



(12) 发明专利

(10) 授权公告号 CN 101292245 B

(45) 授权公告日 2013. 04. 17

(21) 申请号 200680016219. 4

(22) 申请日 2006. 05. 12

(30) 优先权数据

60/681, 182 2005. 05. 13 US

11/432, 856 2006. 05. 11 US

(85) PCT申请进入国家阶段日

2007. 11. 12

(86) PCT申请的申请数据

PCT/US2006/018549 2006. 05. 12

(87) PCT申请的公布数据

W02006/124673 EN 2006. 11. 23

(73) 专利权人 生物辐射实验室股份有限公司

地址 美国加利福尼亚州

(72) 发明人 J·雷纳

(74) 专利代理机构 上海专利商标事务所有限公

司 31100

代理人 李玲

(51) Int. Cl.

G06K 9/00 (2006. 01)

(56) 对比文件

CN 1311891 A, 2001. 09. 05, 全文.

Keertan Dheda, Jim F. Huggett,

Stephen A. Bustin, Margaret A. Johnson,
Graham Rook, and Alimuddin Zumla.
Validation of housekeeping genes for
normalizing RNA expression in real-time
PCR. Bio Techniques Vol. 37. 2004, Vol. 37 pages
112-119, especially page 112.
Kenneth J. Livak and Thomas
D. Schmittgen. Analysis of Relative Gene
Expresion Data Using Real-Time Quantitative
PCR and the 2-delta-delta Ct Method. Methods
2001 Vol. 25. 2001, Vol. 25 402-408, especially
page 402.

审查员 苏珊珊

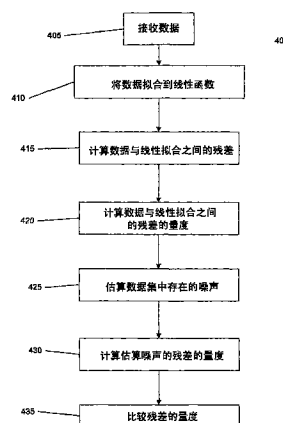
权利要求书 2 页 说明书 6 页 附图 7 页

(54) 发明名称

识别统计线性数据的设备和方法

(57) 摘要

提供了用于处理具有噪声的数据集以判定该数据集是否呈现统计线性品质的方法、装置和系统。基于数据的局部特性计算真实数据信号,而根据真实数据信号对数据中的噪声进行估算。然后,将估算噪声的量度与数据集的线性拟合特性进行比较。



1. 一种处理扩增过程的数据集以判定在所述扩增过程中是否还没有发生扩增的方法，所述方法包括：

接收具有信号分量和噪声分量的原始数据点集，所述原始数据点集是所述扩增过程中的数据点集；

将所述原始数据点集拟合到线性函数以获得经拟合的线性函数；

计算所述原始数据点集与所述经拟合的线性函数之间的残差；

计算所述原始数据点集与所述经拟合的线性函数之间的所述残差的第一量度；

估算所述原始数据点集中存在的所述噪声分量，通过：

i) 通过确定一组经平滑的数据点来计算经平滑的数据集，其中经平滑的数据点的值是基于所述经平滑的数据点局部的多个原始数据点的值；以及

ii) 计算所述经平滑的数据集与所述原始数据点集之间的残差；

计算所述经平滑的数据集与所述原始数据点集之间的所述残差的第二量度；以及

将所述第一量度与所述第二量度进行比较以判定所述原始数据点集是否呈现线性品质，从而识别在所述扩增过程中是否还没有发生扩增，其中所述方法由计算机执行，其中所述原始数据点集表示聚合酶链式反应(PCR)扩增曲线。

2. 如权利要求1所述的方法，其特征在于，拟合所述原始数据点集包括使用最小二乘方拟合。

3. 如权利要求1所述的方法，其特征在于，计算所述经平滑的数据集包括使用低通滤波器。

4. 如权利要求1所述的方法，其特征在于，经平滑的数据点的值是落在包围所述经平滑的数据点的窗口内的所述原始数据点的平均值。

5. 如权利要求1所述的方法，其特征在于，所述第一量度与所述第二量度各自为标准偏差。

6. 如权利要求1所述的方法，其特征在于，比较包括计算所述第一量度与所说第二量度的比值以便判定所述比值是小于还是大于预定值。

7. 如权利要求6所述的方法，其特征在于，所述预定值在1的量级上。

8. 一种处理扩增过程的数据集以判定在所述扩增过程中是否还没有发生扩增的设备，所述设备包括：

用于接收具有信号分量和噪声分量的原始数据点集的装置，所述原始数据点集是所述扩增过程中的数据点集；

用于将所述原始数据点集拟合到线性函数以获得经拟合的线性函数的装置；

用于计算所述原始数据点集与所述经拟合的线性函数之间的残差的装置；

用于计算所述原始数据点集与所述经拟合的线性函数之间的所述残差的第一量度的装置；

用于估算所述原始数据点集中存在的所述噪声分量的装置，通过：

i) 用于通过确定一组经平滑的数据点来计算经平滑的数据集的装置，其中经平滑的数据点的值是基于所述经平滑的数据点局部的多个原始数据点的值；以及

ii) 用于计算所述经平滑的数据集与所述原始数据点集之间的残差的装置；

用于计算所述经平滑的数据集与所述原始数据点集之间的所述残差的第二量度的装

置 ; 以及

用于将所述第一量度与所述第二量度进行比较以判定所述原始数据点集是否呈现线性品质, 从而识别在所述扩增过程中是否还没有发生扩增的装置, 其中所述原始数据点集表示聚合酶链式反应 (PCR) 扩增曲线。

9. 一种聚合酶链式反应 (PCR) 检测系统, 用于处理 PCR 扩增过程的数据集以判定在所述 PCR 扩增过程中是否还没有发生扩增, 所述系统包括 :

用于产生具有信号分量和噪声分量的原始数据点集的装置, 所述原始数据点集是所述扩增过程中的数据点集 ;

用于处理数据以判定所述数据是否呈现线性品质的装置, 通过 :

接收所述原始数据点集的装置 ;

将所述原始数据点集拟合到线性函数以获得经拟合的线性函数的装置 ;

计算所述原始数据点集与所述经拟合的线性函数之间的残差的装置 ;

计算所述原始数据点集与所述经拟合的线性函数之间的所述残差的第一量度的装置 ;

