[19]中华人民共和国国家知识产权局

[51]Int. Cl[7]
G01N 33/00
C12Q 1/68

[12]发明专利申请公开说明书

[21]申请号 99805289.2

[43]公开日 2001年6月6日

[11]公开号 CN 1298486A

[22]申请日 1999.4.22 [21]申请号 99805289.2
[30]优先权
[32]1998.4.22 [33]US [31]60/082,692
[86]国际申请 PCT/IB99/00734 1999.4.22
[87]国际公布 WO99/54724 英 1999.10.28
[85]进入国家阶段日期 2000.10.20
[71]申请人 图像研究公司
地址 加拿大安大略
[72]发明人 彼得·拉姆
罗伯特·纳栋

[74]专利代理机构 中国国际贸易促进委员会专利商标事务所
代理人 鄢 迅

权利要求书1页 说明书57页 附图页数8页

[54]发明名称 评价化学和生物学分析的方法

[57]摘要

公开一种分析过程,用于鉴别从带有重叠分布的样本得到的数据,并用于改进和评估测定阵列中的杂交信号的统计有效性。该过程包括 把数据转换到二个或更多的代表信号和非信号的分立概率密度函数、各离散荧光或其它转接后的独立变量的方法。该系统利用概率密度函 数客观地把杂交信号分配到建好模型的多个分布中的一个分布。相继 的过程评估阵列固有的变异性,并利用评估的偏差为整个杂交阵列以 及为阵列内的离散杂交测定建立可靠性评分和置信界限。

ISSN1008-4274

权 利 要 求 书

1.一种用于对从阵列杂交研究得到的数据做出统计上有效论断的方法，该方法利用大量的基因组样本，每种样本由少量的不足做出准确和有效的统计论断的重复构成，该方法包括通过对从大量样本上得到的各误差估计取平均值对一个样本估计误差的步骤。

2.权利要求 1 的方法，其中利用从整个阵列群体计算的统计量估计该群体内小量重复样本的统计量的分立实例。

3.权利要求 1 或 2 的方法，其中利用测量误差的估计生成统计功效和试验敏感性分析。

4.一种方法，其利用权利要求 1 或 2 获得的误差估计确定局外点。

5.一种用于区别杂交阵列内的诸如信号和非信号的响应类别的方法，其包括对代表着和从阵列杂交研究中得到的数据有关的论断的分布重叠部分进行消卷积的步骤。

6.权利要求 5 的方法，其中通过数学混合建模对阵列内的双分布和多分布中的一个分布建模。

7.权利要求 5 或 6 的方法，其中利用混合建模以得到有关任一独立阵列单元落入建了模型的分布中的一个分布内的概率的论断。

8.权利要求 1、2、5 或 6 中任一要求的方法用于对从在孔板、试管或其它介质之一中进行的生物学和化学测定中得到的数据作出有效论断。

9.权利要求 3 的方法用于对在孔板、试管或其它介质之一中进行的生物学和化学测定中得到的数据作出有效论断。

10.权利要求 4 的方法用于对在孔板、试管或其它介质之一中进行的生物学和化学测定中得到的数据作出有效论断。

11.权利要求 7 的方法用于对在孔板、试管或其它介质之一中进行的生物学和化学测定中得到的数据作出有效论断。

说　　明　　书

评价化学和生物学分析的方法

本发明涉及对从杂交阵列中得到的数据进行客观化分析的评价方法。在一个方面上本发明是一种用于推断由少量数据点构成的重复基因组样本中存在的随机误差的程度的方法，而在另一个方面上本发明是一种用于区分不同类别的探针强度（例如，有信号和无信号）。

基于阵列的遗传分析从在基片上固定大的 cDNA 或寡核苷酸库（探针）着手。用单个标记的序列或用从组织或细胞系信使 RNA（靶标）而来的标记的复杂混合物杂交探针。如本文中所使用的那样，术语“探针”应理解成表示固定在该阵列中的材料，而术语“靶标”表示施加到阵列上的各探针的材料，从而会出现杂交。

存在二种类型的测量误差即随机的和系统的。通过对相同过程或属性进行反复测量可检测出随机误差，并通过统计方法予以处理。低随机误差对应于高精度。不能通过重复测量检测系统误差（漂移或偏差）。低系统误差和高准确度对应。

背景修正涉及从探针中减去探针外某区域的强度。用于计算背景的区域可以接近探针（例如绕探针的一个圆周），或者离开一个距离。例如，可以建立“空白”单元（即，不带有探针材料的单元），这些单元的值可用于背景估算。归一化处理涉及通过某基准强度划分探针。最常用地，该基准是从一组探针或从所有探针的均值得到的。

一旦通过背景消除和归一化处理（或若需要时其它处理）去掉系统误差，任何剩下的测量误差在理论上是随机误差。随机误差反映测量值中的期待统计偏差。测量值例如可包括单个值、多个值的累计（均值，中值）、单个值和累计值之间的差或差值之间的差。为了把二个值可靠地看成是彼此不同，它们的差必须超过一个阈值，该阈值是由与该差相关的测量误差以及错误地得出这二个值不同的规定概率（类型 I 出错率）联立定义的。

主要的兴趣在于二个或多个典型地在不同条件下（例如，有病的对无病的细胞系、加药的和不加药的）量化值之间的差。理想地应从相同量的重复测量值显示出的偏差得到所想要的期待随机误差的估计。这是常规地在其它科学领域中使用这样的估计的原因。但是，在杂交研究中，趋向于使用很少数量的重复（例如，二或三个）。基于如此少的样本的随机误差估计本身是很易变的，这使得对于几乎很大的差值利用标准统计测试进行不同条件间的比较是不准确的和不实用的。

Bassett、Eisen和Boguski在“Gene expression informatics: It's all in your mine”, Nature Genetics, 21, 51-55(1999)一文中认识到这种困难，他们论述：提供基因表达数据的最大挑战性涉及各表达值的量化和鉴定而且鉴定应包括标准的统计显著性检验和置信区间。他们还提出“理想地，重复足够次数的试验从而可以给出和每个转录本水平相关的偏差在经济上是可行的”（P.54）。上述引用中的短语“足够次数”是问题的中心。基于阵列的研究的现有技术状态使得不可能得到标准的统计指数（例如，置信区间、局外描述）并且不可能进行其它科学领域中为例行的标准统计检验（例如 t－检验、方差分析），因为该研究中典型出现的重复次数普遍被认为对于这些目的是不充分的。本发明的关键是克服该困难。

需要统计指数以及统计检验以便可以对不同条件下所观测的探针/靶标交互作用之间差异的可靠性作出评估。这类比较的关键问题是测量值中观测到的差异是否可能只反映随机误差或反映和处理效果相结合的随机误差（即，“真实的不同”）？在缺少用于在这些供选用方案之间作出决定的正式统计处理下，非正式的处理已在现有技术中得到进展。下面概述这些处理方法：

1. 任意阈值。通过一个任意阈值区分不同条件下的观测差异。例如，大于2倍或3倍的差异判定成反映“真正的”差异。
2. 建立与阵列单元的子集相关的阈值。为感兴趣的比率把“基准”基因子集作为比较点。例如，对于基准基因，在时间1测

量时某基因可能示出 2：1 的表达比率，而在时间 2 测量时示出 2.8：1 的比率。

3. 根据背景中的观测偏差建立的阈值。背景值的标准偏差充当和感兴趣的探针值相关的测量误差标准偏差的代表。若某探针强度超过背景标准偏差一规定值（例如，2.5），该探针看成是“重要的”。

上述各处理方法都不是最优的，因为每一种都依赖于用于导出推断规则的相对少的观测次数。而且，置信评定是主观的并且不能针对“机会”统计模型评定。对于这种批评方法 1 和 2 尤其是脆弱的。它们不满足在其它科学领域普遍接收的统计推断标准，因为正式概率模型在决策处理中不起作用。方法 3 受到后一种批评较少，因为测量误差的代表是从背景得到的，尽管如此该方法不是最优的，因为测量误差不是从感兴趣的测量值（即各探针）直接得到的，而且这不会是背景值上的误差必然和探针值上误差幅度相同和/或模型相同的情况。

其它非正式方法也是可能的。例如在上面 2 中说明的方法可修改成估计多次探测过的基准基因的对数变换测量下的标准偏差。由于等式[log(a)-log(b)=log(a/b)]，从而测量误差的这些代表估计可以用来为感兴趣的对数变换探针的差异比率导出置信区间。尽管如此，该方法是差于最优的，因为误差基于代表值并且基于相对少的重复次数。

Chen 等(Chen, Dougherty 和 Bittner)在“Ratio-based decisions and the quantitative analysis of cDNA microarray images”，Journal of Biomedical Optics, 2, 364-374(1997)提出一种分析数学方法，其在零(null)假设下估计非重复的差异比率的分布。类似于本发明，该办法导出一种用于为不同条件下的探针强度差异得到置信区间和概率估计的方法。然而，在如何得到这些估计上它和本发明不同。不象本发明，Chen 等的方法不从重复的探针值得到测量误差估计。替代地，与不同条件下的探针强度比率相关的测量误差是通过比率的零假设散布的数学推导得到的。即，Chen 等推导出若不同条件的探针示出的测量值上的差异都不大于会期待为“机会”的差异下比率的散

布应是怎样的。基于这种推导，他们建立二个条件下的探针强度可靠统计比率的阈值。如其推导那样，该方法只可应用于在二个条件下估计差异。另外，它假定和探针强度相关的测量误差是正态分布的。如其推导那样，该方法不容纳其它测量误差模型（例如，对数正态）。它还假定所有测量值是“真实”探针强度的不偏移的和可靠的估计。即，它假定所有的探针强度都不是应从分析中排除的“局外”值。事实上，Chen 等说明的方法不能进行局外检测。

上面说明的各种方法试图解决和不同条件下差异应该多么大然后才能对于保证“真实的”差异结论把它们看成是足够可靠的有关问题。区分代表着信号的探针值和不代表信号的探针值是一个和阵列内的而不是不同条件下的探针值的鉴定有关的另一个论题。

已提出二种方法，Piétu 等(Piétu, Alibert, Guichand 和 Lamy)在他们的研究中观察到探针强度的直方图呈现双峰分布，见“Novel gene transcripts preferentially expressed in human muscles revealed by quantitative hybridization of high density cDNA array”，Genome Research, 6, 492-503(1996)。他们还观察到所出现的较小值的分布遵循高斯分布。以他们的论文中一种未描述的方式，他们把较小的值“拟合”成高斯曲线并使利用该高斯曲线均值之上的 1.96 的标准偏差的阈值区分信号（大于该阈值）和非信号（小于该阈值）。

Chen 等（上面提到的）描述下面的用于确定一个探针代表信号值或代表无信号值的方法。在阵列的数字化图象中，每个探针区域内的象素按等级排序。通过非参数统计检验（Mann-Whitney U-检验）八个最低象素值的强度和背景进行比较。若该统计检验的结果支持这八个象素值大于背景，停止该过程并把该探针认为是信号。若这八个象素值不在背景之上，部分或所有象素被看成是处于或低于背景。通过去掉所有八个象素并对下八个最低象素值重复该检验或者通过去掉这八个象素中的子集并用随后的最低值中的相同数量顶替重复相同的检验。以这种方式继续该检验，直至把所有的象素评估成处于或低于背

景或者达到象素数量的阈值。在这二种情况下，该探针归类为非信号。

大型格式（图 1 和图 4）是在数年前引入的并得到相当广泛的使用。典型地，探针按直径约为 1mm 的点设置在膜上。容易用机器人生成这些大的点，它们很适宜靶标的同位素标记，因为来自高能示踪分子（例如，32P）的电离辐射的扩散阻止使用小的间隔近的探针。最常见的检测是利用存储磷成象器进行的。

由微加工器件上合成的寡核苷酸组成的微阵列已使用了一段时间。随着近来商业上可购到微阵列制作及检测设备，在其上沉积着单链 cDNA 的微阵列的使用更为广泛。

借助微型和大型基因组阵列，通过检测各测定点处的同位素或荧光标记生成数字数据。该结果是一个或多个数字阵列，它的每个成员量化样品阵列中一个测定点处的杂交程度。杂交程度是对某具体探针互补的序列的表达水平的一种指示。从而，分析可用来确定互补序列的存在和用来量化产生这些互补序列的基因表达。

通过确定哪些具体测定显示杂交水平的有趣变化继续进行分析。典型地，杂交中的变化是以条件间的比率给定的。例如，数据可能是这种形式，即，在肿瘤细胞系中标记的测定 X（代表某特定基因的表达）比正常细胞系强三倍。相关的问题是“如何确定某具体比较的统计显著性？”。

由于测量中存在误差统计显著性规范是重要的，我们可把真实杂交定义成若不存在程序上的和测量上的误差下观测到的量。理想下，相同的探针－靶标配对应该总是赋给我们相同的杂交测量值。有效的杂交值是那些指示真实杂交的值。

事实上，杂交趋向于受到反应条件以及测量误差的严重影响。重复试验的荧光微阵列中的平均偏差系数通常为 25％左右。即，相同探针和靶标下的重复杂交例产生的值围绕均值（真实杂交的最佳估计）可以有相当大的变化。从而，任一数据点可能是或可能不是真实杂交的准确反映。

和现有技术不同，本发明直接从阵列重复试验（在阵列内或多个

阵列间）估计测量误差。本发明可以在以小数量重复（例如，三个）为特征的阵列杂交研究中提供有效的统计推论。在本发明中，由于小样本量造成的统计困难被新颖的处理所克服，其中该处理根据所有探针的所有重复的平均方差为每个探针获得测量误差估计。

根据一个优先方面，本发明假定作为各次试验的同一群体中的一部分的和在阵列处理期间得到类似处理的所有重复共享共有的和/或不变的方差。

根据另一个优先方面，可以为不同的探针类别独立地估计测量误差。可以根据下面说明的消卷积方法或者通过其它统计或试验方法确定这些类别。

本发明和所有现有技术的不同在于：

1. 可应用于任何数量的试验条件而不是限制在仅仅二个条件；

2. 从探针重复经验地估计测量误差；

3. 可检测出局外结果；

4. 可采纳不同的测量误差模型；以及

5. 可估评假定测量误差模型的充分程度。

存在本发明的第二个方面，其处理阵列内探针响应类别的鉴别。阵列内的单元测量可反映多类别的值。例如，某些值可代表信号而另一些值可代表非信号（例如，背景）。作为另一个例子，一些值可代表和疾病状态相关的基因家族，而其它值起源于疾病下未知的会改变的基因。本发明的新颖在于它使用对不同的基础分布所构成的任何混合进行消卷积的数学推导方法，该方法进而又用于把探针值分类为信号或非信号。

具体地，本发明的新颖之处在于它的处理阵列数据内的重叠分布的方法。尤其，本发明对阵列内的双分布或多分布建模。最好通过数学混合建模法做到这一点，该建模法可用于对各种分布以及精确方式下分布间的叠加区域进行消卷积。这和现有技术是不同的，现有技术不能用阵列数据对多于一种的分布建模而且从而也不能对分布之间的叠加区域建模。结果是，现有技术可能丢失具有属于有效信号分布的

可接受概率的数据（例如，带有低信号水平的探针）。本发明分配任何探针属于阵列数据群体的各种分布中的一种分布的概率。

从下面参照附图对目前优选的示意实施例的详细说明可更完整地理解本发明的其它目的、特点和优点，附图中：

图 1 是一个模拟杂交阵列的频率分布图，其示出信号测定和非信号测定二者的混合。背景具有等于零的均值并在该值上变化。从而，该分布中存在正值和负值。这种类型的分布是在尼龙阵列中发现的典型分布。

