下：

a. 估计 g_{ij}. 对于基因 i (i = 1，2，…，12 558)，取 $g_i=\frac{1}{M}\sum_{j=1}^{M}\hat{g}_{ij}$，这里 M=27. 为了保证基因之间的理论相关系数为 0，我们使模拟数据中所有基因的真实表达水平在各张芯片保持恒定，即取 $g_{ij}\equiv g_i$ (j=1，2，…，12 558).

b. 估计 s_{ij}. 对于每张芯片，使用 $\hat{g}_{ij}$对 g_{ij} 进行 3 次多项式回归，回归模型为 $\hat{g}_{ij}=a_jg_{ij}^3+b_jg_{ij}^2+c_jg_{ij}+d_j+e_{ij}$，取回归值 $\hat{g}_{ij}$与 g_{ij}的差别为系统偏差.

c. 模拟产生 e_{ij}. 将 E2A 组芯片实验数据使用 quantiles 方法进行标准化处理，用标准化之后的数据计算测量第 i 个基因表达水平随机误差的方差 σ_i^2 (各种标准化方法处理后，所得 σ_i^2会略有不同，但差别不大，对结论没有什么太大影响). 对于该基因，生成 27 个服从 $N(0，\sigma_i^2)$分布的随机数，记为 e_{ij}，其中 j=1，2，…，27.

1.3.2　表达水平线性相关基因的模拟.

a. 随机误差的模拟. 取 $\sigma_Y^2=\sum_{i=1}^{N}\sigma_i^2\times I(\sigma_i^2)\,/\sum_{i=1}^{N}I(\sigma_i^2)$，其中 σ_i^2 为第 i 个基因表达水平的方差(见 **1.3.1**)，$I(\sigma_i^2)=\begin{cases}1，\sigma_i^2\leqslant 2\\ 0，\sigma_i^2>2\end{cases}$，$N$=12 558. σ_Y^2的直观意义为芯片上所有基因表达水平测量随机误差的平均值，但其中大于 2 的不记在内，因为如此大的方差极有可能是由于生物学原因引起的，并不是由真正的随机误差产生的. 使用 σ_Y^2，对于每个基因，生成 27 个服从 $N(0，\sigma_Y^2)$分布的随机数，作为该基因的随机误差.

b. 真实表达水平的模拟. 取 $g_{ij}=g_i^0+x_j$，其中 g_i^0 (i=1，2，…，200)为第 i 个基因的基本表达水平，在各张芯片上保持不变，x_j (j = 1，2，…，27)则为基因表达水平在各张芯片上的变化，其方差 σ_x^2 的大小反映了基因真实表达水平的变化程度. 由式(2)，通过控制 σ_x^2的大小，可以控制基因之间理论相关系数的大小. 在具体模拟实验中，我们调整 σ_x^2 的值，使 r 取约 0，0.1，…，0.9 等 10 个值.

c. 系统偏差的模拟与 **1.3.1** 所述一致.

2 结 果

2.1 各种标准化方法的比较

在一组芯片模拟实验中，同时包含了大量彼此不相关的基因(12 558 个)和少数彼此之间理论相关系数已知的基因(200 个). 芯片数据经过标准化处理后，这些彼此相关的基因之间的相关系数与理论相关系数的差别是本文最为关注的部分. 较好的标准化方法会使标准化之后的相关系数更加接近于理论相关系数，因此我们可以通过比较标准化之后相关系数与理论相关系数的差别，可以比较各种标准化方法的效果.

对于 200 个已知理论相关系数的基因，共有 C_{200}^2个基因对. 为了度量标准化之后基因之间相关系数与理论相关系数的差别，我们定义距离变量 d 如下：

$$d=\sqrt{\frac{1}{K}\sum_{i=1}^{K}(r_i-r)^2} \tag{3}$$

其中 $K=C_{200}^2$，为所有可能的基因对的数目. r_i 为其中第 i 对基因的 Pearson 相关系数，r 为该组实验的理论相关系数.