估算所述原始数据点集中存在的所述噪声分量的装置, 通过 :

i) 通过确定一组经平滑的数据点来计算经平滑的数据集的装置, 其中经平滑的数据点的值是基于所述经平滑的数据点局部的多个原始数据点的值 ; 以及

ii) 计算所述经平滑的数据集与所述原始数据点集之间的残差的装置 ;

计算所述经平滑的数据集与所述原始数据点集之间的所述残差的第二量度的装置 ; 以及

将所述第一量度与所述第二量度进行比较以判定所述原始数据点集是否呈现线性品质, 从而识别在所述扩增过程中是否还没有发生扩增的装置。

识别统计线性数据的设备和方法

[0001] 相关申请的交叉引用

[0002] 本申请要求 2005 年 5 月 13 日提交的题为“SYSTEMS AND METHODS FOR IDENTIFYING STATISTICALLY FLAT DATA IN A DATA SET(用于识别数据集中的统计平坦 (flat) 数据的系统和方法)”的美国临时申请 60/681,182,以及与其相关且同时提交的题为“DETERMINATION OF BASELINE END CYCLE IN REAL TIME PCR USING DERIVATIVE PEAK ANALYSIS(使用导数峰值分析确定实时 PCR 中的基线末端周期)” (律师案号:002558-072200US)”的序列号 No. 60/680,765 的美国临时申请的优先权,这些申请的公开通过引用全部结合于此。本申请还涉及以下共同受让、同时提交的题为“BASELINING AMPLIFICATION DATA(扩增数据的基线设定)”的美国专利申请:No. ____ (律师案号:002558-072210US)。这些申请的各自公开通过一般引用全部结合于此。

[0003] 技术领域

[0004] 本发明一般涉及数据处理系统和方法,尤其涉及用于识别诸如聚合酶链式反应 (PCR) 的扩增过程的数据集内统计线性数据的系统和方法。

[0005] 背景技术

[0006] 许多试验过程呈现数量的扩增。例如,在 PCR 中,数量可对应于 DNA 链中已被复制部分的数目,在扩增阶段或扩增区的期间,该数量显著增加。其它呈现扩增的试验过程包括细菌生长过程。通过试验设备经由数据信号来检测该数量,并分析其数据点以确定关于扩增的信息。作为数据分析的一部分,获知扩增是否可能发生是重要的;否则,工作将浪费在分析非扩增数据上。如果数据是统计线性的,则未发生扩增。

[0007] 理想情况下,来自扩增检测设备的数据可能是单调且连续的信号,因此很容易识别数据或其部分是否具有统计线性品质。然而,来自扩增设备的信号通常包含噪声,因而使信号品质难以识别。噪声本身出现在来自设备的信号的各个数据点中,并具有在例如 DNA 链的实际数目的真实信号上发生的随机波动。因而,数据需要进行处理以便能识别线性品质。

[0008] 一种用于处理数据以判定其是否为统计线性的典型现有方法是通过线性最小二乘法 (LSQ) 拟合。LSQ 拟合的相关值可用于判定是否充分拟合。通过标准约定,相关值 0 与不良拟合相关,因此数据不是线性,而值 1 表示良好线性拟合。问题在于,存在噪声时,对于看起来统计线性的数据,相关值可能接近 0 或 1。此外,相关值并不对应于可提供额外理解和效能的物理值。因而,相关值不是可以接受的标准,尤其对于噪声极大的数据而言。

[0009] 因此,期望提供用于处理具有噪声的数据集、以及用于识别该数据集是否为统计线性并且克服了上述及其它问题的系统和方法。

发明内容

[0010] 因此,本发明的实施例提供了针对处理数据以判定数据是否呈现统计线性品质的方法和系统。统计线性数据表示数据一般不向下或向上弯曲或者以其它方式显示扩增。这种数据通常呈大致线性,其中较大噪声信号叠加其上。可从实时 PCR 过程或呈现扩增或生

长的其它过程接收该数据。

[0011] 根据本发明的一个示例性实施例,提供了处理数据的方法。该方法通常包括接收具有信号分量和噪声分量的原始数据点集。原始数据集被拟合到线性函数。在一个方面中,该拟合通过对数据集计算线性最小二乘方拟合来实现。该方法还包括计算原始数据集与线性拟合之间的残差,以及计算原始数据集与线性拟合之间的残差量度。在一个方面中,该量度是标准偏差。

[0012] 该方法还通常包括通过计算经平滑的数据集以及计算经平滑的数据集与原始数据集之间的残差来估算数据集中存在的噪声分量。经平滑的数据点基于该经平滑数据点局部的原始数据点的值。在一个方面中,低通滤波器用于计算经平滑的数据集。示例性低通滤波器包括 Savitzky-Golay 滤波器、数字滤波器或数字平滑多项式滤波器。在另一个方面中,经平滑的数据点的值是落在包围该经平滑数据点的窗口内的原始数据点的平均值。

[0013] 该方法通常还包括计算所估算噪声的残差的量度,以及比较量度以便判定原始数据集是否呈现统计线性品质。比较可包括计算第一与第二量度的比值,以便判定比值小于还是大于预定值。在一个方面中,预定值在 1 的量级上。

[0014] 在较佳方面中,在诸如独立计算机、网络附连计算机的处理器或诸如实时 PCR 机器的数据采集设备中实现该方法。实时 PCR 机器的一个示例是由 Bio-Rad Laboratories 提供的 iCycler iQ 系统。

[0015] 根据本发明的另一个示例性实施例,提供了具有适于引导信息处理设备执行处理数据以判定曲线是否呈现线性品质的操作的多个指令的信息存储介质。在一个方面中,信息存储介质是 RAM 或 ROM 单元、硬盘、CD、DVD 或其它便携式介质。

[0016] 根据本发明的另一个示例性实施例,提供了 PCR 检测系统。该 PCR 检测系统包括用于产生具有信号分量和噪声分量的原始数据点集的检测器,并且包括用于处理数据以判定该数据是否呈现线性品质的逻辑。

[0017] 对包括附图和权利要求的本说明书剩余部分的参考将实现本发明的其它特征和优点。以下将参照附图详细描述本发明的其它特征和优点以及本发明各个实施例的结构和操作。在附图中,相似的附图标记表示相同或功能相似的要素。