图 2 由图 2A 和 2B 组成，表示从图 1 的数据集中建模的信号和非信号的离散分布。

图 3 示出来自图 2 的二种分布，带有重叠区域，重叠区域内建模处理产生数据点起源。

图 4 由图 4A 和 4B 组成，示出来自玻璃微阵列上的淋巴细胞系（每次测定为三次重复的均值）以及尼龙膜上的 Clonetech Atlas 阵列的表达值的频率分布。二种情况中都已减掉来自基质的背景。玻璃阵列示出位于可能和非特异杂交混淆的某区域内的值比例相对小。膜阵列示出背景区中的大峰值。膜阵列适宜于建模。玻璃阵列不适宜建模。

图 5 和图 6 是示出处理的优选实施例的流程图。图 5 应用于测量误差模型已知的情况，而图 6 应用于未知该模型的情况。

本发明是一种客观分析阵列数据的统计处理方法，它包括二个过程

a) 各种分布的消卷积。在观测到的数据阵列包括来自二种或更多分布的贡献的情况下，本发明把这些分布消卷积成离散概率密度函数。这允许区别杂交信号和非信号，和/或区别不同标记的分布；

b) 对测定分配置信。

我们的如何区分分布的处理将涉及由信号和非信号组成的数据集。把这些处理方法应用到包含二个或更多标记的贡献的数据集对于本领域普通技术人员是显而易见的。

杂交数据集提供信号元和非信号元二者（图 1）。辨别非信号是必

要的，从而我们可做出有意义的表达比较（信号：信号），同时避免虚假比较（其中包括着非信号）。

假定存在一种或更多的分布。第一个问题是为信号设置阈值。我们的处理方法使用从阵列的方差特性中得到的信息以定义非信号和信号之间的切断点。首先，我们假定阵列分布是二种分布的混合。存在一个较低强度范围的分布（非信号，包括背景和非特异杂交）和一个较高强度范围的分布（信号）（图 2）。

利用建模对这二种分布描述概率密度函数。现在我们建立一组说明每个分布的性质的描述符。为了建立这些描述符，我们做出另一个假定。该假定是，每个分布起源于一个可从四个参数即均值、方差、混合比例和类别（例如高斯、伽马）估计的专用概率密度函数(pdf)。一种普遍认可的从混合分布中推导均值、方差和混合比例的方法是最大似然估计（MLE）。也可使用其它方法。

定义：

最大似然方法：我们提问“在给定每个分布的四个参数（均值、方差、混合比例和分布类别(例如，高斯、伽马)）的值下（由软件或用户产生）如何才能得到实际数据？”。MLE 方法在给出初始值下估计得到实际数据的似然性，并接着在给出略微不同的值下评估该似然性。继续迭代直到到达最大的似然性或直到到达预定义的迭代界限。

概率密度函数：由数学公式定义的一条曲线（例如，高斯曲线）。可以根据该曲线下的面积得到各种范围下的值(例如，X≤100; X≥500）的概率。

MLE 过程生成用于信号分布和非信号分布的 pdf（图 3）。这些分布包括对一个分布或对另一个分布明确无歧义的各区域。它们还包括重叠的区域，正是在重叠区内运行我们的处理以确定各数据点的来源。

利用概率密度函数向各杂交值分配它们的来源的分布。对于任何杂交值，我们可以确定得到一个大于非信号分布值的值或小于信号分布值的值的概率。以这种方式，我们得到二个概率（一个来自非信号

分布的值和另一个来自信号分布的值)。这二个概率的比较告诉我们哪个分布更有可能是该数据值的来源地。

研究表 1 中报告的值，这些值是从附录 A 中讨论的模拟数据中取出的。有三件事要注意：

1. 较大的值不太可能来自非信号分布（见列 2）而是更可能来自信号分布（见列 3）。
2. 列 2 和列 3 中的概率表示这二个分布中的哪个更可能是某特定杂交值的来源。例如，值 40 或更大值来自非信号分布的概率是.2107。值 40 或更小值来自信号分布的概率是.0995。我们的方法确定为 40 的值更可能来自非信号分布。
3. 可从概率函数获得信号和非信号杂交的判定值。在我们的例子中，小于 49 的值归类为非信号，而大于 49 的值归类为信号。

表 1 不同杂交值来源的概率

值	来源于非信号分布的概率	来源于信号分布的概率	更有可能的来源分布
40	.2107	.0995	背景
45	.1740	.1258	背景
49	.1493	.1482	背景
50	.1436	.1540	信号
60	.0980	.2148	信号
70	.0669	.2788	信号

检验拟合优良度。本发明建立描述实际数据的多个模型。可以根据χ平方统计利用拟合优良度参数评估各模型。该检验可以是自动的，并且软件标记出其中的建模产生坏拟合的情况。

何时建模是恰当的

建模过程假定杂交数据点阵列可分析成多个分布，每个分布具有足够的成员以允许准确的建模。使用尼龙阵列时通常是这种情况，这种阵列含有大的非信号分量（图 4）。许多玻璃阵列在性质上是相当不

同的。背景趋于低得多，信噪比较高。从而，不可能或不必对每个干净的阵列建立非信号分布的模型。在带有单个标记的一个干净的玻璃阵列下，我们可假定单种（信号）分布而免掉建模，并利用简单信号准则辨别有用的测定（例如，具有信噪比>3：1的测定）。

分布建模概述

本发明使用建模过程以把数据矩阵消卷积成二个或更多的概率密度函数。然后把各杂交数据分配到来源的最有可能分布。本发明的优点在于该建模过程提供一种客观方法，用于把各杂交值分配到信号分布或非信号分布、分配到一种或另一种标记或者分配到任何其它的消卷积后的分布。该过程可包括拟合优良度检验，若建模的结果是可疑的该检验会警告我们。

分配置信

每个杂交测定是一个估计。即，若我们多次重复该测定，我们将得到绕均值变化的一组值。所有这些值估计一个真实杂交值。一些测定值是该真实值的良好估计，而另一些不是。差的估计复盖可能的真实值的一个大区间。好的估计覆盖一个窄区间。在定义置信限制下，本发明生成围绕各观测值的各区间。我们可以具有各真实值位于这些区间内的高置信（例如>95％）。我们还可以利用这些区间确定各测定值之间的差异的置信度。若区间重叠，对于这些差异我们具有低的置信。若区间不重叠，我们具有高置信。从而，本发明对差异杂交的每个情形提供置信评分（见下节）。

观点1：用户输入误差估计。用两种方法之一我们得到误差幅度。若我们处理单成员阵列（无重复），用户可输入存在多大误差（按比例或常数）的估计。例如，持家基因可告诉我们该阵列具有25％的测量误差。

观点2：利用标准偏差或偏差系数从重复中确定误差。还可直接地从重复中确定测量误差。重复过程的优点是和平均值相关的误差按

$1/\sqrt{n}$的因子减小，其中 n 是重复次数，我们可利用与该变化性有关的信息为整个阵列提供总有效性参数（式 1）。

$$\sigma_\chi=\sqrt{\frac{\sum(x-\bar{x})^2}{N-1}} \qquad (1)$$

其中 N 是重复次数。

对于具有比例性测量误差的测量（杂交阵列的特征），偏差系数是变异性的一种有用量测。和各个值相关的百分比测量误差（相对于其均值）按如下估计：

百分比 $CV_x=100\frac{\hat{\sigma}_x}{\bar{x}}$

观点 3：利用从重复中得到的方差估计确定高不可靠的测定。各重复上的变异性估计随测定不同而不同。若它们变化过大，应丢掉该测定。我们应该如何设定丢掉某测定的准则？

我们检查变异性的变异。由此，我们可以确定其变异性超过某值的重复。通过计算方差值的方差和设置一个客观方差准则（例如 3SD 单位）指示局外值来确定该值。

在带有寄生误差（例如，100±10，1000±10）的情况下，标准偏差是每个数据点处的变化的最佳估计量。误差的绝对值保持为常数。

在带有比例误差（例如，100±10，1000±100）的情况下，偏差系数是更有用的变异性量测。标准偏差按测量值的幅值改变。

原始计数杂交测定典型地会呈现比例误差，而对数变换测定会呈现寄生误差。在这种基础上选择适当的统计。

综述该过程，我们从整个阵列中得到重复试验的平均 SD 或 CV。然后在下个步骤中我们利用该平均值。

对于寄生误差模型，通过式 2 完成该平均处理：

$$\hat{\sigma}_g^2=\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}(y_{gij}-\bar{y}_{gi})^2/(m-1) \qquad (2)$$

其中下标 g 代表分组或条件（例如，药剂，对照）。出于示意目的本文对二个分组建模，尽管该讨论可推广到任何数量的分组，下标 i 代表排成阵列的探针（n 是排成阵列的探针的总数量），而下标 j 代表重复

（m是重复次数）。式2是本发明的一个关键特性，其描述一种可以从整个阵列的分组估计离散重复分组的偏差性质。在给出观测数据下，该方法估计群体方差的期望值。其它使用基于整个阵列的各重复集上的方差的信息的各种方法也是可能的（例如，最大似然方法）。后一种方法对$\hat{\sigma}_g^2$的不同值计算得到观测数据的似然性。产生最高似然性的$\hat{\sigma}_g^2$的估计被选为群体方差的估计。在这二种方法中，在选择接着要应用到每个重复集上的群体方差值中利用整个阵列的各重复上的方差导出本发明的新颖性。

观点 4：利用从整个阵列或一组基准测定上导出的置信界限估计各个测定值的变异性。百分比CV提供各重复值围绕它们的均值的变异性的一种量测。各次重复的均值是测定的真实值的最好估计。但是，平均值具有它所关联的测量误差。和平均值相关的标准偏差称为均值的标准误差并按如下计算：

$$\hat{\sigma}_{\bar{x}} = \frac{\hat{\sigma}_x}{\sqrt{N}}$$

其中N是重复次数。

当测量误差为比例型时，变异性的量测是均值的百分比CV，后者按如下计算：

$$\text{百分比 } CV_{\bar{x}} = 100\frac{\hat{\sigma}_{\bar{x}}}{\bar{x}}$$

本发明取重复测定，并计算来自各重复的测量误差。在假定大多数或所有测定值区间上CV或SD相等时，该处理方法良好工作。此外，可确定具有不可用的过高百分比CV或SD的测定并且若它们确实不可靠可在进一步分析中删除掉。

各阵列上差异表达的情况

大多数建模处理要求大量的数据点。在一些情况下，比较阵列上的杂交值不能提供大量的差异杂交的测定。而是大量的测定具有类似的比率（通常为1：1），仅仅存在少量的差异杂交（例如4：1）情况。

利用阵列上杂交比率，本发明采用不要求大量数据点的分布式建模形式。

当存在重复时，生成用于杂交比率的置信界限。若我们已经评估与分子和分母比相关的百分比误差，根据下述公式估计和该比相关的百分比误差是一件简单的事情：

百分比误差 A/B=1--

$$\text{百分比误差 A/B} = 100\sqrt{\left(\frac{\hat{\sigma}_{\bar{x}_A}}{\bar{X}_A}\right)^2 + \left(\frac{\hat{\sigma}_{\bar{x}_B}}{\bar{X}_B}\right)^2}$$

其中 $\hat{\sigma}_{\bar{x}_A}/\bar{X}_A$ 是阵列 A 的重复均值的比例误差。本发明利用该公式计算用于任何 A/B 比的置信界限。

当不存在重复时，估计杂交比率的置信界限。

本发明具有向单例测定分配置信界限的优点。可由用户输入该估计。在用户输入的变异性基础上指定界限。

处理举例

已知测量误差模型

在一个优选方面，本发明假定已通过应用所需的已知过程（例如，背景修正、归一化）使系统误差最小或建立了模型。在另一个优选方面下，可在已对系统误差建立模型并对离散数据点实行偏移消除系统误差下使用本发明。本处理过程也可用于未建模的含有系统误差的数据，但结果会不那么有效。

为了便于陈述，下面的讨论假定在多个阵列上重复探针。然而，本处理方法等同地应用于在阵列内存在重复的例案。

二种常见的误差模型是“寄生型”和“比例型”。与测量量无关带有固定方差的误差模型称为“寄生模型”。具有和测量量成比例的方差的误差模型称为“比例模型”。后一种模型违反许多统计检验假设的固定方差假定。在该情况下，（对任何常用基底的）对数变换把误差模型从比例型改变到寄生型。在本文中讨论的处理中，可向每个阵列单元

施以对数变换。取决于误差模型，可设想其它变换或不变换。

图 5 和 6 是示出该处理的优选实施例。其它动作顺序是可设想的。例如，涉及消卷积以及分类过程的框 5 至框 7 可插入在框 2 和框 3 之间。即，在该替代实施例中，消卷积先于重复测量的误差估计。

在图 5 示出知道测量误差模型时的处理的概述。下面的各段编号成与该图中功能框号码相对应。

1.根据误差模型变换数据

若需要，在框 1 中变换原始数据，以满足后续统计检验所要求的假定。

2.计算重复的均值以及标准偏差

量化每组探针重复（例如，通过读出重复 cDNA 的荧光强度）并对探针值取平均以生成各组的平均值。按任何其它相关的描述性统计学，对每个重复探针组计算方差的无偏估计。

3.进行模型检查

在本发明的一个关键方面中，每组重复的平均变异性是根据该阵列内所有重复组的变异性的。然后可在诊断检查中利用该统计。各种误差模型和诊断检查是可采用的。诊断检查包括图形的（例如，分位点－分位点曲线以检查残留假定的分布）和形式的（例如，χ^2 检验；Kolmogorov-Smirnov 检验；比较均值、偏斜度和与误差模型的期待值有关的观测残留的峰态的各种检验）的统计检验，若满足误差模型的各假定，可建立用于去掉局外残留观测的阈值（例如，离开均值的±3 倍的标准偏差）。在去掉局外数后可重新检查模型的各种假定而且可对每个重复组重新计算平均变异性。接着可在框 8 中使用该变异性量测。

4.满足模型的各假定？

在框 4，判定残数的分布对于继续进行数据分析是否是充分的。若是，转到框 5。若不是，转到框 9。

5.需要消卷积？

在框 5，判定是否可能需要对值的混合分布的消卷积。若需要，转到框 6。若不需要，转到框 8。

6.对混合分布消卷积

在本发明的一个关键方面里，该处理的输入数据是从各个单观测或（最好是）从各重复上得到的单元强度。在一个优选方面下，E－M算法以及任何使该应用更灵活的修改（例如，允许对非常规的分布建模；允许使用先验信息，例如负值为非信号）提供一种对基础分布建模的常用算法。也可能采用其它对混合消卷积的方法。

7.应用分类规则

在给出框 6 中得到的各分布参数下，感兴趣的是把各次观测分类到一类或另一类（例如，信号和非信号）。可以用题目为“利用概率密度函数把杂交值分配到它们的来源分布上”一节中所说明的过程对各观测分类。

8.统计检验

一旦确定测量误差，进行标准统计检验并提供置信区间。这些检验应包括相关的和独立的 t 检验、相关的和独立的方差分析（ANOVA）以及其它标准检验。应对不同条件下的重复均值进行这些比较。一旦完成这些检验，处理结束。这被当成正常结束。

9.生成警告

若不满足误差模型假定，生成警告并结束处理。这被当成不正常结束。然后有三种可能的解决办法。可通过 Box-Cox 或其它过程人工变换原始数据。可以重新开始处理，从而能评估新模型的各假定。备择地，可应用图 6 中示出的最优策略。最后，可通过非参数的试验方法如仿真或其它过程估计误差分布。