我们分别取 r 为约 0，0.1，…，0.9 共 10 个值，对于每个值进行 20 次模拟实验. 对于每次模拟实验，我们都使用 loess，invariantset，qspline，mean，quantiles 等方法对数据进行标准化，然后计算每种方法标准化后基因之间相关系数和理论相关系数 r 之间的距离，分别记作 d_{lo}，d_{in}，d_{qs}，d_{me} 和 d_{qu}. 另外，为了观察标准化方法对相关系数的影响，我们还计算了标准化之前实际相关系数与理论相关系数的距离，记做 d_n.

图 2 简要概括了理论相关系数 r 取不同的值时，各种方法所得 d 的值. 图 2 中每个点的横坐标对应于理论相关系数，纵坐标对应于该理论相关系数下 20 次实验所得 d 值的平均值.

从图 2 可以看出，loess 方法在各种情况下都有最好的表现，而 mean 方法则稍逊于其他各种方法，invariantset，quantiles，qspline 等方法所示曲线十分相近，表明这 3 种方法具有相近的效果.

另外，通过对比 200 次实验的具体结果，可以对这些标准化方法进行更为详细的评估(表 2). 在 200 次模拟实验中，有 155(160)次实验 quantiles 方法所得的距离小于 invariantset(qspline)方法，这说明在大多数情况下，quantiles 方法效果要优于 invariantse (qspline) 方法. 而对比 qspline 和 invariantset 两种方法，200 次实验中只有 119 次 invariantset 方法的表现优于 qspline，因此总体而言，二者效果十分接近.

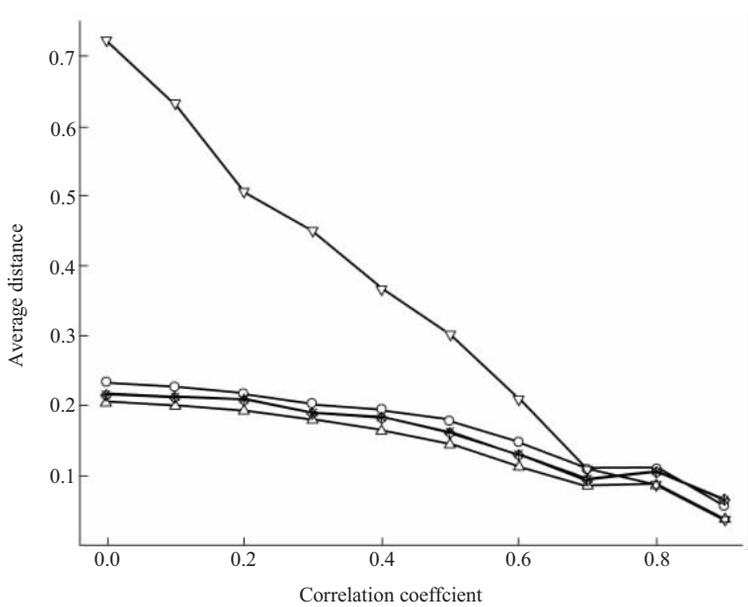

Fig. 2 Average distance between real and theoretical correlation coefficients

Different combinations of points and lines in the figure correspond to different normalization methods (the highest line in the figure corresponds to distances before normalization). ○—○: mean; △—△: loess; +—+: invariantset; ×—×: qspline; ◇—◇: quantiles; ▽—▽: before normalization.

Table 2 Comparison of normalization methods

Method	loess	quantiles	invariantset	qspline	mean
loess	–	0/200	0/200	0/200	0/200
quantiles	200/0	–	45/155	40/160	23/177
invariantset	200/0	155/45	–	81/119	25/175
qspline	200/0	160/40	119/81	–	24/176
mean	200/0	177/23	175/25	176/24	–

The first row and the first column contain the names of 5 different normalization methods, and the comparison results are in the crosses. For example, in the 3rd row and the 4th column, the figure is 45/155. This means that in 200 simulated experiments, we have 45 times $d_{qu}>d_{in}$ and 155 times $d_{in}>d_{qu}$.