附图说明

[0018] 图 1 示出了 PCR 扩增曲线的一个示例。

[0019] 图 2 示出了呈现噪声和统计线性品质的实时 PCR 数据集。

[0020] 图 3 示出了呈现噪声和扩增的实时 PCR 数据集。

[0021] 图 4 示出了根据本发明的一个实施例的处理数据集以判定该数据集是否呈现统计线性品质的方法。

[0022] 图 5A 示出了对呈现统计线性品质的数据集进行的线性拟合。

[0023] 图 5B 示出了对呈现扩增品质的数据集进行的线性拟合。

[0024] 图 6A 示出了根据本发明的一个实施例的实时 PCR 数据的经平滑数据集。

[0025] 图 6B 示出了根据本发明的一个实施例的实时 PCR 数据的估算噪声。

[0026] 图 7 示出了根据本发明的一个实施例的处理实时 PCR 数据的系统。

具体实施方式

[0027] 本发明提供了用于处理数据集并识别该数据集是否为统计线性、以及将这种线性数据集与包含扩增信号的数据集区别开的技术。在较佳方面中,对于处理来自 PCR 生长或扩增过程的数据以便在进一步分析该数据之前识别和移除统计线性数据,本发明尤其有用。然而,应该理解,本发明的技术适于处理可能包括噪声的任何数据集或曲线,尤其是除此之外可能呈现诸如细菌生长过程的生长或扩增的曲线。

[0028] 图 1 示出了 PCR 曲线 100 的一个示例,其中绘制了典型 PCR 过程的强度值 110 相对于周期数 120。值 110 可以是任意感兴趣的任意物理量,而周期数可以是与该过程的时间或步骤数相关联的任意单位。这些扩增曲线通常具有线性区域 130、紧随其后的扩增区域 140 以及之后的渐近区域 150,如图 1 所示。可能还存在诸如向下弯曲数据的附加类型的品质。扩增区域可具有指数、反曲 (sigmoidal)、高阶多项式或其它类型的逻辑函数或模拟生长的逻辑曲线。

[0029] 为了理解所涉及的试验过程,识别扩增区域 140 的位置和形状是很重要的。例如,在 PCR 过程中,可能期望识别扩增的开始,它发生在基线区域 (线性区域 130) 末端。识别位置的步骤是识别可能的扩增区域是否确实存在,因为 PCR 过程可能不呈现任何扩增。然而,由于实时 PCR 数据具有噪声,所以可能难以识别数据集是否可能呈现扩增或等价于它是不是统计线性。

[0030] 例如,图 2 示出了由具有包括信号和噪声的数据点 240 的数据集构成的实时 PCR 曲线 200 的线性区域 230。注意,即使对于产生恒定信号的设备,也必须将该数据分割成数据点以便分析。噪声导致数据点的波动。总体而言,数据通常以线性方式向上移动 (即正斜率)。然而,由于曲线 200 在点与点之间完全非线性,所以一般的线性品质无法通过在沿曲线的任意一点上直接分析曲线 200 来判定。直接分析曲线 200 可能会错误地判定数据不呈现统计线性品质。本发明的实施例可有效地判定数据是否呈现统计线性品质。

[0031] 此外,区分具有线性品质的数据曲线和具有扩增品质的数据曲线是很重要的。图 3 示出了呈现扩增的实时 PCR 曲线 300。最初,数据在区域 330 中呈现线性品质,而在随后周期中,区域 340 中存在扩增。本发明的实施例是稳定并一致地精确区分仅具有线性品质的 PCR 曲线 200 和诸如 PCR 曲线 300 的可能具有扩增区域的 PCR 曲线。

[0032] 图 4 示出了根据本发明的一个实施例的处理数据以判定该数据是否呈现统计线性品质的方法 400。数据集由数据点构成,并且表示具有信号分量和噪声分量的曲线。

[0033] 在步骤 405,首先收集或接收数据集。可通过任意机制来接收该数据集。例如,该数据集可通过驻留在诸如 iCycler iQ 设备的 PCR 数据采集设备或类似的 PCR 分析设备中的 (执行指令的) 处理器来采集。可在收集数据时将数据集实时地提供给处理器,或者可将其存储在存储器单元或缓冲器中并在试验完成之后提供给处理器。类似地,可经由到采集设备的网络连接 (例如 LAN、VPN、内联网、因特网等) 或直接连接 (例如 USB 或其它直接接线或无线连接) 将数据集提供给诸如台式计算机系统的分离系统,或者设置在诸如 CD、DVD、软盘等的便携式介质上提供给独立计算机系统。在接收或采集到数据集之后,可对该数据进行分析。

[0034] 在步骤 410,对数据集计算线性拟合。通常,拟合定义了衡量数据集与拟合之间的一致或差异的优质函数 (merit function) S,其中优质函数的较小值通常表示用于拟合的

较好参数。例如,在线性最小二乘方拟合中,优质函数是数据值 Y 与拟合函数 $f(x_i)$ 之间差值的平方,其中对于 N 个数据点, $S = \sum_{i=1}^N (Y_i - f(x_i))^2$ 。在 PCR 过程中, Y 是数据强度,而 x 是周期数。图 5A 示出了 PCR 曲线 200 的线性拟合 510。图 5B 示出了 PCR 曲线 300 的线性拟合 550。

[0035] 对于不同数据点,优质函数可包括对优质函数的不同加权基值 (contribution) 或归一化因子。在取差值之前,优质函数还可按比例换算数据点值或对数据点取函数。可在一 x 值的数据与不同 x 值的 $f(x)$ 之间取差值。例如,优质函数中的一项可表示从数据曲线到线性拟合的线的长度,其中该线与线性拟合垂直。这发生在不同的周期数处,除非线性拟合的斜率为 0。本领域技术人员将意识到可使用的许多不同的优质函数。

[0036] 在步骤 415,计算数据与线性拟合之间的残差 R 。残差 R 是与数据点偏离线性拟合的误差相对应的一组值。例如,残差可以是每个周期数的线性拟合值与实际数据点的差值,表示成 $R_i = Y_i - f(x_i)$,它是残差的标准形式。在一些实施例,残差与用于确定线性拟合的优质函数的值相关。在其它实施例,残差是不同的值。在图 5A 中,误差 520 用于计算曲线 200 与线性拟合 510 之间的残差 R 的值。在图 5B 中,误差 560 用于计算曲线 300 与线性拟合 550 之间的残差 R 的值。