未知测量误差模型

当测量误差模型不知时，图 6 中所示的处理除了如何选择误差模型外和知道误差模型下的处理相同。在该案例下，根据计算机密集型优化过程选择误差模型。在从框 1 到框 3 的循环中数据经历大量连续变换。这些变换例如可基于本领域普通技术人员熟悉的 Box-Cox 或其它类型变换。根据误差模型假设选择最优变换。若该最优变换接近某

基于理论的变换（例如，对数变换），后者可能是优先的。该处理继续通过和已知误差模型下方式相同的剩余步骤。

附属的附录 A 是一篇技术论文，其进一步公开本发明的各个方面。

尽管出于说明目的示出本发明的一种优选实施例，本领域普通技术人员理解，在不背离本发明的范围和精神不许多增添、修改和替代是可能的。

附录 A

统计信息学

用于分析阵列数据的处理方法

引言 18

表达研究的类别 18

表达数据 20

一种新方法 20

分布的辨别 21

建模处理 22

何时建模是适当的 24

实际样品的建模 26

分布建模概述 27

可靠性和置信区间 27

处理：存在重复 28

处理：不存在重复 29

分析差异表达 30

处理：测量误差已知 30

处理：测量误差未知 30

图示选择 32

参考文献 33

引言

基于阵列的表达分析（ABEx）用于在药物发现和诊断中发现增加的应用。尽管 ABEx 工具是复杂的，但是我们认为技术和商业工具的快速发展将有助于增加这种技术的应用。

某些 ABEx 发展在于材料、工艺和仪器设备。在微构造、杂交过程、排列和检测方式，有很大优点。另一发展方向与从阵列-信息中抽取含义有关。与含义有关的问题有两层。

. 观测的有效性：如果验证我们的观测是真实的？

. 结果有效性：我们的观测在生物的生物结果方面是否有意义？例如，在某些条件下表现出相似活性的基因族中，我们能够验证一致的“表达模式”吗？能够将特定基因的变化表达联系于变换/后变换事件吗？

所有的 ABEx 用户将面对这些问题。结果有效性是本研究的目标。经验证的观测有效性是得出与结果有效性有关的结论的前提条件。

我们认为，完整的 ABEx 系统必须检测和指定高密度阵列的杂交数据的有效性。因此，我们开发了“统计信息学”工具。统计信息学（SI）是一套分析方法，提供对 ABEx 数据点的可靠估计。统计上的可靠数据最可能是有效的。

表达研究的类别

ABEx 研究有三大类。

单条件：　单个条件下的表达，不与对照物比较

诊断：　　单个条件下的表达，与标准对照物比较

比较：　　直接与各条件下的表达比较

单条件研究（如 Pietu 等，1996）报告道“我们在这一组织上发现了 X 高度表达的结果，某些结果在其他组织中没有表达。”这些报告正变得不同，因为在表达观测与组织条件之间很难建立一种偶然的联系。

诊断的和比较的研究对样本条件直接进行比较。在诊断情况下，与独立标准化的条件进行比较。在比较情况下，在一种环境下对多个

条件进行总结。

表达数据

所有类型的 ABE_x 研究必须产生可靠的表达值。注意本文中使用的术语可靠的是分析的，这和生物学上不同。可靠的表达值是具有规定的（并且最好是小的）误差方差的表达值。存在许多用来降低阵列数据中的误差方差的策略。

- 多次观测。已经长期采用重复（对每个探针采取多例）案例以使测定内的偏差效应最小。通过测定的多次重复并取平均值或去掉高变化的案例，我们得到比取单例下更高的可靠性。
- 利用相同探针上的多个标记比较不同条件。通过允许直接比较不同的与单个探针杂交的条件（例如癌对正常）多个荧光标记可使误差方差为最小(例如,DeRisi 等,1996;Shalon, Smith 和 Brown, 1996)。在这种情况下，用条件之间的比表达数据。
- 参照均值、中间值或一组基准基因。阵列中的每个成员可以以某全局参数为基础。理论上，这种处理通过去掉对绝对强度值的依赖使样本间的偏差为最小。
- 匹配对－不匹配对。阵列中的每个序列具有一个通常有一个碱基对差异的成对序列。用“良好匹配”对不匹配序列的比率或者用相减值（匹配－不匹配）表达数据。相减去掉非特异的杂交和背景(在二种条件下应相同),而归一化为有关的探针提供内部基准。

一旦产生一批数据，下个步骤是规定表达中的多大变化是有意义的。例如，可看到这样的叙述:“可检测表达中的 2: 1 变化”。对这种类型的叙述的准确证明是多样的。最常用的方法采用从基准基因导出的变异性估计。库中包含一组基准或“持家”基因，它们的杂交是已知的。该组中的方差用来为阵列中的其它成员建立偏差准则。

一种新方法

我们提出一种我们称为“统计信息学”（SI）的统计处理方法，用于分析 ABE_x 数据。SI 包括二个主要部分。

a) 分布的消卷积。若阵列数据包括来自二个或更多分布的贡献（例如，信号/非信号，多种荧光），我们把这些分布消卷积成不同的概率密度函数。这允许辨别信号和非信号，和/或辨别不同标记的贡献。

b) 表达值的可靠性。我们观测到的表达值中的一部分是良好估计(可靠的)。另一些受到误差的严重影响（不可靠的）。我们对任何表达值计算可靠性。

现有技术不包含 SI 的优点。这些优点包括:

- 接收利用任何偏差减小策略生成的数据;
- 是基于模型的，与利用由阵列建立的基准材料不同;
- 简单易使用，从而可分析一般阵列;
- 为计算各数据点的可靠性提供一种客观方法。

分布的辨别

许多 ABE_x数据阵列由多个分布构成。例如，一个杂交数据集具有信号单元和非信号单元二者（图 1，2）。必须分辨出非信号，从而可做出有意义的表达比较（信号: 信号），同时避免假比较（任何包含非信号的比较）。

图 1：模拟阵列的频率分布，示出信号测定和非信号测定二者的混合。背景平均值为零，并在该零值上下变化。从而，该分布中存在正值和负值。这种类型的分布在尼龙膜阵列中是典型的。

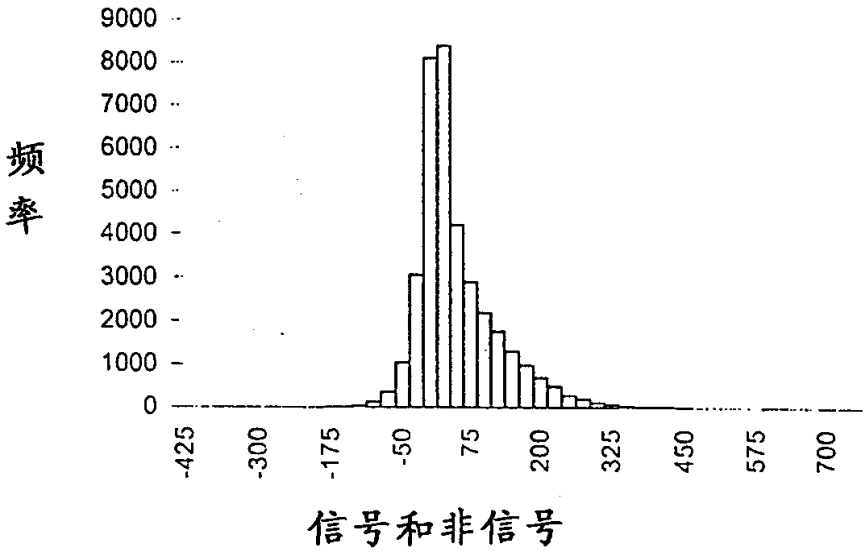
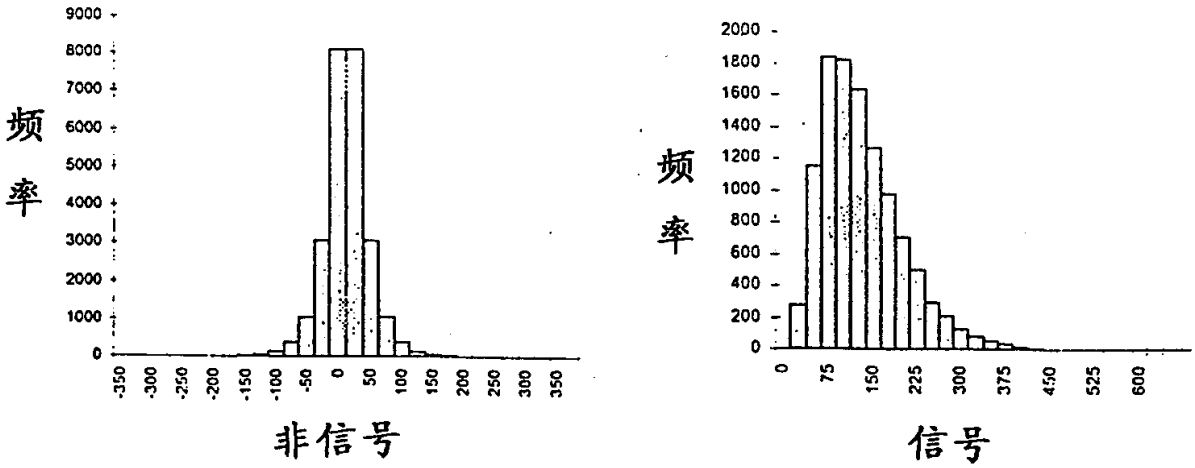


图 2：从图 1 的数据集中生成的信号分布和非信号分布



建模处理

步骤 1：利用建模为二种分布描述概率密度函数。我们建立一组规定每个分布的性质的描述符。为了建立这些描述符，我们假定每个分布源于一个专用的概率密度函数(pdf)，可以从四个参数即均值、方差、混合比例和类别（例如，高斯，伽马）估计 pdf。一种普遍认可

的从混合分布中推导均值、方差和混合比例的方法是最大似然估计（MLE）。也可使用其它方法。

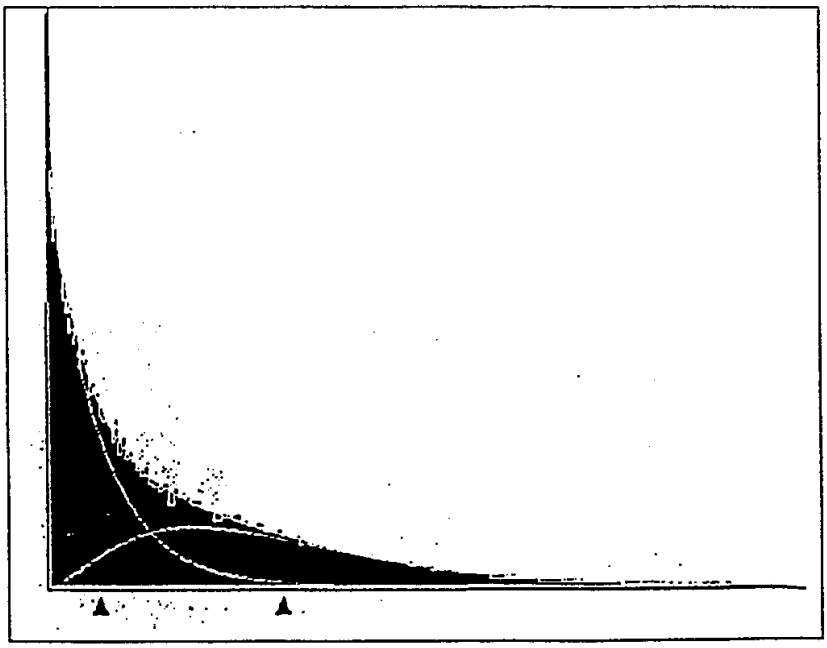
定义

最大似然方法：我们提问“在给定每个分布的四个参数（均值、方差、混合比例和分布类别(例如，高斯、伽马)）的值下（由软件或用户产生）如何才能得到实际数据？”。MLE 方法在给出初始值下估计得到实际数据的似然性，并接着在给出略微不同的值下评估该似然性。继续迭代直到到达最大的似然性或直到到达预定义的迭代界限。

概率密度函数：由数学公式定义的一条曲线。可以根据该曲线下的面积得到各种范围下的值（例如，x>100; x<500）的概率。

MLE 过程生成用于信号分布和非信号分布的 pdf（图 3）。这些分布包括对一个分布或对另一个分布明确无歧义的各区域。它们还包括重叠的区域，正是在重叠区内运行我们的处理以确定各数据点的来源。

图 3：信号分布和非信号分布的概率密度函数，示出重叠的区域。在该区域内，我们的处理对来源的分布确定杂交值。



步骤 2：利用概率密度函数把杂交值分配到它们的来源的分布上。

对于任何杂交值，我们可以确定得到一个大于等于非信号分布值或小于等于信号分布值的值的概率。以这种方式，我们得到二个概率（来自非信号分布的值的概率和来自信号分布的值的概率）。比较这二个概率告诉我们哪个分布更有可能是该数据值的来源地。

研究表 1 中报告的值，这些值是从附录 A 中讨论的模拟数据中取出的。有三件事要注意：

1. 较大的值不太可能来自非信号分布（见列 2）而是更可能来自信号分布（见列 3）。
2. 列 2 和列 3 中的概率表示这二个分布中的哪个更可能是某特定杂交值的来源。例如，值 40 或更大值来自非信号分布的概率是.2107。值 40 或更小值来自信号分布的概率是.0995。我们的方法确定为 40 的值更可能来自非信号分布。
3. 可从概率函数获得信号和非信号杂交的判定值。在我们的例子中，小于 49 的值归类为非信号，而大于 49 的值归类为信号。

表 1 不同杂交值的来源的概率

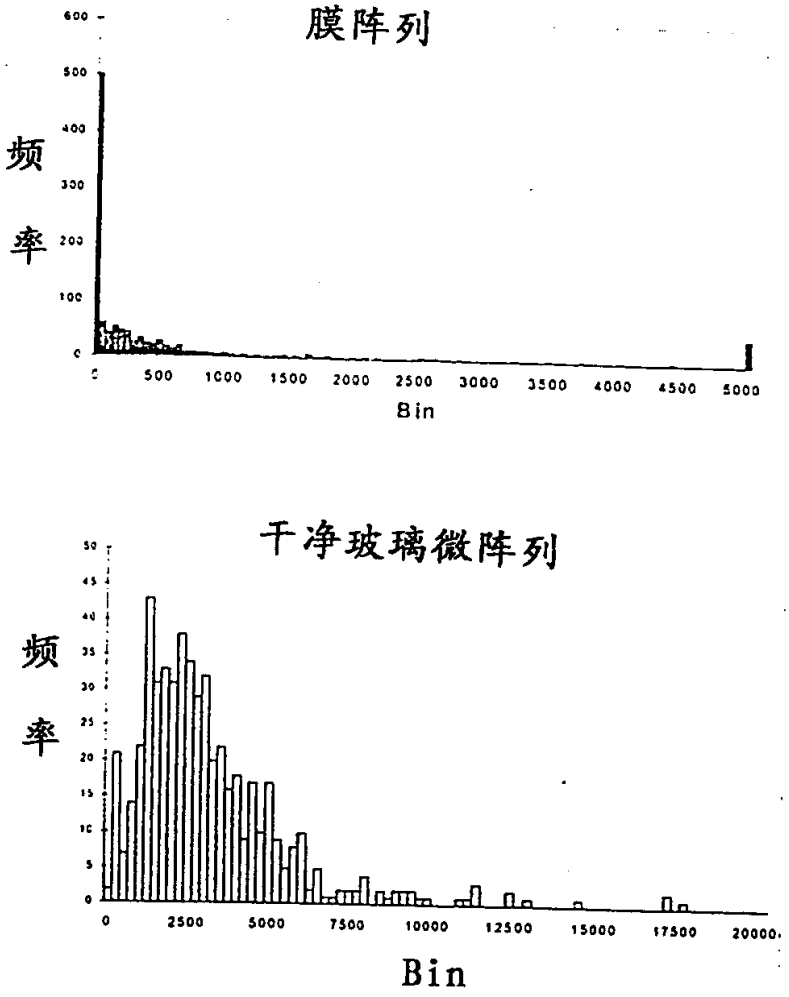
值	来源于非信号分布的概率	来源于信号分布的概率	更有可能的来源分布
40	.2107	.0995	非信号
45	.1740	.1258	非信号
49	.1493	.1482	非信号
50	.1436	.1540	信号
60	.0980	.2148	信号
70	.0669	.2788	信号
78	.0493	.3308	信号