2.2 标准化方法对相关系数的影响

与未标准化的数据相比，标准化处理降低了基因之间实际相关系数与理论相关系数 r 的差别(对比图 2 中的实线和其他各条曲线). 对于在各张芯片上表达水平变化不大的基因，它们之间的理论相关系数很低，标准化之前的相关系数与 r 相差很大，标准化处理对恢复这部分基因的相关性结构具有最大效果(图 2 的左边一部分). 而对于表达水平在各张芯片上变化很大的基因对，系统偏差对其相关系数的影响比较弱，标准化方法的效果也就不那么明显了(图 2 的右边一部分). 对于覆盖基因组大部分基因的芯片实验来说，芯片上大部分基因的表达水平变化都不显著，因此，标准化方法在实际中最重要的作用是使这些基因的相关系数重新回到 0 左右，而同时保留真正高相关的基因对.

对于图 2 中理论相关系数为 0 的点，它们对应的基因的表达水平在各张芯片上保持不变. 而在实际的芯片实验中，有大量基因的表达水平都保持恒定. 因此，通过观察这部分模拟基因相关系数的变化，可以对这部分基因之间相关系数在标准化前后的变化有所了解. 标准化之前，这部分基因之间的相关系数往往相当高，标准化处理则能很大程度上消除这种非生物学原因引起的相关性.

2.3 和其他研究结果的比较

本文的比较认为 loess 方法在恢复基因之间相关系数方面最具优势，这个结果在其他的一些研究中得到了部分印证. 在 2006 年，Kim 等[22]使用不同的标准化方法处理相同的数据，然后对处理过的数据进行聚类分析，他们发现，loess 方法最有利于得到一个具有鲁棒性的聚类结果. Choe 等[13]在 2004 年使用 6 张芯片进行了 spike-in 实验，然后用实验数据来评价各种数据处理方法，他们发现，loess 方法和其他方法相比(包括 quantiles 方法，invariantset 方法)，更有利于挑选差异表达基因. 这些研究结果都从一个侧面说明了 loess 方法具有相对的优越性，和本文的结果是一致的.

3 讨 论

3.1 本文对以往工作的改进

基因芯片数据标准化方法种类繁多，且对后续工作影响深远，因此是一个进行比较研究的热点，但是一直较少有工作以标准化方法对基因之间的相关性结构的影响作为标准. 这可能是由于基因之间相互关系十分复杂，其表达水平之间的相关性难以把握，并且基因之间的真实相关系数也无从得知所致.

有研究工作避开了特定基因对之间相关系数的问题，转而研究基因对之间的平均相关系数[18]. 这在一定程度上回答了标准化方法对基因相关性结构影响的问题，但是还需要进一步改进. 因为芯片数据经过标准化处理之后，即使基因表达水平之间的平均相关系数与期望值差别很小，也有可能会有某些基因表达水平之间的相关系数与真实相关系数差别很大. 如果是这样的话，则还不能判定该标准化方法是一个很好的标准化方法.

通过表达数据模拟，则可以具体观测到标准化方法对特定基因对相关系数的影响，从而可以更加直观地比较各种处理方法. 不但如此，本文所采用的方法还有助于探讨标准化方法对基因相关系数影响的一些规律，从而回答一些研究者提出的问题.

如 Qiu 等[17]提出标准化方法可能会破坏基因之间的相关性结构，消除了可能由生物学原因引起的高相关性. 本文认为标准化能够在一定程度上保留和恢复由生物学原因引起的相关性，并不会对由生物学原因引起的相关性产生根本性的破坏，众多使用芯片数据推断基因之间相互关系的成功案例则支持了本文的观点.

3.2 模拟方法的合理性

要使用模拟数据研究标准化方法对基因之间相关性的影响，首先要保证模拟数据能够切实有效地代表真实数据. 为此，我们研究了前文所述模拟方法中 2 个步骤：a 彼此线性无关基因的模拟和 b 彼此线性相关基因的模拟.

3.2.1 线性无关基因的模拟与系统偏差.