[0037] 在步骤 420 中,计算数据与线性拟合之间的残差量度 σ_1 。该量度是由一组残差值得到的单个值。在一个实施例,残差是标准偏差,表示成 $\sigma_1 = \sqrt{\frac{1}{N} \sum_{i=1}^N R_i^2}$ 。一些实施例可具有各个残差值的加权值,而其它实施例可将各个残差值或所有残差值代入到附加函数或其它函数。本领域技术人员将意识到可使用的许多不同的量度。

[0038] 在步骤 425 中,计算数据集中出现的估算噪声分量。假设数据由真实信号和噪声的两个分量构成。因而,噪声是真实信号与实际数据点之间的差。然而,真实信号不可能直接测量,因为当检测信号时,总是添加有或出现噪声。

[0039] 真实信号被估算为由经平滑的数据点构成的经平滑数据集。图 6A 示出了 PCR 曲线 300 的经平滑的数据集 670。经平滑的数据点的值是基于经平滑的数据点局部的多个原始数据点的函数 G 。术语局部涉及数据点的 x 值与被计算的数据点相距多远。例如,如果一个点与另一个点相差预定数目 (窗口) 的周期,则该点在另一个点局部。三个或五个周期的窗口被证明为适当的,但是也可使用其它窗口,诸如 10 或 20 个周期或以上。还可使用可变窗口值,即可使用不同窗口来计算各个经平滑的数据点。此外,可使用具有分数个周期的窗口,例如其中内插分数个数据点。窗口也可以关于数据点不对称,即可使用数据点之前的一个点以及该数据点之后的三个点。一旦 x 值之差逼近所使用的总范围 - 即周期总数,则该点不再是局部。

[0040] 在一个实施例中,函数 G 是移动平均值或低通滤波器。例如,函数 G 可对指定周期数内的原始数据点取平均值,即居中平均值。因此,在一个实施例中, $G(x_L) = \frac{1}{2K+1} \sum_{i=L-K}^{L+K} Y_i$,

其中 L 是所计算的经平滑的数据点的下标 (index),而 K 为所用窗口。

[0041] 同时,在步骤 425,计算经平滑的数据与原始数据之间的残差。该残差被定义成估算噪声。经平滑的数据与原始数据之间的残差可以以相同方式定义成原始数据与线性拟合

之间的残差,或者这些残差可以以不同方式定义。图 6B 示出了与 PCR 曲线 300 和经平滑的数据集 670 相关联的估算噪声分量 680。噪声分量 680 与信号 670 的叠加形成数据曲线 300。

[0042] 在步骤 430,计算经平滑的数据与原始数据之间的残差量度 σ_2 。 σ_2 值用作固有噪声幅度的量度。在一个实施例中, σ_2 是标准偏差。可以以相似或不同的方式定义量度 σ_1 和 σ_2 。

[0043] 在步骤 435,将第一量度 σ_1 与第二量度 σ_2 进行比较以判定数据集是否呈现线性品质。在一个实施例中,取 σ_1 与 σ_2 的比值。如果比值小于或大于预定值,则数据被判定为呈现线性品质。例如,如果 σ_1/σ_2 小于 1 的量级上的值,例如 1.5,则该数据被判定为线性。等效地,可使用表达式 $\sigma_1 < c_0 * \sigma_2$ 。该表达式说明数据与线性拟合之差的量度必须小于数据中出现的估算噪声的量的常数倍。在某些实施例中, c_0 的值可变化。

[0044] 常数 c_0 与噪声定义以及其它值并不唯一的事实相关。 c_0 的值可通过检查大量数据集来获得,以便于获得该数的合理值。研究表明当使用标准残差的标准偏差时,值 1.5 适于作为常数 (c_0)。当使用其它残差或残差的量度时,其它值可能更适合。通常, c_0 的值在 1 的量级上时应当工作良好。

[0045] 一旦数据已被识别为统计平坦(线性),例如没有向下或向上弯曲或者以其它方式显示扩增,则放弃对该数据的进一步分析。

[0046] 在某些方面中,用于控制处理器以实现本发明的数据处理技术的代码和指令存储在计算机可读介质或信息存储介质上,诸如 RAM 或 ROM 单元、硬盘驱动器、CD、DVD 或其它便携式介质。

[0047] 图 7 示出了根据本发明的一个实施例的系统 700。所示系统包括在样品容器 710 内的诸如细菌或 DNA 的样品 705。来自样品的诸如荧光强度值的物理特征 715 由检测器 720 来检测。包括噪声分量的信号 725 从检测器 720 发送到逻辑系统 730。来自信号 725 的数据可存储在本地存储器 735 或外部存储器 740 或存储设备 745 中。在一个实施例中,模-数转换器将模拟信号转换成数字形式。

[0048] 逻辑系统 730 可以是或可包括计算机系统、ASIC、微处理器等。它还可包括显示器(例如监视器、LED 显示器等)和用户输入设备(例如鼠标、键盘、按钮等)或与它们耦合。逻辑系统 730 和其它组件可以是独立或网络连接计算机系统的一部分,或者它们可直接附加到或结合到热循环设备。逻辑系统 730 还可包括在处理器 750 中执行的优化软件。

[0049] 根据一个实施例,逻辑系统 730 包括用于处理数据和识别统计平坦数据的指令。尽管这些指令也可在诸如软盘、CD、DVD 等的任何软件存储介质上提供,但是较佳地将指令下载并存储在存储器模块 735、740 或 745(例如硬盘驱动器或诸如本地或附加 RAM 或 ROM 的其它存储器)中。应该理解,可通过诸如 C、C++、Java、Visual Basic 及其它、或诸如 VBScript、JavaScript、Perl 的任何脚本语言、或诸如 XML 的标记语言来实现用于实现本发明的诸方面的计算机代码。此外,各种语言和协议可用于根据本发明的诸方面的数据和命令的外部 and 内部存储以及传输。

[0050] 应该理解,本文所述的过程是说明性的,而且变化和更改是可能的。顺序描述的步骤可并行执行,步骤的次序可改变,并且步骤可被更改或组合。

[0051] 虽然本发明作为示例并就具体实施例而得到描述,但是应该理解本发明并不局限

于所公开的实施例。相反,它旨在涵盖对于本领域技术人员而言显而易见的各种更改和相似配置。因此,对所附权利要求书的范围应当给予最宽泛的解释,以便包括所有这些更改和相似配置。

