



(12) 发明专利

(10) 授权公告号 CN 107109490 B

(45) 授权公告日 2022. 12. 02

(21) 申请号 201580069073.9
(22) 申请日 2015.10.16
(65) 同一申请的已公布的文献号
申请公布号 CN 107109490 A

(43) 申请公布日 2017.08.29

(30) 优先权数据
1418366.9 2014.10.16 GB
1418379.2 2014.10.16 GB
1507742.3 2015.05.06 GB

(85) PCT国际申请进入国家阶段日
2017.06.16

(86) PCT国际申请的申请数据
PCT/GB2015/053083 2015.10.16

(87) PCT国际申请的公布数据
W02016/059427 EN 2016.04.21

(73) 专利权人 牛津楠路珀尔科技股份有限公司
地址 英国牛津

(72) 发明人 斯图尔特·威廉·里德
加文·哈珀 克莱夫·加文·布朗
丹尼尔·约翰·特纳
安德鲁·约翰·赫伦
克里斯托弗·詹姆斯·怀特

(74) 专利代理机构 北京康信知识产权代理有限公司 11240
专利代理师 张英 宫传芝

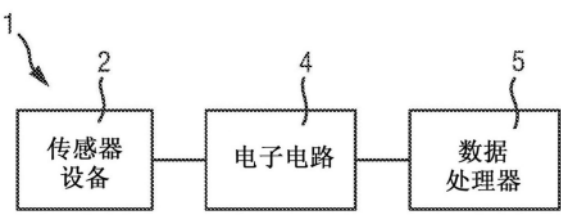
(51) Int.Cl.
G12Q 1/6869 (2018.01)
G01N 33/48 (2006.01)

(56) 对比文件
CN 104066850 A, 2014.09.24
WO 2014096830 A1, 2014.06.26
审查员 蒋红云

权利要求书4页 说明书36页
序列表11页 附图13页

(54) 发明名称
聚合物的分析

(57) 摘要
一种聚合物的分析,一种生化分析系统通过从包括纳米孔的传感器元件,在聚合物移位穿过纳米孔期间,采集聚合物的测量值来分析聚合物。当聚合物已部分移位时,使用来源于参考序列的参考数据来分析一系列测量值以提供相似性的量度。响应于相似性的量度,传感器元件可以选择性操作以排出聚合物,并且从而使纳米孔可用于接收另外的聚合物。其中生化分析系统包括传感器元件的阵列,并且以多路复用方式从选择的传感器元件采集测量值,响应于相似性的量度,生化分析系统停止从当前选择的传感器元件采集测量值并开始从新选择的传感器元件采集测量值。



1. 一种控制用于分析聚合物的生化分析系统的方法,所述聚合物包含聚合物单元的序列,其中,所述生化分析系统包括至少一个包含纳米孔的传感器元件,以及所述生化分析系统可操作为在所述聚合物移位穿过所述传感器元件的所述纳米孔期间,从传感器元件采集聚合物的连续测量值,

其中,所述方法包括,当聚合物已部分移位穿过所述纳米孔时,使用来源于聚合物单元的至少一种参考序列的参考数据来分析在所述聚合物部分移位期间采集自所述聚合物的一系列测量值,以提供已部分移位的聚合物的聚合物单元的序列与所述至少一种参考序列之间的相似性的量度,以及

响应于所述相似性的量度,操作所述生化分析系统以排斥所述聚合物并采集来自另外的聚合物的测量值。

2. 根据权利要求1所述的方法,其中,至少一个所述传感器元件可操作为排出正在移位穿过所述纳米孔的聚合物,并且操作所述生化分析系统以排斥所述聚合物并采集来自另外的聚合物的测量值的步骤包括,操作所述传感器元件以从所述纳米孔中排出所述聚合物,并且在所述纳米孔中接收另外的聚合物。

3. 根据权利要求2所述的方法,其中,至少一个所述传感器元件可操作为通过施加足以排出所述聚合物的排出偏压来排出正在移位穿过所述纳米孔的聚合物,操作所述传感器元件以从所述纳米孔中排出所述聚合物的步骤通过施加排出偏压进行,以及操作所述传感器元件以在所述纳米孔中接收另外的聚合物的步骤通过施加足以使另外的聚合物能够移位穿过其中的移位偏压进行。

4. 根据权利要求1所述的方法,其中,所述生化分析系统包括传感器元件的阵列,并可操作为以多路复用方式从所选择的传感器元件采集聚合物的连续测量值,以及操作所述生化分析系统以排斥所述聚合物并采集来自另外的聚合物的测量值的步骤包括,操作所述生化分析系统以停止从当前选择的所述传感器元件采集测量值并开始从新选择的传感器元件采集测量值。

5. 根据权利要求4所述的方法,其中,所述测量值包括从所述传感器元件采集的电测量值,并且所述生化分析系统可操作为以电多路复用方式从所选择的传感器元件采集聚合物的连续测量值。

6. 根据权利要求5所述的方法,其中,所述生化分析系统包括:

检测电路,包括多个检测通道,各自能够采集来自传感器元件的电测量值,所述阵列中的传感器元件的数目大于检测通道的数目;和

开关布置,能够以多路复用方式选择性地所述检测通道连接至各个传感器元件。

7. 根据权利要求4至6中任一项所述的方法,其中,所述传感器元件可控制为排出正在移位穿过所述传感器元件的所述纳米孔的聚合物,并且所述方法进一步包括,当操作所述生化分析系统以停止从当前选择的所述传感器元件采集测量值时,还控制当前选择的所述传感器元件以排出所述聚合物,并且由此使所述纳米孔可用于接收另外的聚合物。