步骤 3：检验拟合优良度。本发明建立描述实际数据的多个模型。可以根据 X 平方统计利用拟合优良度参数评估各模型。该检验可以是自动的，并且软件标记出其中的建模产生坏拟合的情况。

何时建模是恰当的

建模过程假定杂交数据点阵列可分析成多个分布，每个分布具有足够的成员以允许准确的建模。利用尼龙阵列下通常是这种情况，这种阵列含有大的非信号分量（图 4）。许多玻璃阵列在性质上是相当不同的。背景趋于低得比，信噪比较高。从而，不可能或不必对每个干净的阵列建立非信号分布的模型。在带有单个标牌的一个干净的玻璃阵列下，我们可假定单种（信号）分布而免掉建模，并利用简单信号准则辨别有用的测定（例如，具有信噪比>3: 1 的测定）。

图 4: 示出二种非信号比例的数据分布，上方是 32P 标记的 Clonetech Atlas 尼龙阵列，存在大的非信号分量。下方是 Cy3 标记的玻璃微阵列（肌肉组织），非信号分量很低。



实际样品的建模

对该情况总结如下：

我们已经证明建模对于理论分布良好有效。

我们已经示出膜阵列具有理论分布性质。

干净玻璃微阵列可能不具有允许建模的足够非信号点。

建模对玻璃有用吗？为了回答该问题，我们检查一些比起我们的优秀淋巴细胞库阵列不那么干净的微阵列。事实上，这样阵列具有许多膜性质（图 5、6）。从而，对于广大的各种阵列（包括荧光微阵列）建模是有用的。

图 5：集合着脊髓库的三次重复的荧光微阵列图象。淡红点代表非信号。亮红点属于重叠区域，其中建模把这些点确定为信号或非信号中的一种。其它颜色无歧义地是信号。

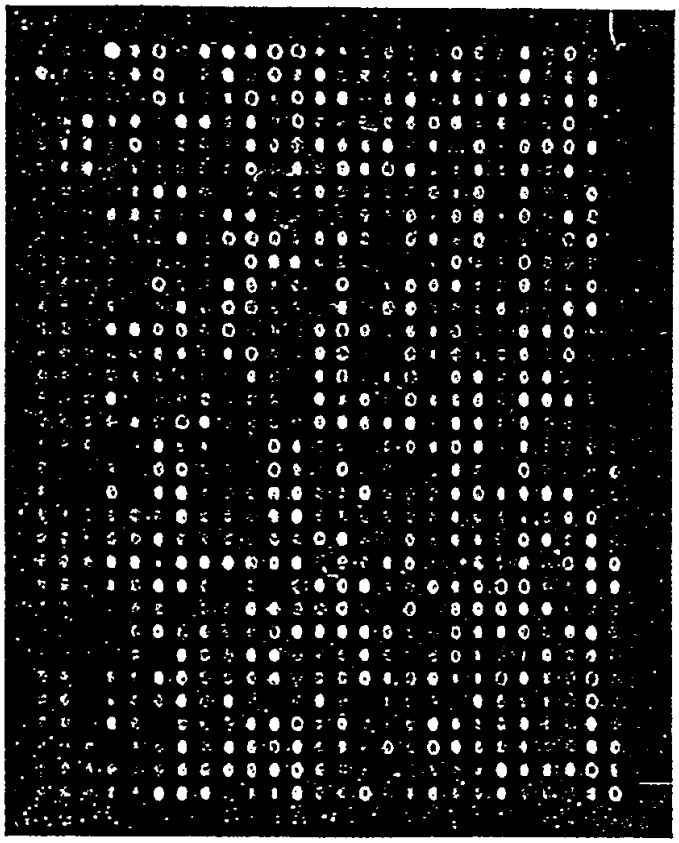
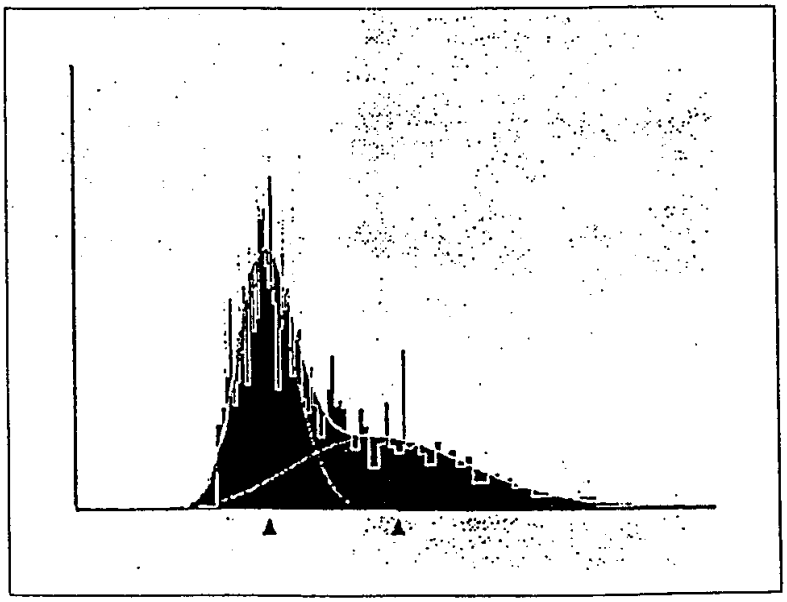


图 6：图 5 阵列的建模。红线示出信号和非信号的分布。蓝线示出强度 bin。绿线表示对实际数据的模型拟合。该模型没有明显地不

同于数据（χ^2检验）。



分布建模概述

我们利用建模过程把数据矩阵消卷积成二个或更多的概率密度函数。然后把各杂交数据分配到来源的最可能分布。建模的优点是:

- 不需要建立估计非信号的基准阵列;
- 把杂交值客观地分配到信号或非信号分布、分配到一种或另一种标记、或分配到任何共它消卷积后的分布。

该处理可包括拟合优良度检验，若建模的结果是可疑的该检验会警告我们。

可靠性和置信区间

每个杂交测定是一个估计。即，若我们多次重复该测定，我们将得到绕均值变化的一组值。所有这些值估计一个真实杂交值。一些测定值是该真实值的可靠估计，而另一些不是，规定任何给出的表达值是可靠的一个范围是有用的。

置信区间对真实值分类。在为杂交定义置信界限下，我们把观测值用作为估计，并生成这些估计周围的各个区间。在给出观测值 X 以

及该观测值的可靠性估计下，我们可给出一个区间，用 X 估计的真实杂交值应在该区间内。该区间是用某特定置信（例如，>95%）描述的。我们还可以利用该区间数据规定测定值之间或表达比率之间的差异的置信度。若区间重叠，对于这些差异我们具有低的置信。若区间不重叠，我们具有高置信。

处理：存在重复

若存在重复，可以直接确定测量误差。重复的附加优点是与平均值相关的误差以 $1/\sqrt{n}$ 的因子减小，其中 n 是重复次数。

步骤 1：利用从重复中导出的方差估计辨别高不可靠的测定。重复上的变异性估计随测定不同而不同。若它们变异太多，应弃掉该测定。我们如何为弃掉某测定设置准则？

我们检查变异性的变异性。由此，我们可识别它们的变异性超过某值的那些重复。该值是通过计算方差值的方差确定的，并设定指明局外点的客观偏差准则（例如，3SD 单位）。

步骤 2：利用均值的标准误差或偏差系数对各可接受的测定确定误差估计。通过各重复的平均值估计测定值。该处理可以利用均值的标准误差（$\hat{\sigma}_{\bar{x}}$，式 1）或均值的偏差系数（$CV_{\bar{x}}$，式 2）估计各次重复的测定误差。

式 1：某给定测定的多次重复的平均值的标准误差

$$\hat{\sigma}_{\bar{x}} = \hat{\sigma}_x / \sqrt{N}$$

其中 $\hat{\sigma}_x$ = 重复的标准偏差，

且 N = 重复的次数

式 2：某给定测定的多次重复的平均值的偏差系数

$$\text{百分比 } CV_{\bar{x}} = 100(\hat{\sigma}_{\bar{x}} / \bar{x})$$

在寄生误差情况下（例如，100±10，1000±10），标准偏差是每个数据点附近的变化的最佳估计量。

在比例误差情况下（例如，100±10，1000±100），偏差系数是变化度的最有用量测。标准偏差按比例地随测量值的幅值变化。

原始计数杂交测定典型地会呈现比例误差，而对数变换测定会呈现寄生误差。在这种基础上选择适当的统计。

综述该过程，我们在整个阵列中得到重复试验的平均 SD 或 CV。然后在下个步骤中我们利用该平均值。

步骤 3：计算置信区间。测定的误差估计允许我们建立每次测定周围的置信区间。我们希望具有的置信越高，归类真实值的区间越充分。在某具体置信级别上的可能值的范围称为“置信区间”。95％或 99％的置信是典型的置信级别。

置信级别，我们的区间包括真实值的可能性。

置信区间：该区间的实际值。

步骤 4：利用置信区间比较测定。任何二次测定的测量值几乎肯定彼此不同。这些差异可能简单地反映测量误差的作用或者反映真实值的实际差别。我们利用置信区间给出观测差异是真实的概率。

若二次测定的置信区间不重叠，我们确信在选定的级别上（例如，95％或 99％）测定的真实值彼此不同。若置信区间重叠，我们不确信这些真实值不同。

我们的处理方法的优点是：

- 利用标准统计过程从重复中计算误差。
- 直接从阵列数据计算置信区间。
- 利用客观准则描述置信区间。
- 借助误差概率给出表达比较。

处理：不存在重复

若不存在重复，不能得到统计上导出的可靠性估计。然而，我们仍需要误差估计以构建置信界限。这些误差估计是以各种方式建立的。我们可以在阵列中建立某些重复测定，并从中估计误差（例如，DeRisi 等，1996）。替代地，用户可以输入作为他的数据集的特征的某误差值。

一旦确定误差估计，可计算置信界限并且可借助概率确定表达值间的比较。

分析差异表达

大多数建模处理需要大量的数据点。通常，比较不同条件下的杂交值不提供大量的差异表达的测量。相反，趋于存在具有类似比率（通常 1：1）的大量测定，差异表达的案例（例如 4：1 相对少。这在准确建模上造成困难。

幸运的是，我们利用比率的一些性质在无须大量数据点下进行分布建模。

处理：测量误差已知

利用重复或者用户的输入估计测量误差为表达比率生成置信区间。若我们已经评估与某分数的分子和分母相关的百分比误差，估计和该比率相对的测量误差是一种简单的事情：

式 3：杂交比率的百分比误差（存在重复）

$$\text{百分比误差 A/B} = 100\sqrt{\left(\frac{\hat{\sigma}_{\bar{x}_A}}{\bar{x}_A}\right)^2 + \left(\frac{\hat{\sigma}_{\bar{x}_B}}{\bar{x}_B}\right)^2}$$

其中（$\hat{\sigma}_{\bar{x}_A} / \bar{x}_A$）=阵列 A 中各重复均值的比例误差

在式 3 中使用原始杂交值。当各阵列内不同测定上的测量误差为相同比例时，式 3 对所有的 A/B 比率产生相同的百分比误差。

当不存在重复但可得到测量误差估计（例如，来自基准值或已有研究；参见附录 A）时采用类似过程。

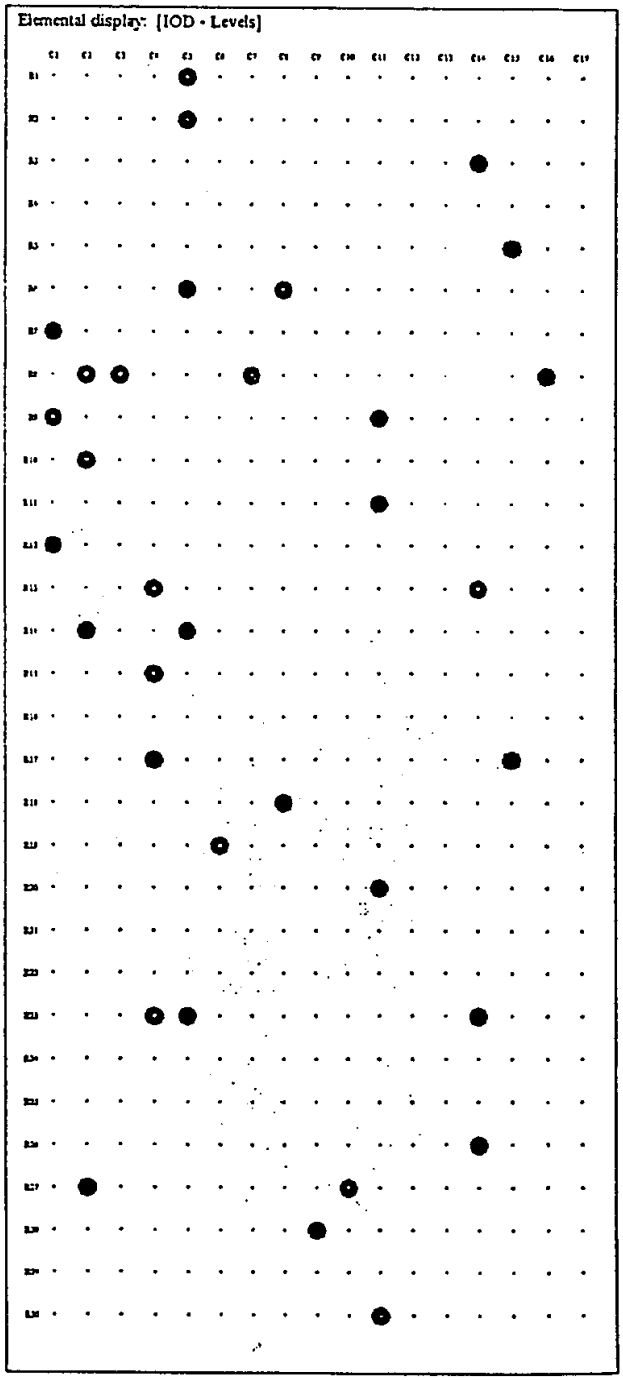
处理：测量误差未知

选择 2。利用从分布中导出的测量误差估计为表达比率生成置信区间。利用从非差异表达的值的变异性导出的估计为比率推导置信区间。

我们确定对数变换后的杂交比率的中间 50%（它们假定成近似地按高斯分布分布）的变异性。从该量测中得到非差异表达的这些比率的变异性的估计。接着如步骤 2 讨论那样使用该估计以对所有的比率