使用 **1.3.1** 中所述方法生成线性无关基因的模拟芯片数据时，由于所有基因的表达水平都保持恒定不变，芯片数据所体现出来的基因之间的相关性主要是由系统偏差引起的. 而在实际芯片实验中，也只有一小部分基因的表达水平发生了显著变化. 因此，在忽略那些表达水平发生变化基因的情况下，由 **1.3.1** 所述方法生成的模拟数据需要与真实数据具有相似的相关性结构，才能说明模拟方法能较好地体现系统偏差的影响.

为此，我们对照 SJCRH 数据库中的 9 组数据，使用 **1.3.1** 中所描述的方法模拟产生了相对应的 9 组模拟数据，然后分别从实验数据和模拟数据中各随机抽取 10 000 对基因，计算其 Pearson 相关系数，并比较二者之间相关系数分布. 以 E2A 组为例，图 3 显示，二者的分布极为相像. 图 4 使用 q-q 图对比了所有 9 组模拟数据和真实数据的相关系数分布，结果表明，模拟数据和真实数据具有十分相似的相关性结构，因此 **1.3.1** 所述的模拟方法较好地体现了系统偏差的影响.

值得注意的是，图 4 的 q-q 图也体现了模拟数据相关系数的分布和真实数据相关系数分布的细微差别. 这一方面是因为真实数据中包含部分彼此之间确有线性相关性的基因，而在图 4 的模拟数据中，则忽略了这些基因. 另外一方面，模拟数据对系统偏差采用了多项式近似，这不可避免地与真实系统偏差会有所区别，于是也会造成模拟数据与真实数据的相关系数具有不同分布. 但总体而言，图 4 表明，这些因素并不影响 **1.3.1** 所述方法的可行性.

3.2.2 线性相关基因模拟方法的合理性.

进行表达水平线性相关基因模拟的理论基础是公式(2)，这个公式在概率论中可以严格证明. 公式(2)反映了决定基因表达水平之间相关性的 2 个因素：一个是基因真实表达水平的变化强度(即 σ_X^2)，另外一个是芯片测量的随机误差的大小(即 σ_Y^2). 基因表达水平变化越明显，芯片测量的随机误差越小，则基因之间所体现出的相关性就会越强，相反，如果芯片测量的随机误差大，则会掩盖基因表达水平之间的相关性.

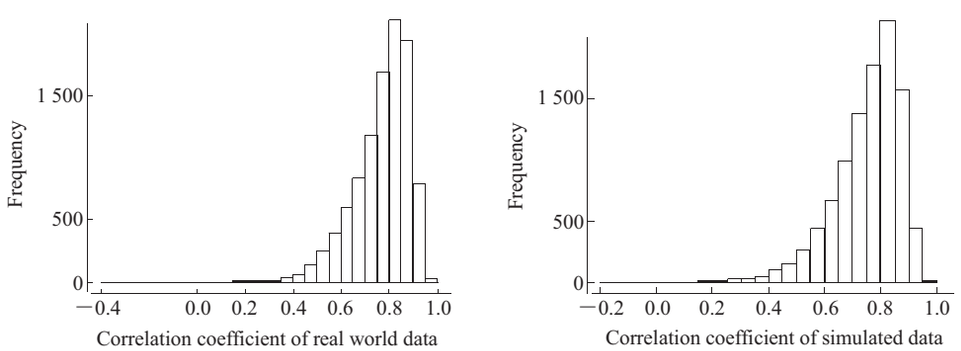

Fig. 3 Comparison of the distribution of correlation coefficient

10 000 gene pairs were randomly selected from the E2A group in the SJCRH database to compute their Pearson correlation coefficients, and the result is shown in the left histogram. In the right histogram, the simulated data were used. The similarity of these histograms reveals that the simulated data grasped the correlation structure of real world data.