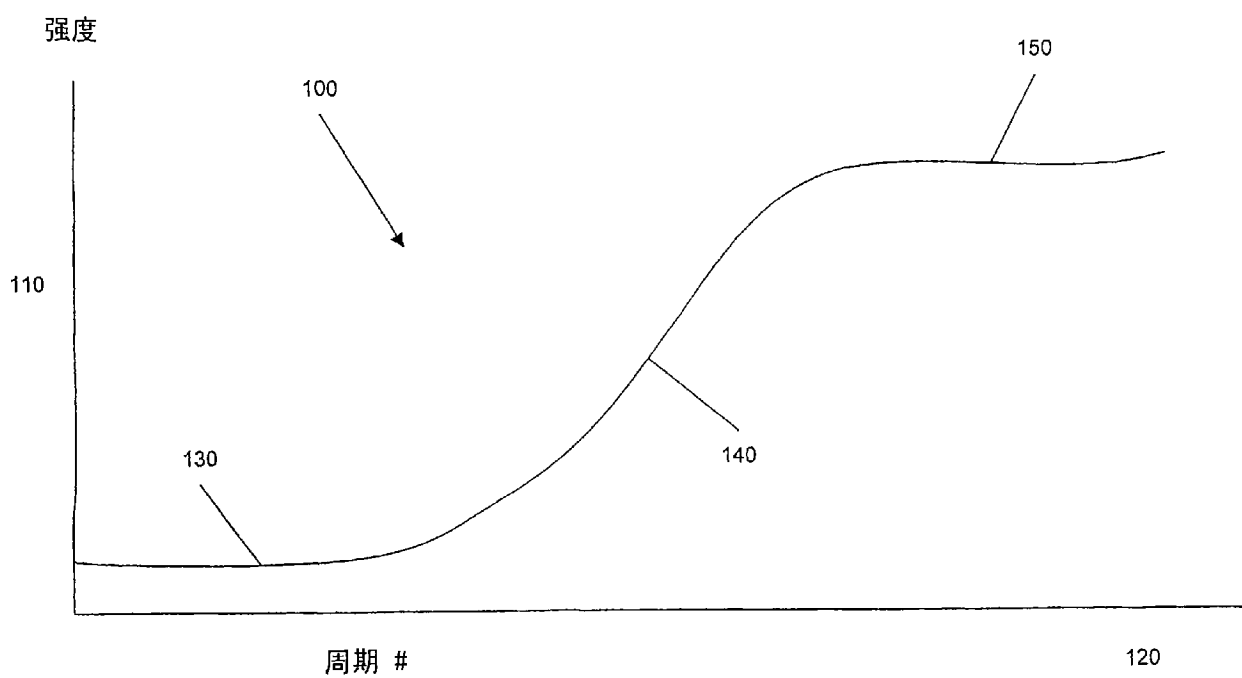


图 1

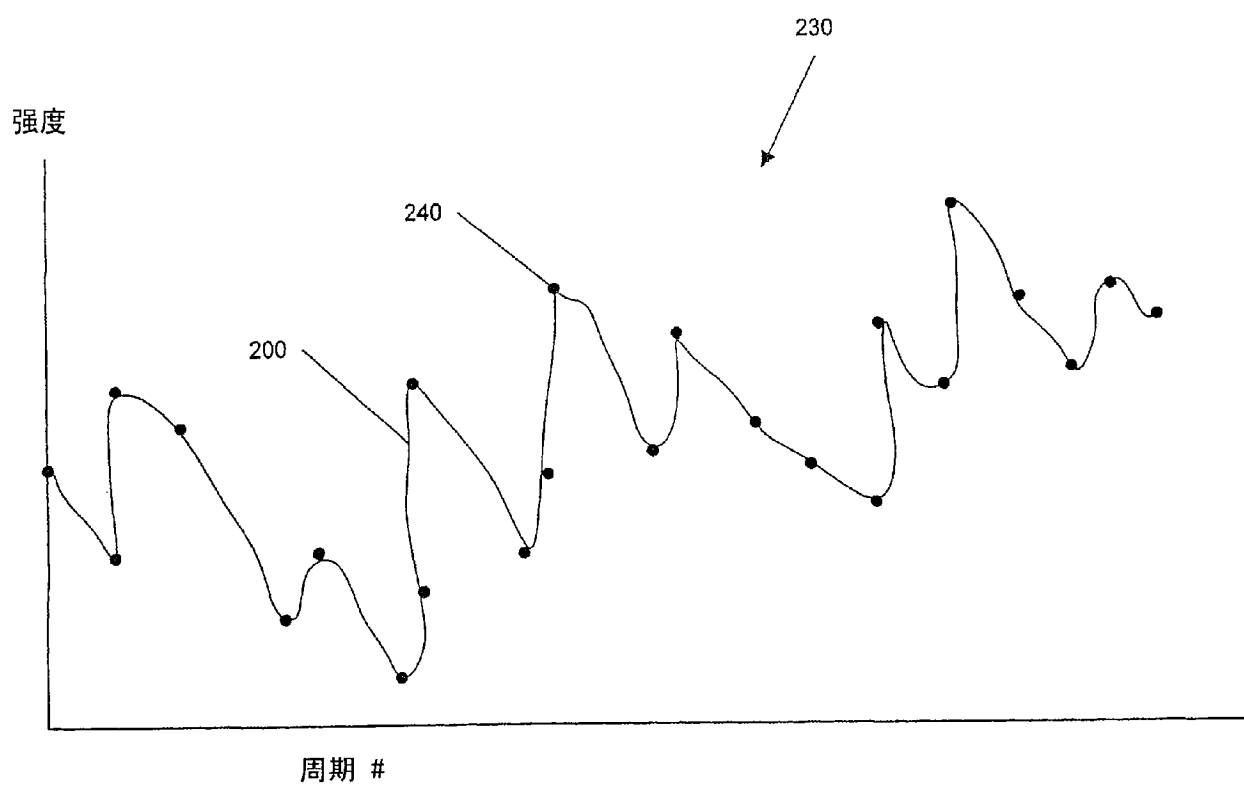


图 2