分配置信界限（图 7）。

图 7：基于比率评估的置信。位置 1，1 处的测定和所有其它测定比较。黄色表示 95％置信(p<.05)处的增加表达，而红色表示 95％处的置信(p<.01)。



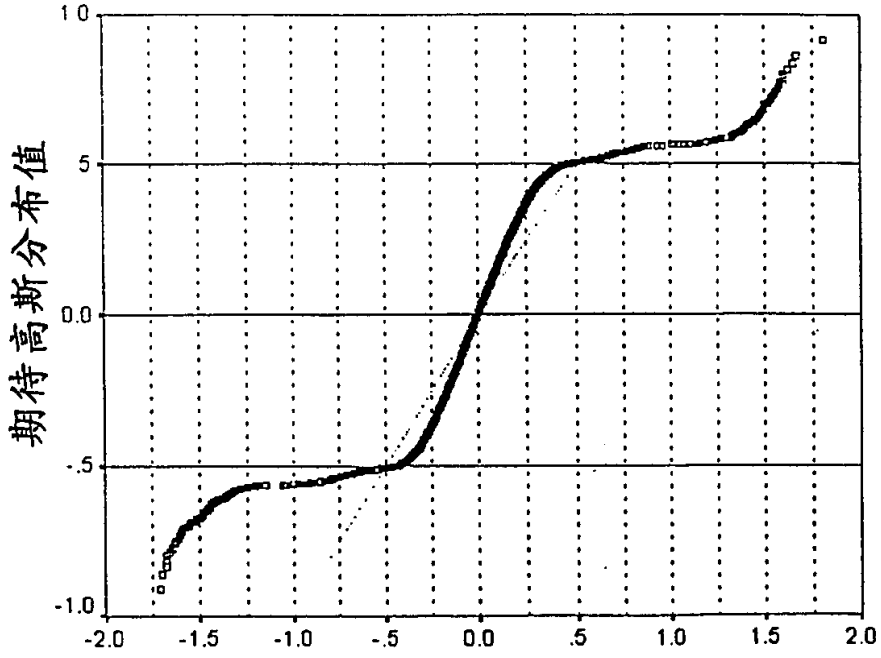
图示选择

利用 Q－Q 曲线对差异表达比率建模。Q－Q 曲线是图形统计学族中的一员。它以易于理解的形式把频率数据变换式 pdf。我们利用 Q－Q曲线对一个阵列中的各测定除以另一个阵列中的各测定的比率(实际上是原始数据比率取对数）建立模型。这些比率应表示三种部分重叠的分布：

- 在各阵列上无差异的值（比率 1：1）；
- 在各阵列上增大的值（比率>1）；
- 在各阵列上减小的值（比率<1）；

组成无差异值的分布的对数取值应是正态的。从而，我们可以使用该分布的中央部分对复盖着数据中所观测到的范围的完整部分建模。从而不落入该分布的观测值落入差异表达的分布中（图 8）。

图 8：Q－Q 曲线比较差异表达比率的分布（红线）和高斯分布（绿线）。其中接近描述期望值的直线的观测值落入在各阵列上无差异的值的分布上。其中偏离期望值的观测值落入差异表达的分布。



该处理方法的优点是它可在任何比率下进行，甚至我们不具有测量误差的直接估计时。

参考文献

DeRisi, J., Penland. L., Brown, P.O., Bittner, M.L., Mcltzer, P.S.,Ray, M., Chen, Y., Yan, A.S. and Trent, J.M. 使用 cDNA 微阵列分析人癌症的基因表达模式 Nuture Genetics 14:457-460(1996).

de Saizieu. A., Certa, U., Warrington, J., Gray, C., Keck, W. and Mous. J. 通过总 RNA 杂交寡核苷酸阵列进行细菌转录本成像 Nature Biotechnology 16:45-48(1998).

Nguyen. C., Rocha. D., Granjeaud, S., Baldit, M., Bernard, K., Naquet. P. and Jordan, B.R. 通过阵列 cDNA 克隆的定量杂交分析鼠科胸腺中的差异基因表达 Genomics 29:207-216(1995).

Pietu. G., Alibert, O., Guichard, V., Lamy, B., Bois, F., Leroy, E., Mariage-Smason, R., Houlgatte, R., Soulare, P. and Auffray, C. 通过高密度 cDNA 阵列的定量杂交揭示优先在人肌肉中表达的新基因转录本 Genome Research 6:492-503(1996).

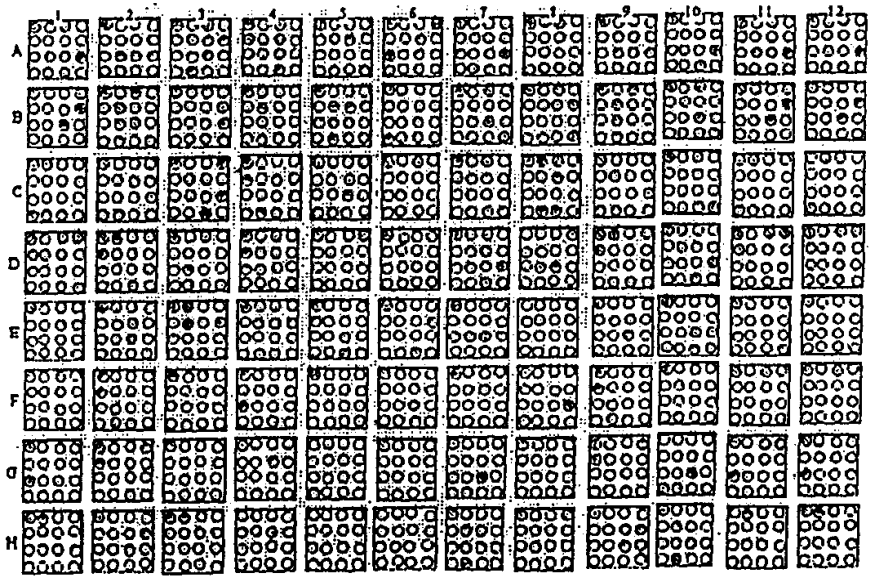
Schena. M., Shalon. D., Davis, R.W. and Brown. P.O. 用互补 DNA 微阵列定量检测基因表达模式 Science 270:467-470(1995).

Shalon. D., Smith, S.J. and brown, P.O. 采用二色荧光探针杂交分析复杂 DNA 样品的 DNA 微阵列系统 Genome Research 6:639-645(1996).

统计信息学
杂文阵列的定量分析

Peter Ramm
Imaging Research公司





观测有效性

- 观测反映“真实的”杂交值。

结果有效性

- 观测对所述生物具有生物学结果。

统计信息学

- 一组提供有关观测有效性的可靠性数据的处理方法。

表达研究的类别

单列：不直接和对照比较

诊断：单例和良好定义的对照比较

对比：不同条件下的表达的直接比较

比较表达

在所有类别的研究中，我们提问“表达中的哪些改变可以和对照条件区分和/或彼此区分”？

例子

- 在该条件下多少测定表现高水平的表达？
- 3：1（试验对对照）是可靠的差异吗？
- 4：1和3：1有区别吗？

改进表达值的有效性

- 有效观测反映“真实”表达水平。
- 误差损害我们做出有效观测的能力。

- 减少误差的方法是生物学的。
  - 多次点样
  - 荧光标记
  - 多种标记
  - 归一化
  - 匹配-不匹配型操作

- 确定误差的方法是统计学的。
  - 统计信息学

统计信息学

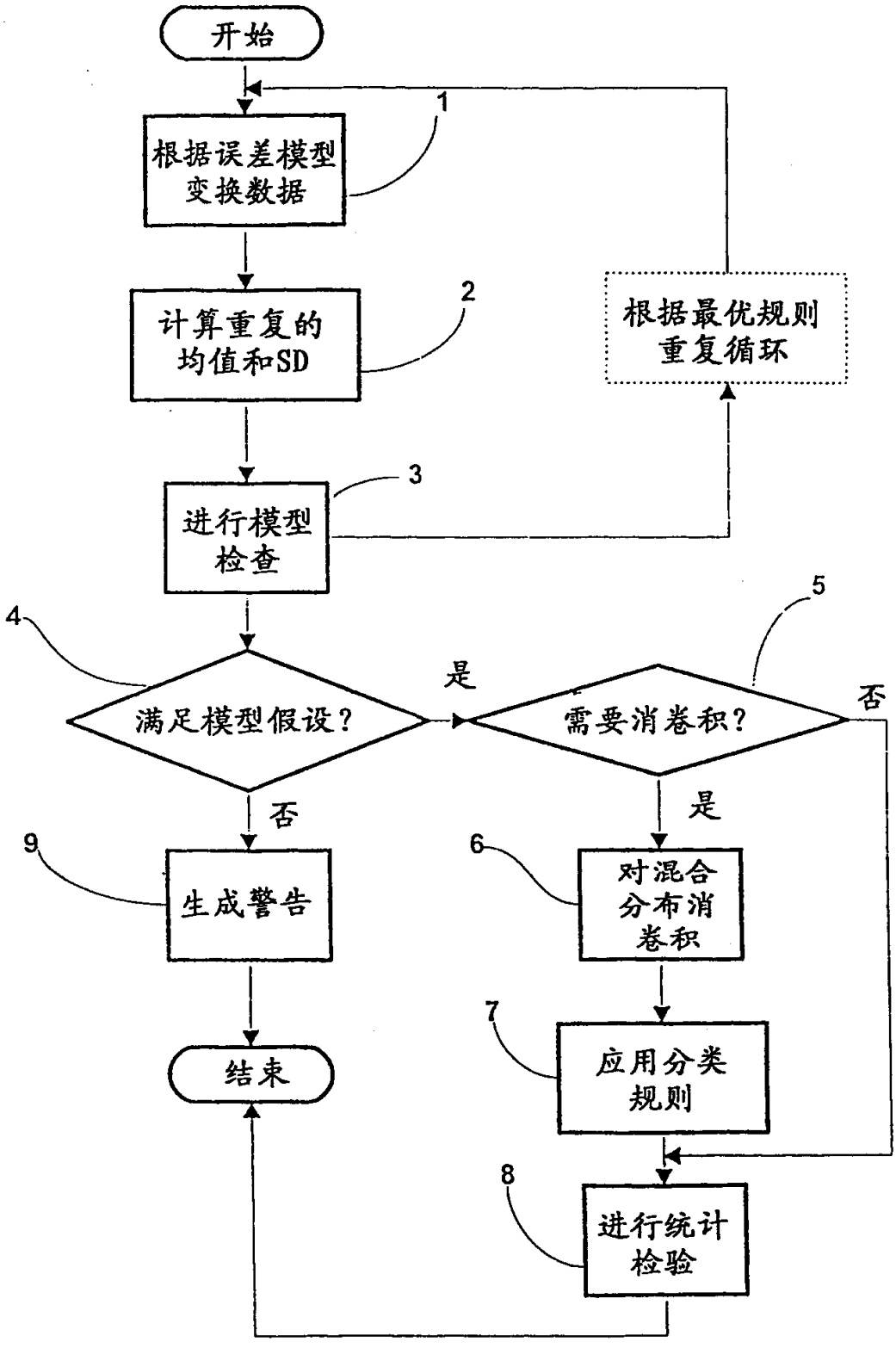
组成

- 对重叠的分布消卷积
- 估计表达值的可靠性

优点

- 通用应用-可分析大多数阵列
- 基于模型-无需阵列中的基准材料
- 客观-利用标准统计处理过程

图6


开始
1
根据误差模型变换数据
2
计算重复的均值和SD
根据最优规则重复循环
3
进行模型检查
4
满足模型假设?
是
5
需要消卷积?
否
否
是
9
生成警告
6
对混合分布消卷积
7
应用分类规则
结束
8
进行统计检验

任意的：在非显然有效下设定某准则.

分布的：利用从阵列中导出的方差准则.
例如，位于均值的 0.5SD 单位内的测定是非信号.

对二种分布建立模型

- 假设阵列分布实际上为二种分布－信号和非信号.
- 假设每种分布来自某专用概率密度函数（pdf）
- 可从四个参数，即均值、方差、混合比例和类别（例如，高斯、伽马等），估计 pdf.
- 设定分布类别，并利用最大似然估计自动地得到均值、方差和各个分布的比例.

定义

最大似然方法:

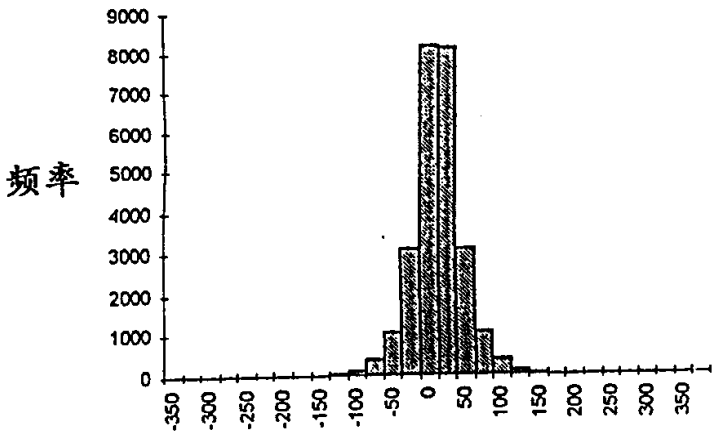
- 输入实际数据分布;
- 使用方差、混合比例、分布误差和类别的初始值;
- 生成对实际分布建模的概率密度函数。

概率密度函数:

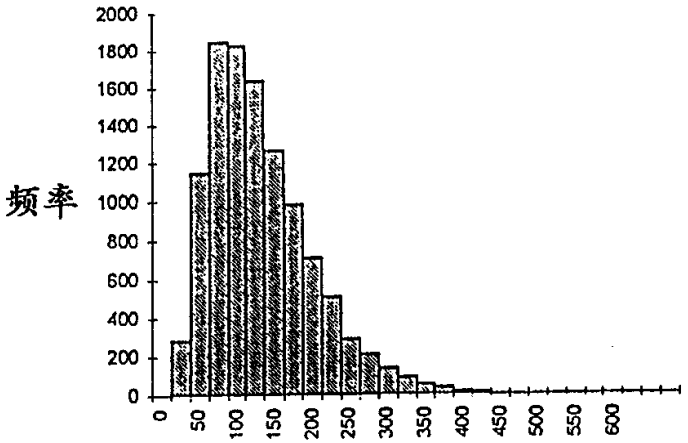
- 由数学公式定义的一条曲线（例如高斯曲线）.