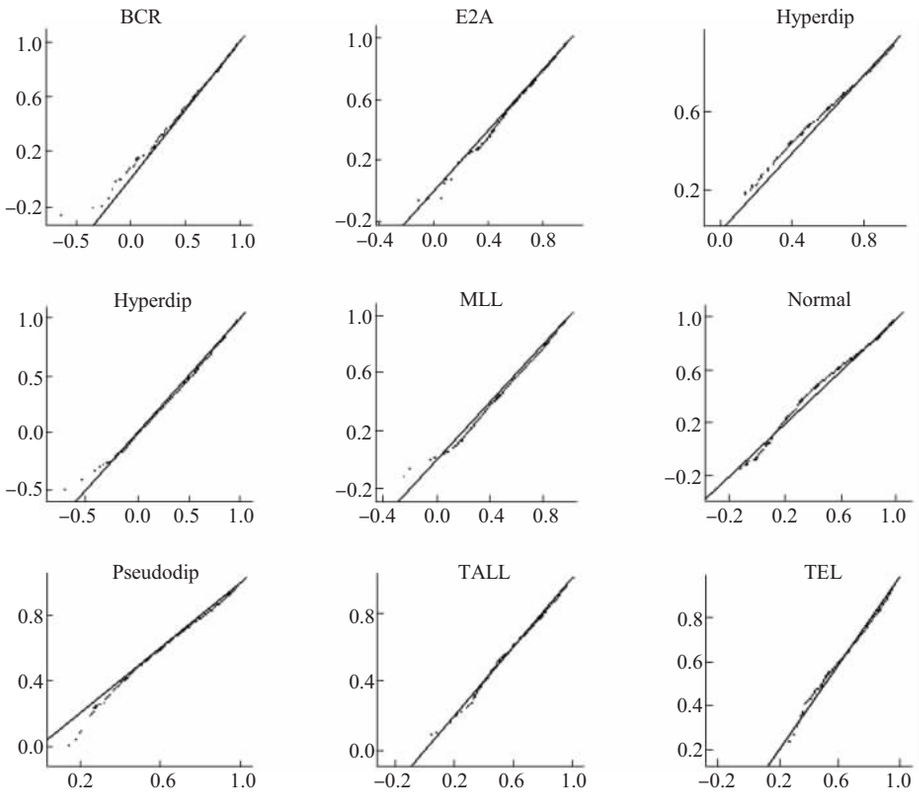

Fig. 4 Comparison of the correlation coefficient distribution of real world data and simulated data

The *x*-coordinates in these plots represent the correlation coefficients of real world data; while *y*-coordinate represent the simulated data. The title of each plot is the corresponding group name in the SJCRH database. In all these 9 q-q plot, points lie close to lines $y = x$, which means the correlation coefficient distribution of real world data and simulated data are very similar.

通过公式(2)，我们可以估计实际芯片实验中，未经标准化的芯片数据，其基因间的平均相关系数是多少. 估计结果的准确与否可以比较容易地验证.

我们以 SJCRH 数据库中的 Hypodip 组芯片为例简要说明进行这种估计的原理和过程. 对于该组芯片，计标准化之前第 i 张芯片上所测量的基因表达水平的平均值为 m_i，其中 i=1，…，11. 由于系统偏差的原因，一般而言，对于任意的 $i \neq j$，都有 $m_i \neq m_j$(图 5). 这种系统偏差会引起未标准化基因的表达水平之间呈现出一定的相关性. 虽然 m_i 的上下波动并不是由生物学原因引起的，但是如果我们在形式上把它看成是基因表达水平的真实变动，则可以利用公式(2)来估计对于该组芯片，用未标准化的数据计算出来的基因之间相关系数的平均值会有多大.

在这种情况下，基因表达水平的变化可以近似用 m 的方差 σ_m^2来衡量. 另外，利用 **1.3.2** 中介绍的对芯片测量随机误差大小的估计方法，可以估计出该组芯片测量的随机误差的方差为 σ_n^2. 于是，由公式(2)，可以预计对于 Hypodip 芯片组，标准化之前基因表达水平之间相关系数的平均值$\bar{r}_{Hypodip}=\sigma_m^2/(\sigma_m^2+\sigma_n^2)$.

为了验证估计结果的准确性，我们从未标准化的 Hypodip 组真实芯片数据中随机抽取了 10 000 对基因，实际计算了它们表达水平之间的 Pearson 相关系数，并对这 10 000 个相关系数求取了平均值. 结果表明，通过这种方法实际计算出来的相关系数的平均值与预测的平均值$\bar{r}_{Hypodip}$非常接近.