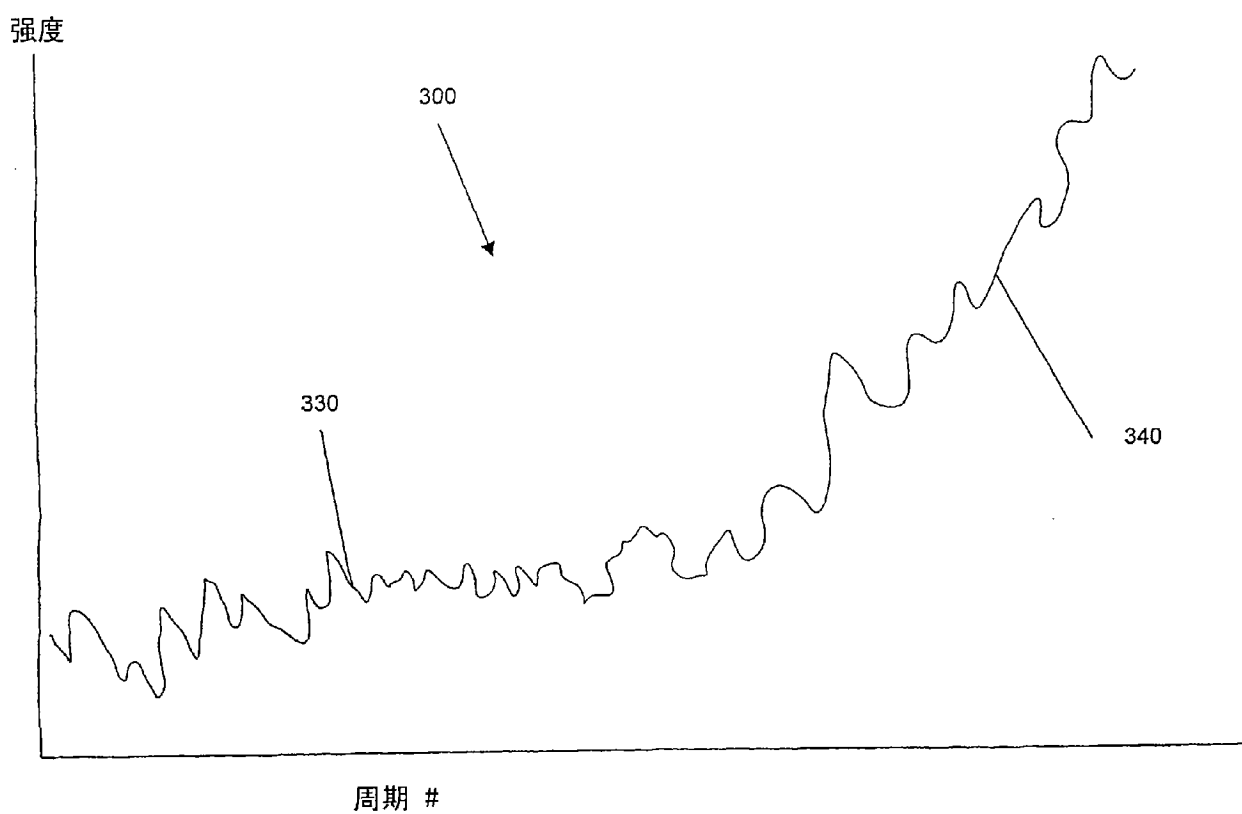


图 3

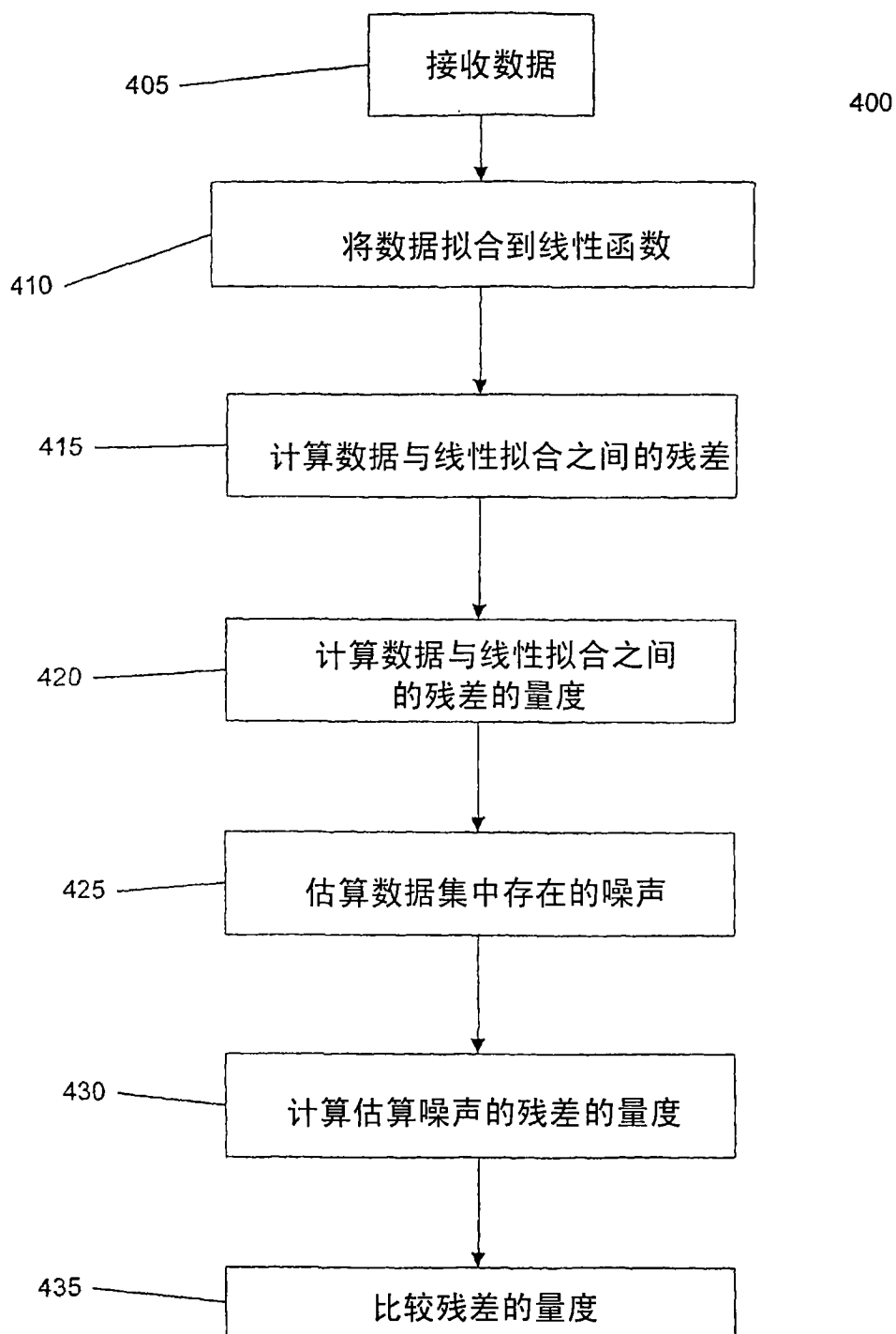


图 4

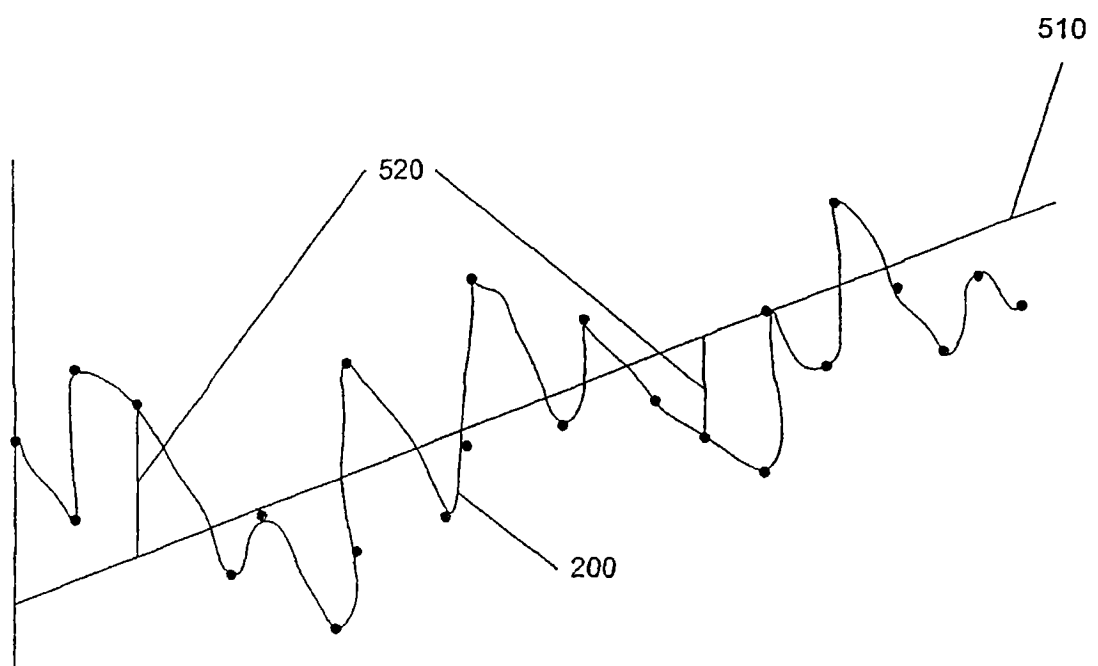