建模的结果

MLE 算法生成二个带有重叠区的高散分布。



非信号



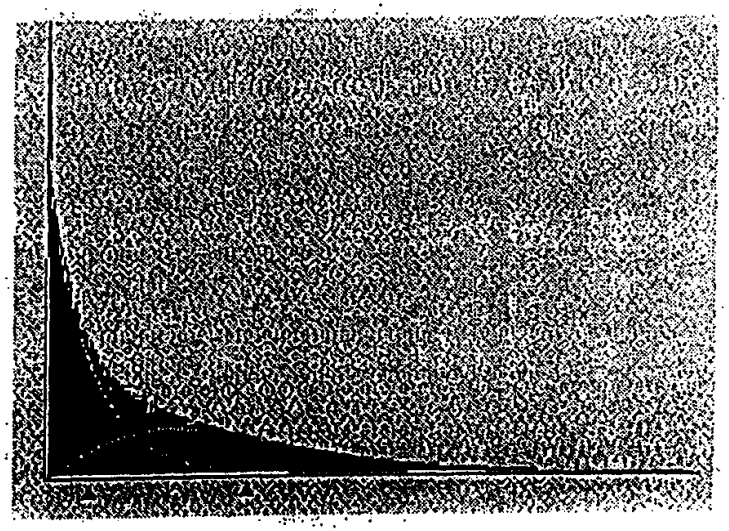
信号

向来源分布分配值

- 计算大于等于非信号分布中的值的和小于等于信号分布中的值的概率。

- 把该值分配到最有可能的来源

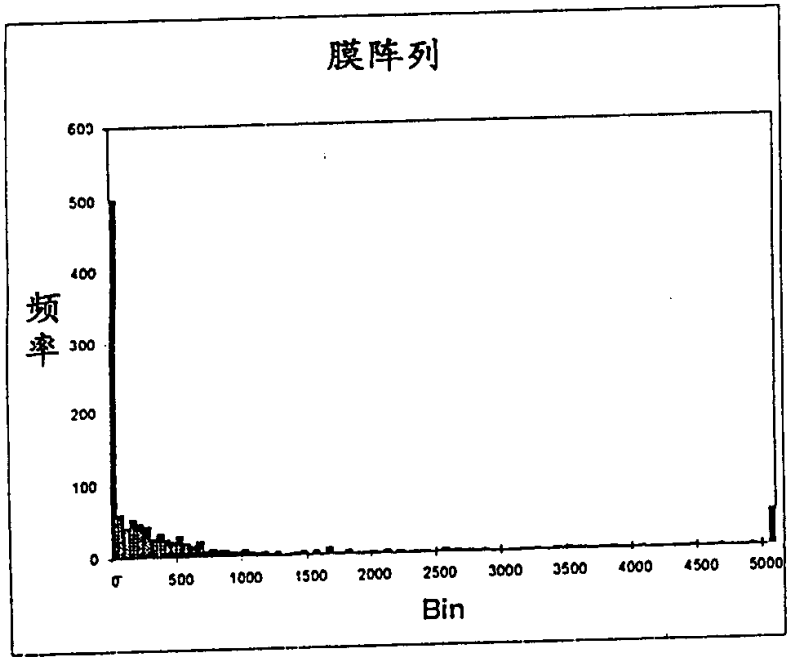
值	来源于非信号分布的概率	来源于信号分布的概率	最有可能的来源分布
40	.2107	.0995	非信号
45	.1740	.1258	非信号
49	.1493	.1482	非信号
50	.1436	.1540	信号
60	.0980	.2148	信号
70	.0669	.2788	信号

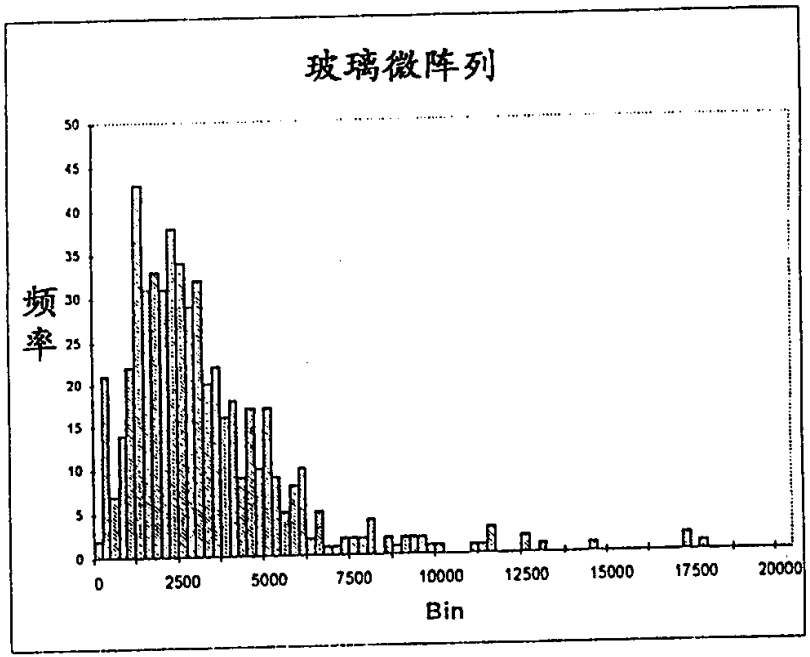


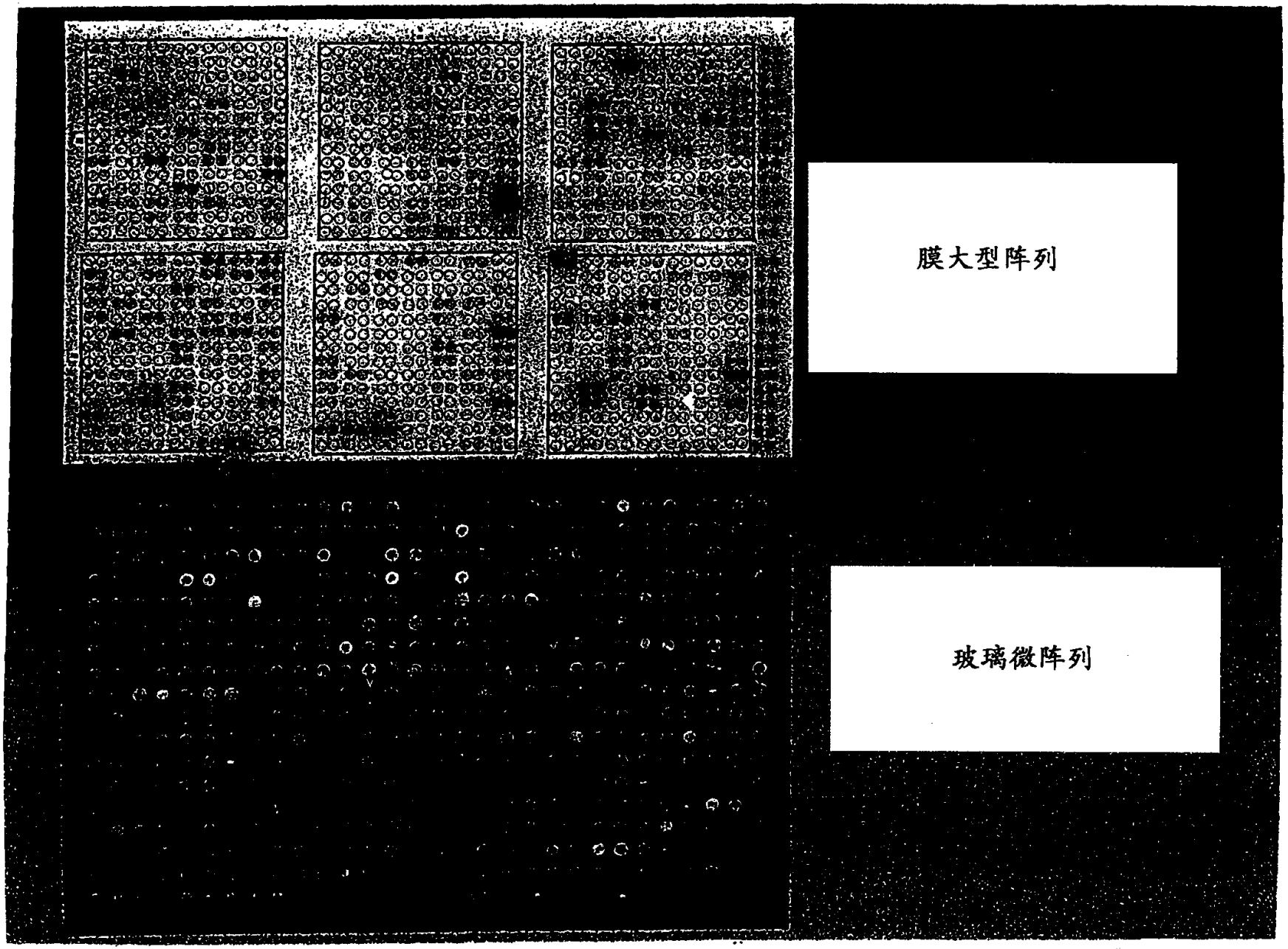
检查拟合优良度

- 我们的模型用于描述真实数据。
- 利用根据 χ^2 统计的一个参数评估优良度。
- 检查可以是自动的，而且软件可标记建模可疑的情况。

建模需要取决于样本

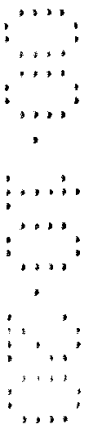

膜阵列
频率
600
500
400
300
200
100
0
0
500
1000
1500
2000
2500
3000
3500
4000
4500
5000
Bin


玻璃微阵列
频率
50
45
40
35
30
25
20
15
10
5
0
0
2500
5000
7500
10000
12500
15000
17500
20000
Bin


膜大型阵列
玻璃微阵列

分布建模小结

- 我们利用MLE过程定义分离的消卷积分布。
- 可使用拟合优良度过程以确定建模过程对给定阵列是适当的。



使用置信区间

任何测定值（例如100的表达）估计真实杂交。

由于误差，一些测定是可靠的估计，而一些不是。

较好估计提供真实杂交值周围的较窄区间。

置信区间用于规定真实值所落入的区间。例如，某为100的观测表达值在95%的置信下估计真实值位于50-150的区间内。

我们可利用置信区间在比较表达值时描述统计显著性。

比较非比率的值

步骤 1: 利用从重复中得到的方差估计确定高不可靠的测定。
问题 – 如何对舍弃设定准则?
我们计算方差值的偏差，并设定一个客观偏差准则以指示局外值。

步骤 2: 利用标准偏差或 CV 确定可接受测定的误差。

$\hat{\sigma}_{\bar{x}} = \hat{\sigma}_x / \sqrt{N}$,
其中 $\hat{\sigma}_x$ = 重复的标准偏差，
且 N = 重复次数

百分比 $CV_{\bar{x}} = 100(\hat{\sigma}_{\bar{x}}/\bar{x})$.

步骤 3: 计算置信区间。我们利用误差估计建立每个测定周围的置信区间。
- 置信区间: 区间值。
- 置信等级: 该区间包含真实值的概率。

步骤 4: 借助规定的置信比较各测定值。

该方法的优点

- 利用标准统计过程从重复中计算误差。
- 从阵列数据直接计算置信区间。
- 利用客观准则描述置信区间。
- 对不同于基线的表达案例赋予概率估计。

分析差异表达比率

方法 1:利用从重复中导出的误差或用户输入的误差的表达比率生成置信界限

百分比误差 $A/B = 100\sqrt{\left(\frac{\hat{\sigma}_{\bar{x}_A}}{\bar{x}_A}\right)^2 + \left(\frac{\hat{\sigma}_{\bar{x}_B}}{\bar{x}_B}\right)^2}$