对于 SJCRH 数据库中的其他各组芯片，我们也进行了类似的计算和比较(图 6). 从图 6 可以看出，公式(2)的预测结果相当准确. 预测值和实际计算值的相关性高达 0.961. 这说明我们采用公式(2)作为理论基础模拟生成表达水平线性相关的基因，不但理论上是可靠的，实际上也是很准确的.

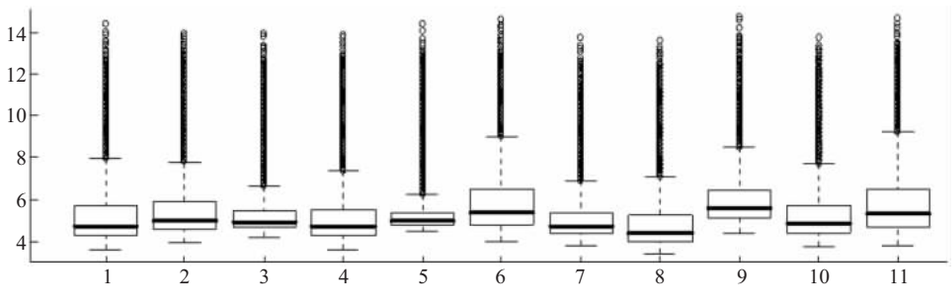

Fig. 5 Different chip has different average expression level before normalization

Each box-plot in this figure corresponds to a chip and illustrates the distribution of gene expression levels. The whole figure corresponds to the Hypodip group in SJCRH database.

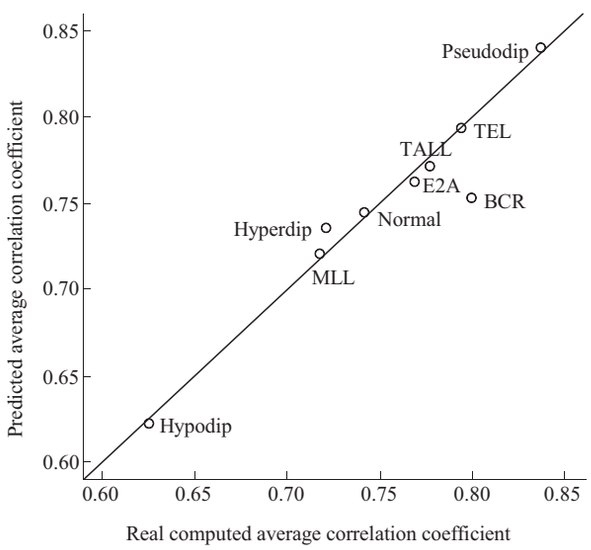

Fig. 6 For expression data before normalization, their average correlation coefficients can be precisely predicted

Each point in the figure corresponds to a group of chips, and the x-coordinate is the real computed average correlation coefficient, while the y-coordinate is the predicted average correlation coefficient. The line in the figure is $y = x$.

参 考 文 献

1 Imbeaud S, Auffray C. The 39 steps in gene expression profiling: critical issues and proposed best practices for microarray experiments. Drug Discovery Today, 2005, **10** (17): 1175～1182

2 Allison D B, Cui X Q, Page C P, *et al*. Microarray data analysis: from disarray to consolidation and consensus. Nature Reviews Genetics, 2006, **7**: 55

3 Wilson D L, Buckley M J, Helliwell C A, *et al*. New normalization methods for cDNA microarray data. Bioinformatics, 2003, **19** (11): 1325～1332

4 Hoffmann R, Seidl R, Dugas M. Profound effect of normalization on detection of differentially expressed genes in oligonucleotide microarray data analysis. Genome Biol, 2002, **3** (7): RESEARCH0033.