图 5A

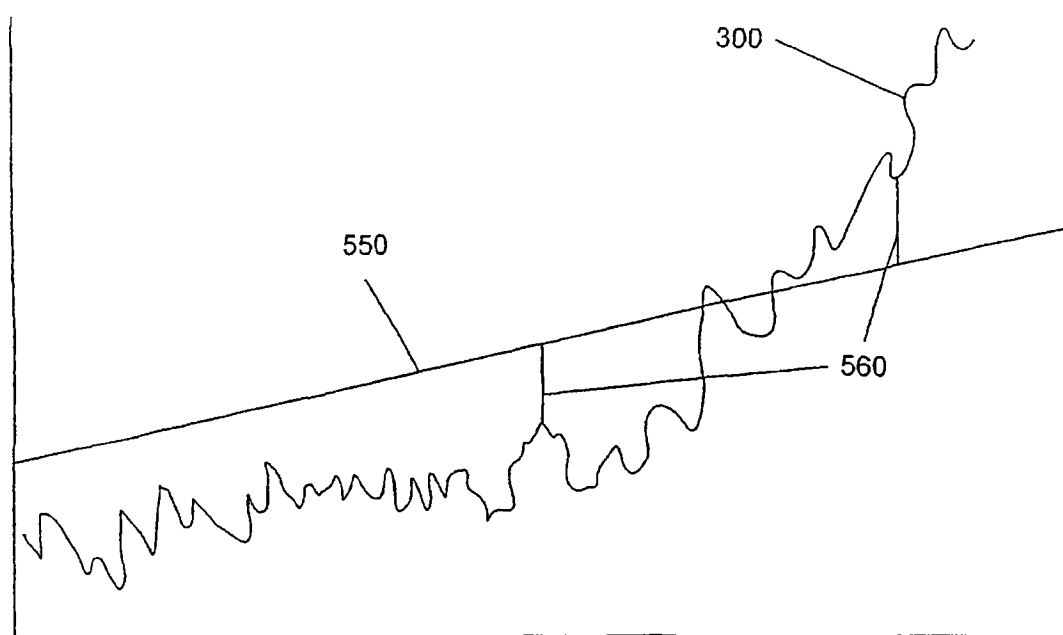


图 5B

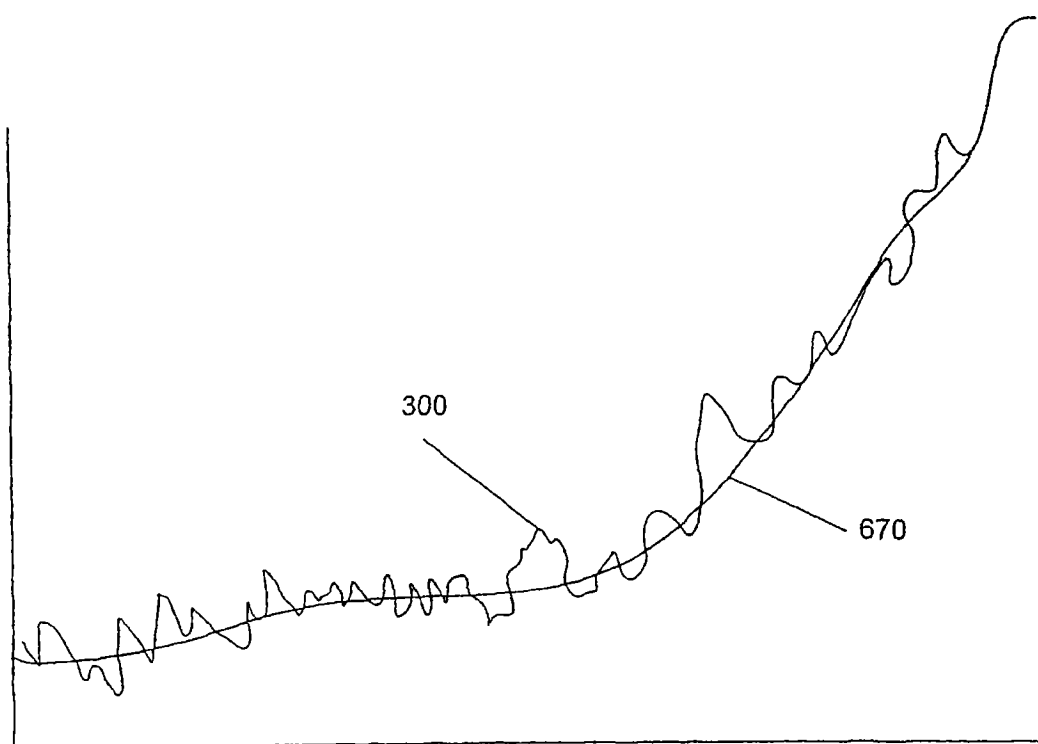