其中 $(\hat{\sigma}_{\bar{x}_A}/\bar{x}_A)$ = 阵列A中每个重复均值的比例误差。

方法 2:利用来自比率分布的误差估计为表达比率生成置信界限。

- 对数变换比率的中间50%代表非差异表达的比率。可假定它们具有高斯分布。
- 从该分布得到变化性的估计，并（如上）利用它向所有比率分配置信区间。

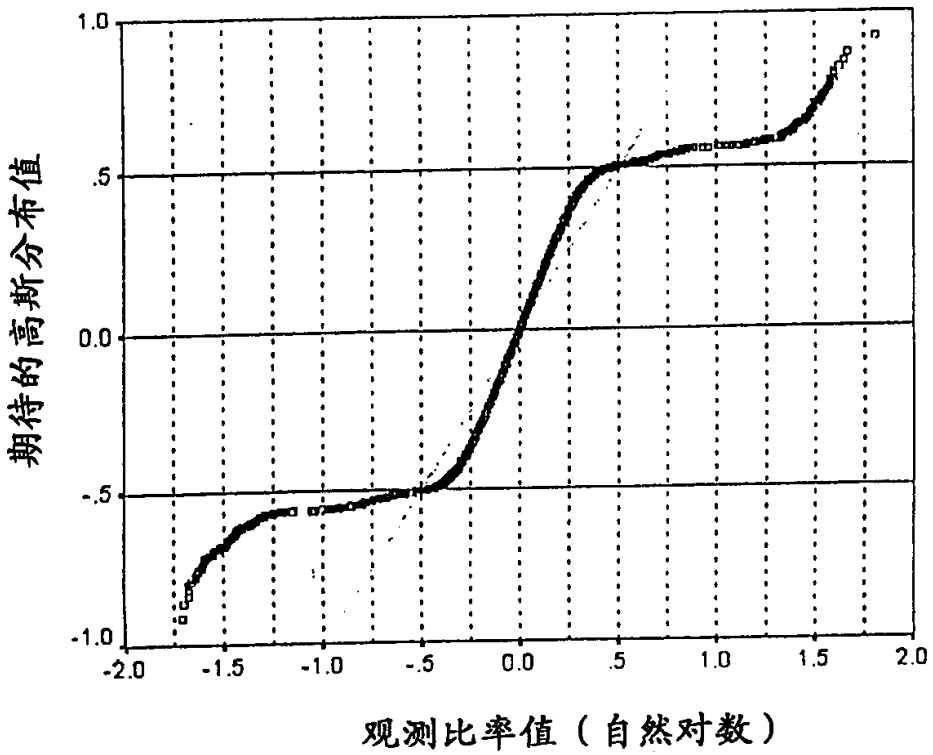
差异表达比率的变换

差异表达：图示表法

利用 Q-Q 曲线对差异表达建模。对二种条件下的比率建模。
这些比率代表三个部分重叠的分布。

- 不同条件下无差别的值（比率 1：1）
- 不同条件下增大的值（比率>1）
- 不同条件下减小的值（比率<1）

该过程可利用任何比率进行，甚至无误差估计。



典型玻璃阵列



该阵列集合脊椎库的三次重复。

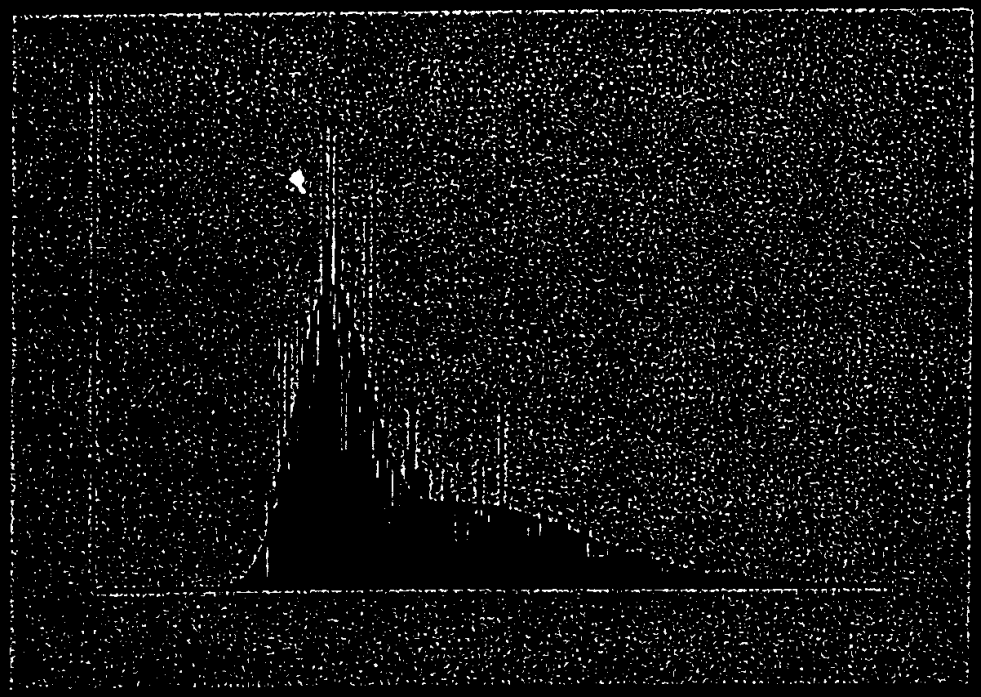
浅红色点位于非信号分布中。亮红点是信号和非信号的重叠区域。

我们如何鉴别信号点？

荧光微阵列的建模

- 蓝线表示实际数据
- 红线表示二个pdf
- 绿线表示对整个阵列的建模

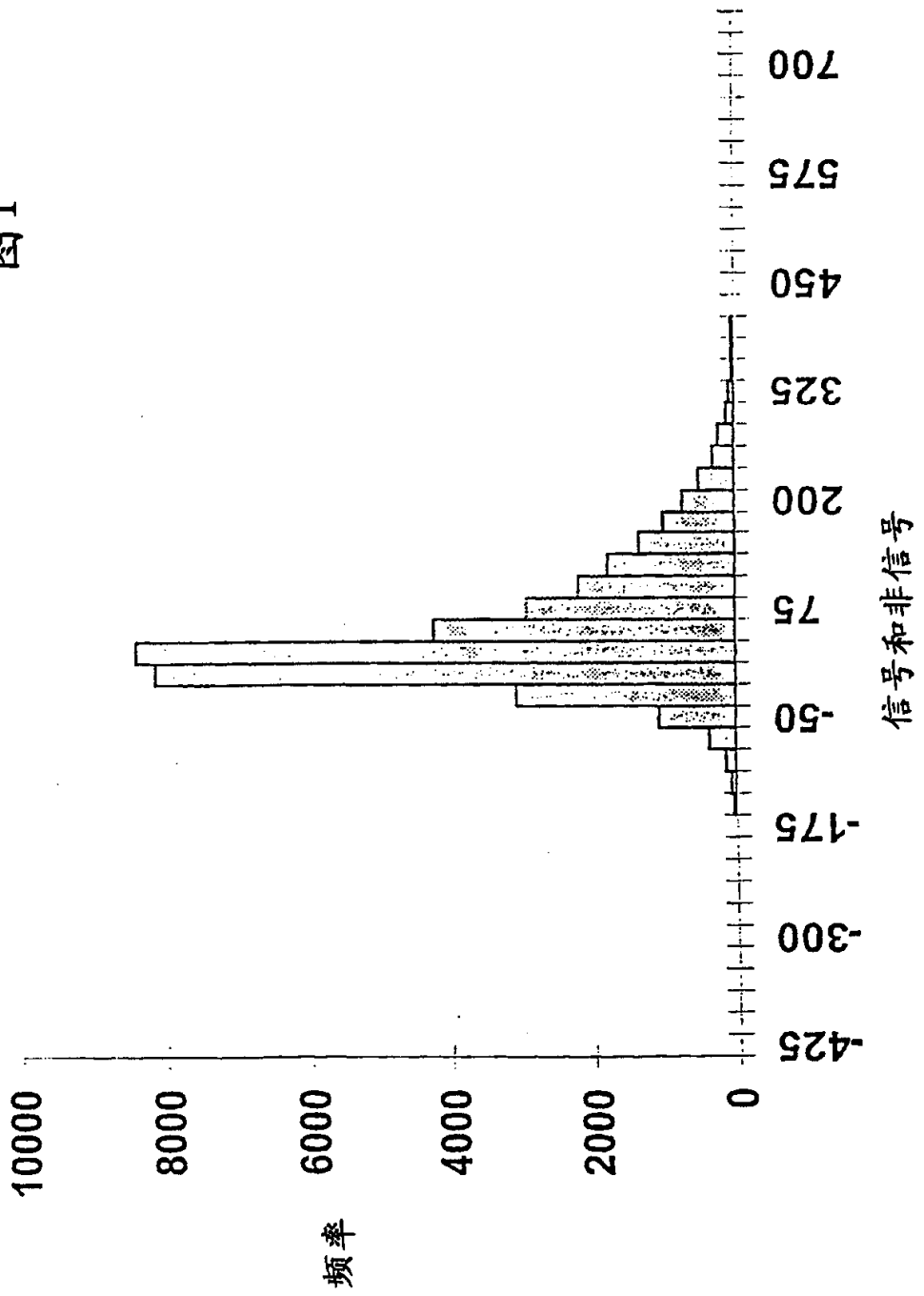
χ^2拟合优良度量测表明实际数据的频率和模型无差异



00.10.20

说 明 书 附 图

图1



00-10-20

图 2A

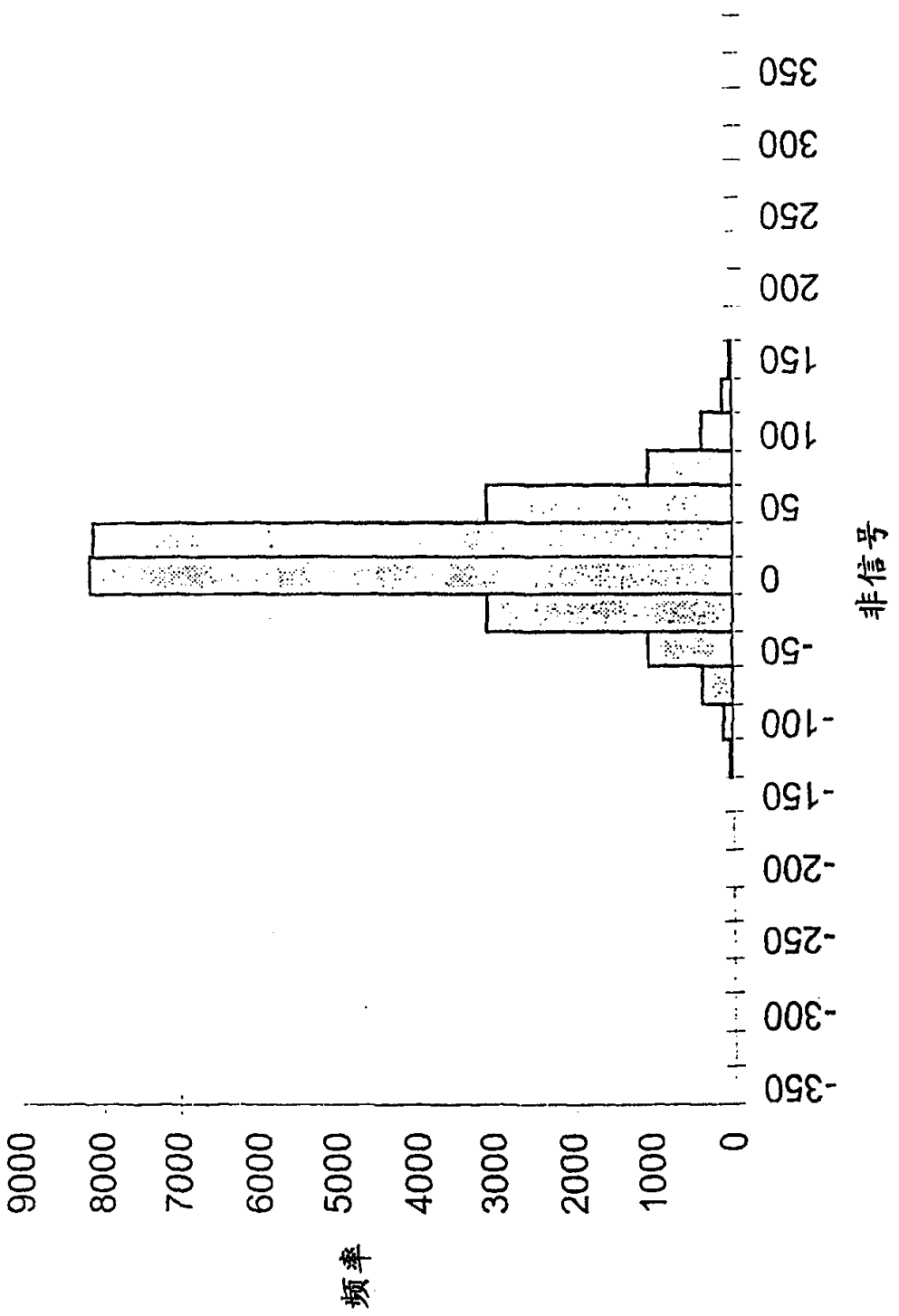
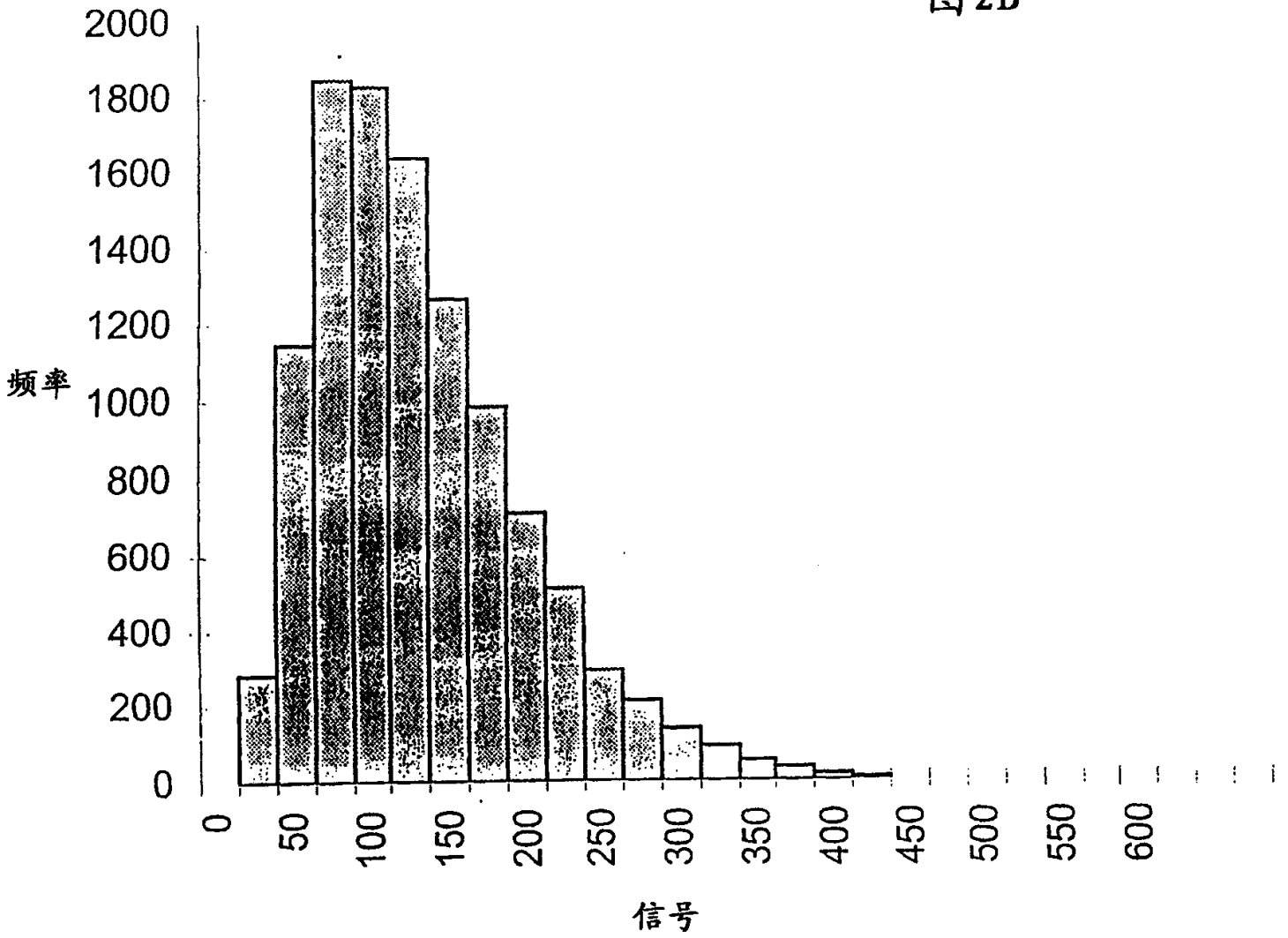

频率
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
非信号
350
300
250
200
150
100
50
0
-50
-100
-150
-200
-250
-300
-350

图2B


2000
1800
1600
1400
1200
1000
800
600
400
200
0
频率
0
50
100
150
200
250
300
350
400
450
500
550
600
信号