5 Lee P D, Sladek R, Greenwood C M T, *et al*. Control genes and variability: absence of ubiquitous reference transcripts in diverse mammalian expression studies. Genome Res, 2002, **12** (2): 292～297

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

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

8 Workman C, Jensen L J, Jarmer H, *et al*. A new non-linear normalization methods for reducing variability in DNA microarray experiments. Genome Biol, 2002, **3** (9): research0048

9 Li C, Hung Wong W. Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. Genome Biol, 2001, **2** (8): REASERCH0032

10 Park T, Yi S G, Kang S H, *et al*. Evaluation of normalization methods for microarray data. BMC Bioinformatics, 2003, **4**: 33

11 Wu W, Xing E P, Myers C, *et al*. Evaluation of normalization methods for cDNA microarray data by k-NN classification. BMC Bioinformatics, 2005, **6**: 191

12 Qin L X, Kerr D F. Empirical evaluation of data transformations and ranking statistics for microarray analysis. Nucleic Acids Research, 2004, **32**(18): 5471～5479

13 Choe S E, Boutros M, Michelson A M, *et al*. Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset. Genome Biol, 2005, **6** (2): R16

14 Ryden P, Andersson H, Landfors M, *et al*. Evaluation of microarray data normalization procedures using spike-in experiments. BMC Bioinformatics, 2006, **7**: 300

15 Segal E, Shapira M, Regev A, *et al*. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nature Genetics, 2003, **34** (12): 166～176

16 Eisen M B, Spellman P T, Brown P O, *et al*. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA, 1998, **95** (25): 14863～14868

17 Qiu X, Brooks A I, Klebanov L, *et al*. The effect of normalization on

the correlation structure of microarray data. BMC Bioinformatics, 2005, **6**: 120

18 Ploner A, Miller L, Hall P, *et al*. Correlation test to assess low-level processing of high-density oligonucleotide microarray data. BMC Bioinformatics, 2005, **6**: 80

19 Yeoh E J, Ross M E, Shurtleff S A, *et al*. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by expression profiling. Cancer Cell, 2002, **1** (2): 133～143

20 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

21 Irizarry R A, Warren D, Spencer F, *et al*. Multiple-laboratory comparison of microarray platforms. Nature Methods, 2005, **2** (5): 345～350

22 Kim S Y, Lee J W, Bae J S. Effect of data normalization on fuzzy clustering of DNA microarray data. BMC Bioinformatics, 2006, **7**: 134

The Comparison of Different Normalization Methods in Microarray Data*

TAN Xiao-Jun[1)], ZHANG Yong-Xin[3)], QIAN Min-Ping[1,2)], ZHANG You-Yi[3)**], DENG Ming-Hua[1,2)**]

([1)]*Center for Theoretical Biology, Peking University, Beijing* 100871, China;

[2)]*Key Laboratory of Pure and Applied Mathematics, School of Mathematics, Peking University, Beijing* 100871, China;

[3)] *Institute of Vascular Medicine, Peking University Third Hospital and Key Laboratory of Molecular Cardiovascular Science, Ministry of Education, Beijing* 100083, China)

Abstract Correlation coefficient between the expression levels of two genes plays an important role in the inference of their relationship in microarray experiments. Gene expression data before normalization often present high correlation coefficients among a large proportion of genes. Some of these high correlations are caused by changes in gene expression levels. However, most of them are caused by systematic errors. It is intended to eliminate superficial high correlations induced by systematic errors and at the same time, preserve high correlation coefficients stem from gene interactions. Although there are a number of comparisons among different normalization methods, less work focused on evaluating the effect of normalization procedures on correlation coefficients among genes and which method does the best in restoring gene correlation structure. Some gene expression data were simulated with reference to real world gene expression data. With the help of these simulated data, it was determined which normalization method does the best in restoring gene correlation structure. In addition, it was shown that the simulated data and the real world data have the same gene correlation structure, so the conclusion drawn from simulated data can be applied to the real world. For 5 normalization methods compared here, it can be concluded that the loess method is the most appropriate one in eliminating superficial correlation coefficients.

Key words microarray, normalization, correlation coefficient

---

*This work was supported by grants from The National Natural Science Foundation of China (30570425, 30400552), The National Key Basic Research Project of China (2003CB715903, 2006CB503806), and supported in part by Microsoft Research Asia (MSRA).

**Corresponding author . Tel: 86-10-62767562, E-mail: zhangyy@bjmu.edu.cn;dengmh@pku.edu.cn

Received: December 12, 2006 Accepted: February 28, 2007