图 6A

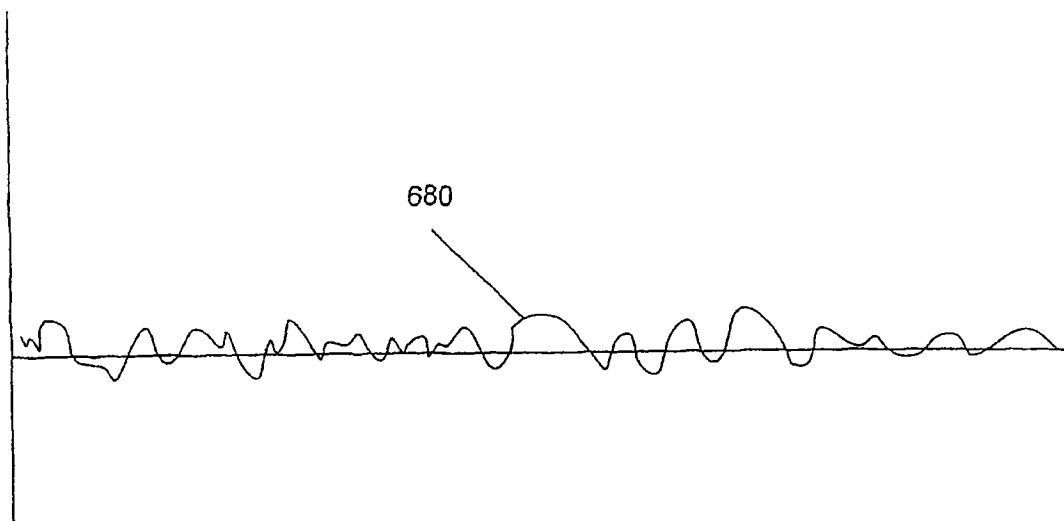


图 6B

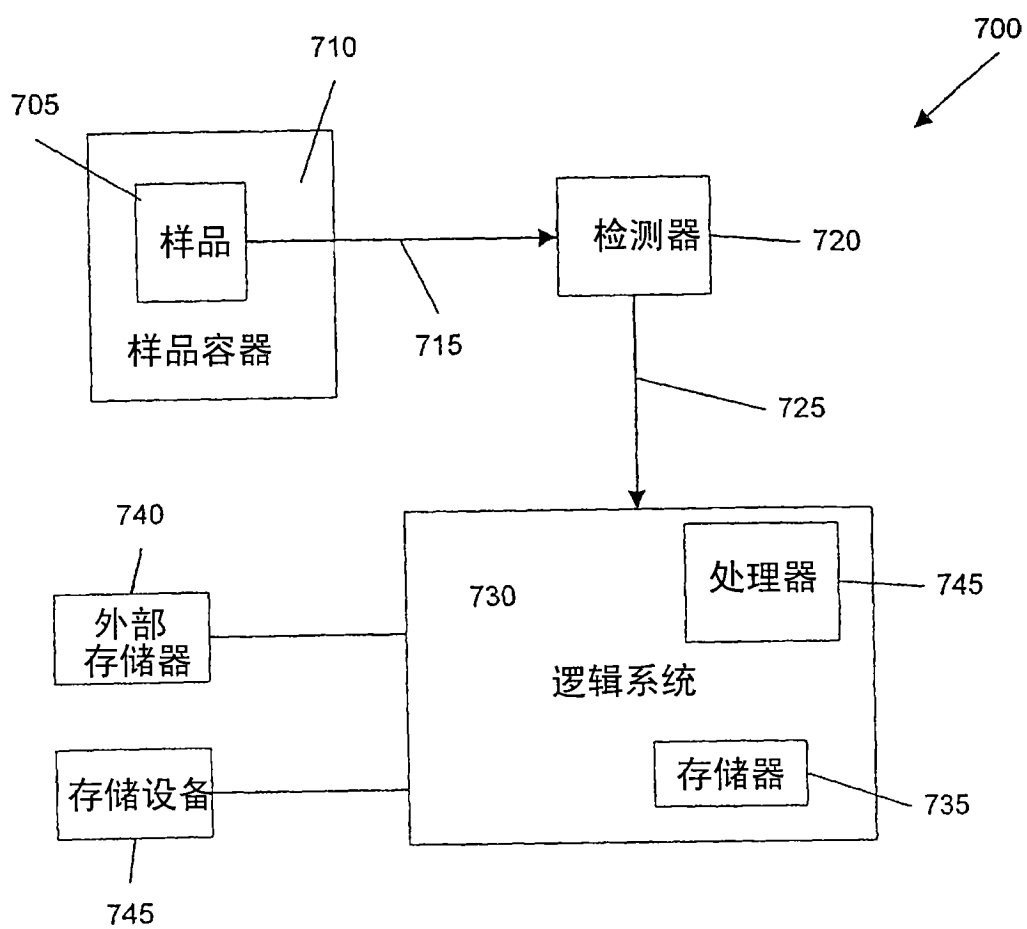


图 7