00.10.20

图3

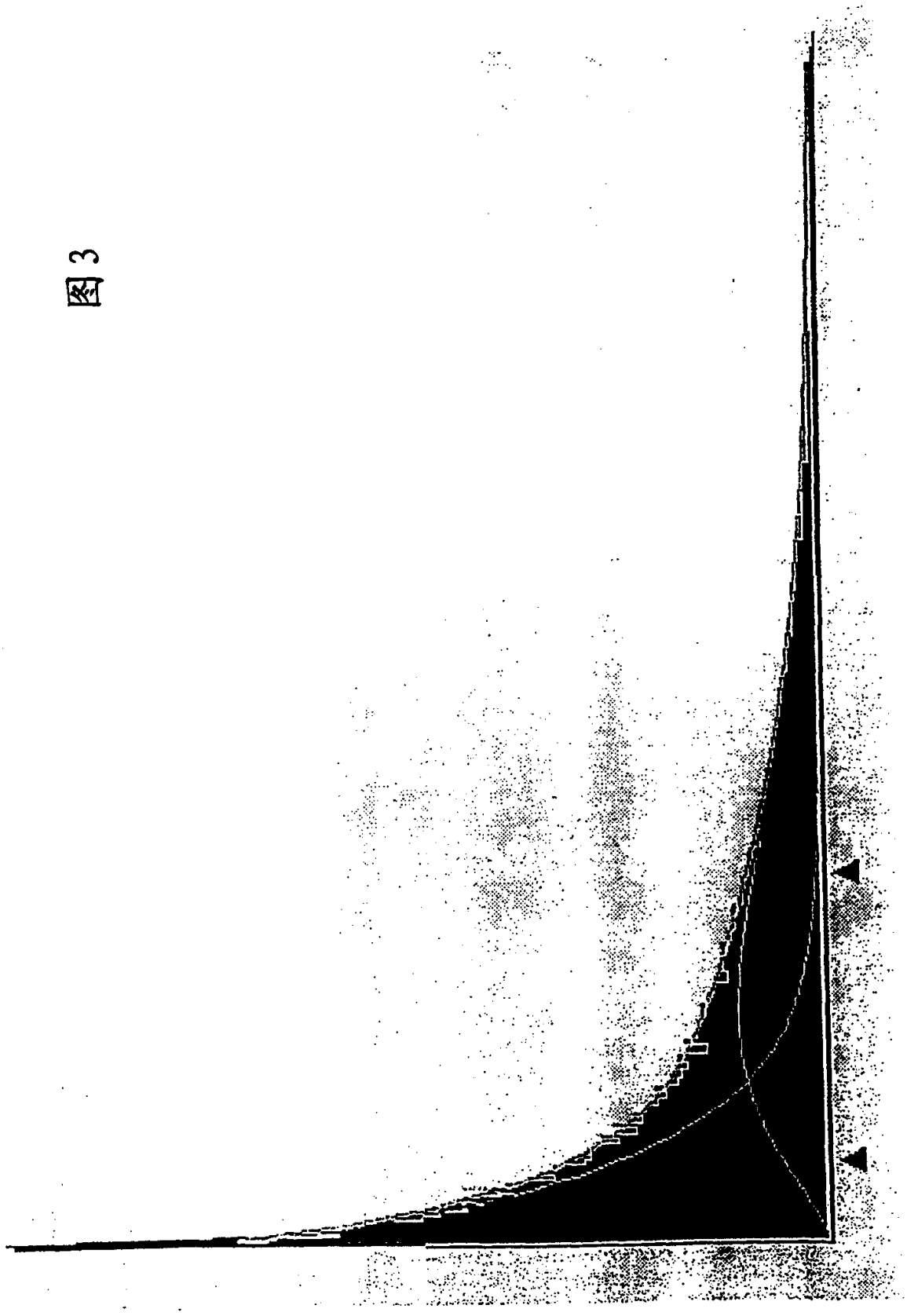


图4A

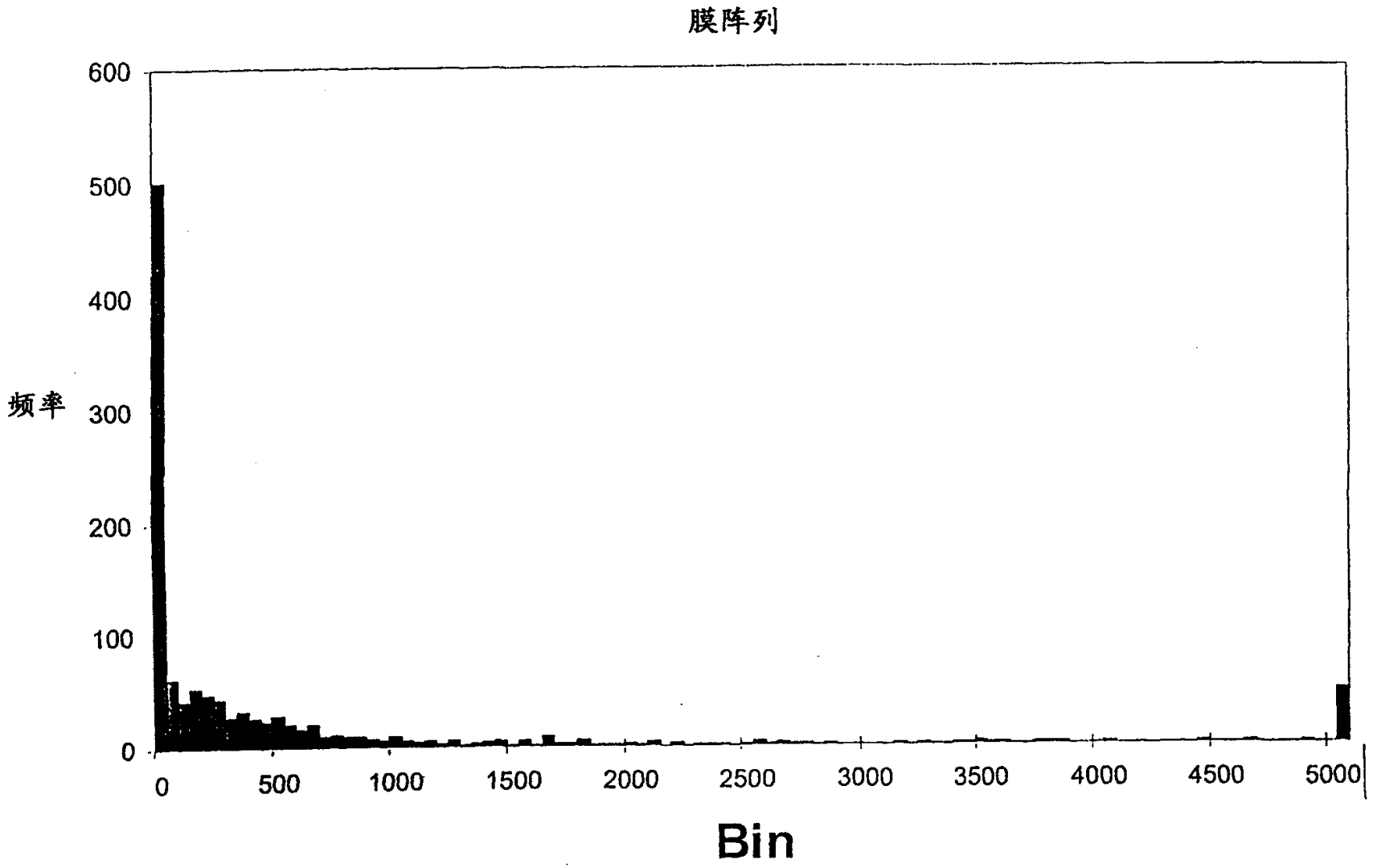

膜阵列
600
500
400
300
200
100
0
频率
0
500
1000
1500
2000
2500
3000
3500
4000
4500
5000
Bin

图4B

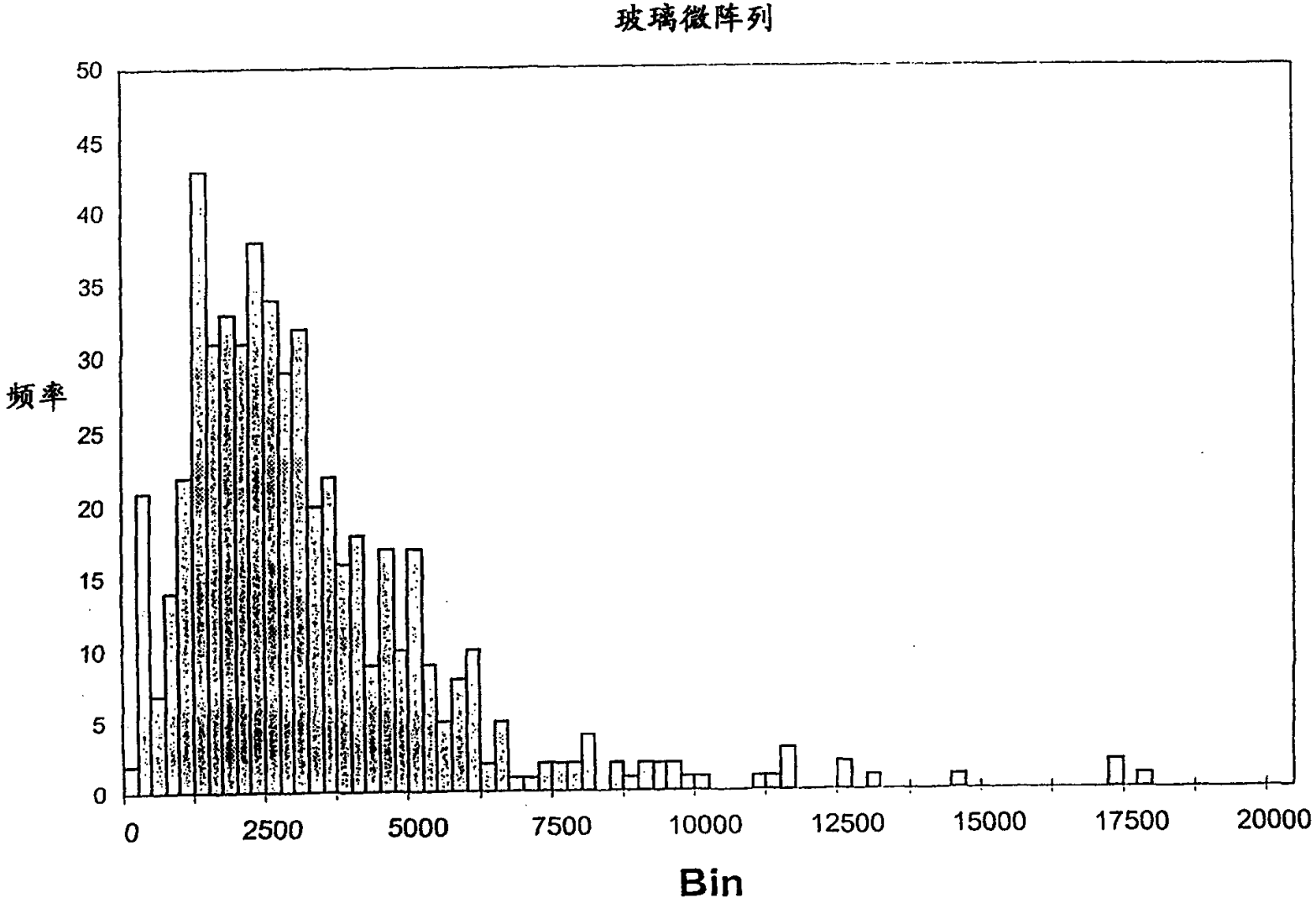

玻璃微阵列
频率
50
45
40
35
30
25
20
15
10
5
0
0
2500
5000
7500
10000
12500
15000
17500
20000
Bin

图5

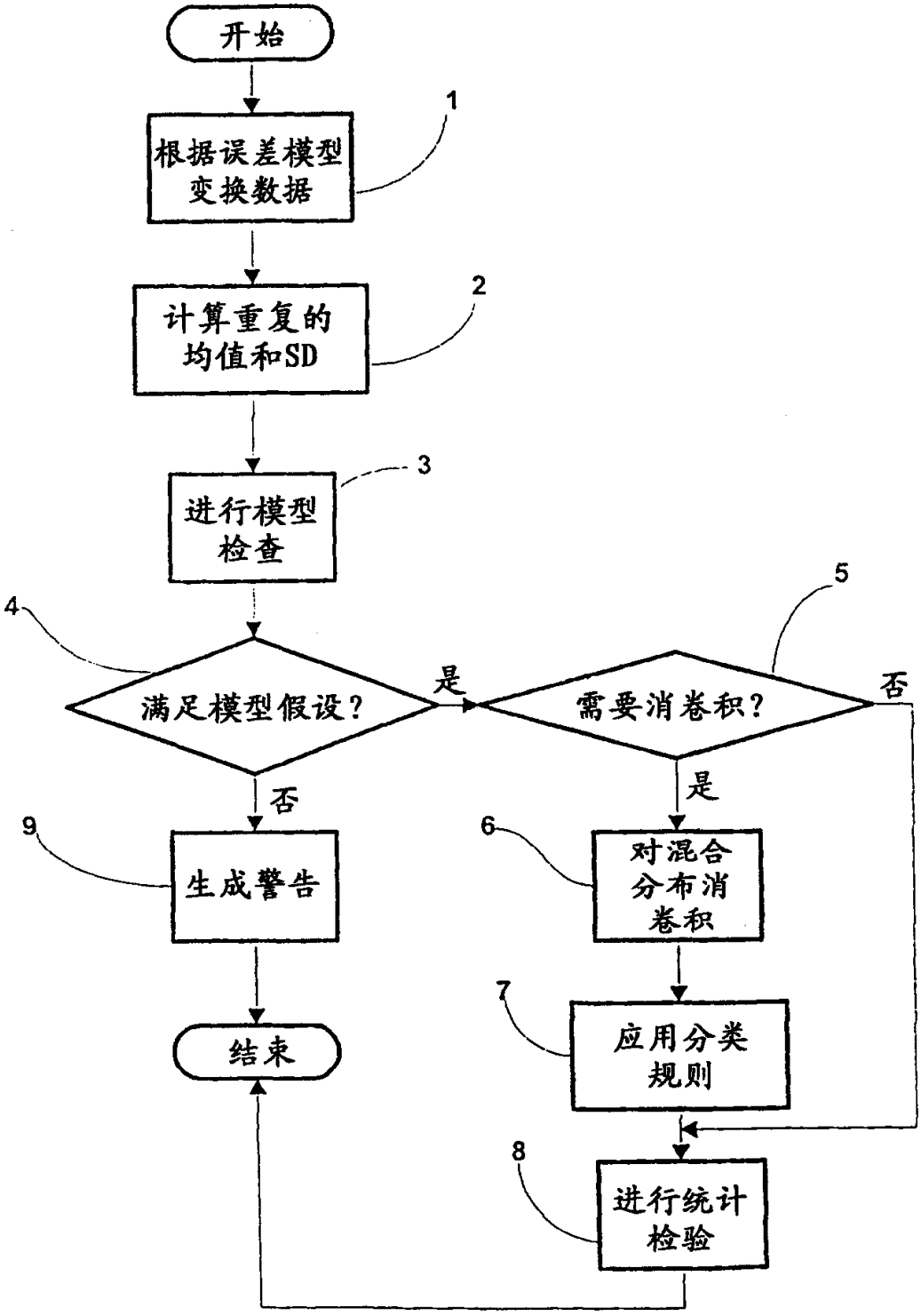
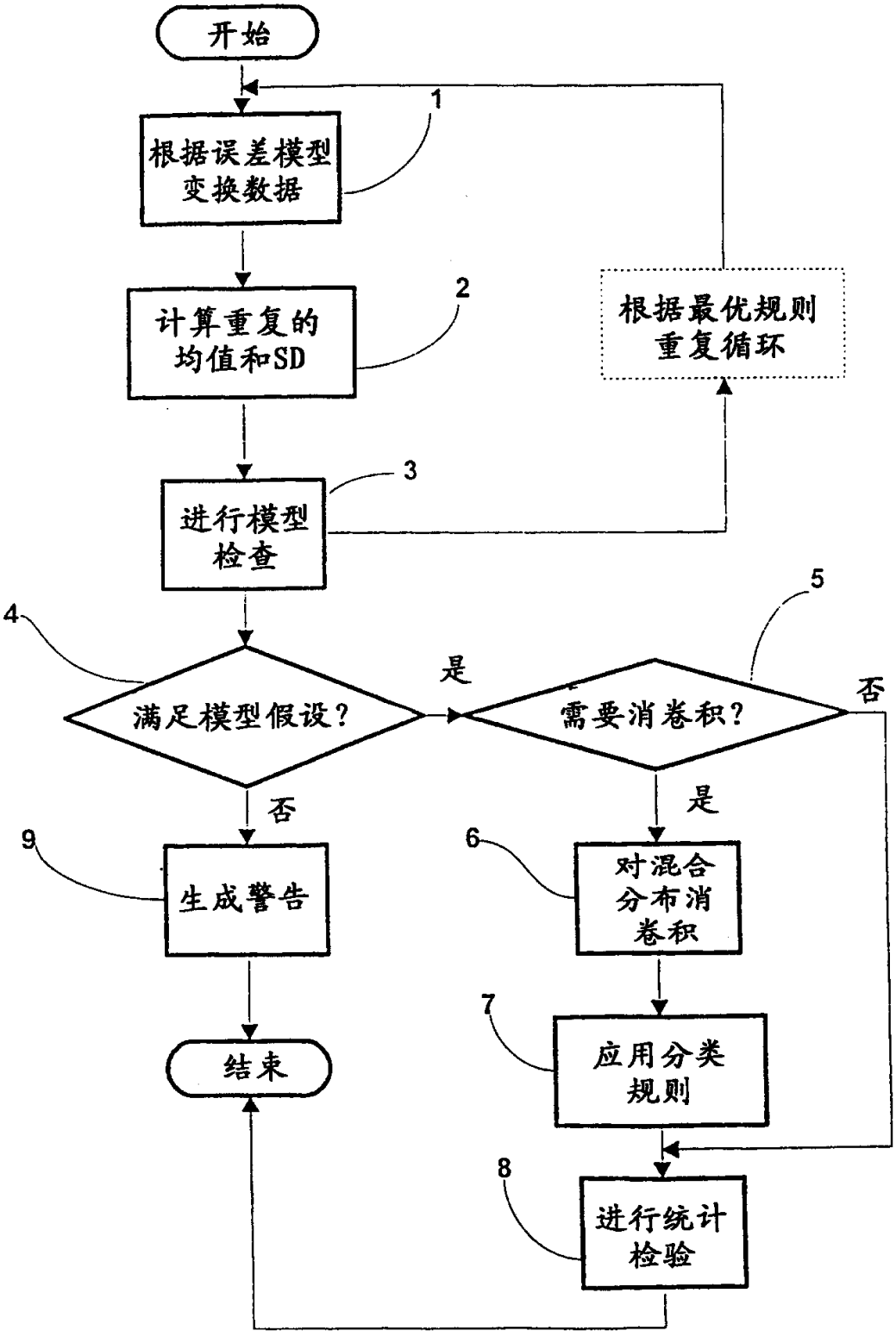

开始
1
根据误差模型变换数据
2
计算重复的均值和SD
3
进行模型检查
4
满足模型假设？
是
5
需要消卷积？
否
是
否
9
生成警告
6
对混合分布消卷积
7
应用分类规则
结束
8
进行统计检验

图6


开始
1
根据误差模型变换数据
2
计算重复的均值和SD
3
进行模型检查
根据最优规则重复循环
4
满足模型假设?
是
5
需要消卷积?
否
否
是
9
生成警告
6
对混合分布消卷积
7
应用分类规则
8
进行统计检验
结束