8. 根据前述权利要求中任一项所述的方法,其中,聚合物单元的所述至少一种参考序列是不期望的序列,所述参考数据来源于聚合物单元的所述至少一种参考序列,以及选择性操作的所述步骤响应于相似性的量度来进行,所述相似性的量度表明已部分移位的所述聚合物是不期望的序列。

9. 根据权利要求1至7中任一项所述的方法, 其中, 聚合物单元的所述至少一种参考序列是靶标, 所述参考数据来源于聚合物单元的所述至少一种参考序列, 以及选择性操作的所述步骤响应于相似性的量度来进行, 所述相似性的量度表明已部分移位的聚合物不是靶标。

10. 根据权利要求1至7中任一项所述的方法, 其中, 聚合物单元的所述至少一种参考序列是聚合物单元的已测量的序列, 所述参考数据来源于聚合物单元的所述至少一种参考序列, 以及选择性操作的所述步骤响应于相似性的量度进行, 所述相似性的量度表明已部分移位的聚合物是聚合物单元的已测量的序列。

11. 根据权利要求1至7中任一项所述的方法, 其中, 聚合物单元的所述至少一种参考序列包括多个靶标, 所述参考数据来源于聚合物单元的所述至少一种参考序列, 以及选择性操作的所述步骤响应于相似性的量度进行, 所述相似性的量度表明已部分移位的聚合物是所述靶标中的一个。

12. 根据前述权利要求中任一项所述的方法, 其中,

来源于聚合物单元的至少一种参考序列的所述参考数据表示从生化分析系统采集的实际的或模拟的测量值, 以及

分析在所述部分移位期间采集的所述聚合物的一系列所述测量值的所述步骤包括:

比较一系列所述测量值与所述参考数据。

13. 根据权利要求1至11中任一项所述的方法, 其中, 来源于聚合物单元的至少一种参考序列的所述参考数据表示时序特征的特征向量, 所述时序特征的特征向量表示通过生化分析系统采集的所述测量值的特性, 以及

分析在所述部分移位期间采集的所述聚合物的一系列所述测量值的所述步骤包括:

由一系列所述测量值衍生出表示所述测量值的特性的时序特征的特征向量, 以及

比较衍生的所述特征向量与所述参考数据。

14. 根据权利要求1至11中任一项所述的方法, 其中,

来源于聚合物单元的至少一种参考序列的所述参考数据表示所述至少一种参考序列的所述聚合物单元的一致性, 并且

分析在所述部分移位期间采集的来自所述聚合物的一系列所述测量值的所述步骤包括:

分析一系列所述测量值以提供已部分移位的聚合物的聚合物单元的序列的所述聚合物单元的一致性的估算值, 以及

比较所述估算值与所述参考数据以提供所述相似性的量度。

15. 根据权利要求1至11中任一项所述的方法, 其中,

所述测量值取决于k聚体, 所述k聚体是聚合物的k个聚合物单元, 其中k是正整数;

所述参考数据表示参考模型, 所述参考模型处理测量值作为对应于聚合物单元的所述参考序列的k聚体状态的参考系列的观察值, 其中, 所述参考模型包括:

用于k聚体状态的所述参考系列中的所述k聚体状态之间的转换的转换加权; 和

对于每个k聚体状态, 当观察所述k聚体状态时, 用于所观察的不同测量值的发射加权, 以及

分析在所述部分移位期间采集自所述聚合物的一系列所述测量值的所述步骤包括, 将

所述模型拟合至一系列所述测量值以提供相似性的量度作为所述模型与一系列所述测量值的拟合。

16. 根据前述权利要求中任一项所述的方法, 其中, 所述测量值取决于k聚体, 所述k聚体是聚合物的k个聚合物单元, 其中k是正整数。

17. 根据前述权利要求中任一项所述的方法, 其中, 所述纳米孔是生物孔。

18. 根据前述权利要求中任一项所述的方法, 其中, 所述聚合物是多核苷酸, 以及所述聚合物单元是核苷酸。

19. 根据前述权利要求中任一项所述的方法, 其中, 所述聚合物穿过纳米孔的移位以棘轮方式进行。

20. 根据前述权利要求中任一项所述的方法, 其中, 所述测量值包括电测量值。

21. 一种用于分析聚合物的生化分析系统, 所述聚合物包含聚合物单元的序列, 其中, 所述生化分析系统包括至少一个包含纳米孔的传感器元件, 并且所述生化分析系统可操作为在所述聚合物移位穿过所述传感器元件的所述纳米孔期间, 从传感器元件采集聚合物的连续测量值,

其中, 所述生化分析系统布置为当聚合物已部分移位穿过所述纳米孔时, 使用来源于聚合物单元的至少一种参考序列的参考数据分析在所述聚合物的部分移位期间采集的所述聚合物的一系列测量值, 以提供已部分移位的聚合物的聚合物单元的序列与所述至少一种参考序列之间的相似性的量度, 以及

所述生化分析系统布置为响应于所述相似性的量度, 排斥所述聚合物并采集来自另外的聚合物的测量值。

22. 一种控制用于分析包含聚合物单元的序列的聚合物的生化分析系统的方法, 其中, 所述生化分析系统包括至少一个包含纳米孔的传感器元件, 并且所述生化分析系统可操作为在所述聚合物移位穿过所述传感器元件的所述纳米孔期间, 从传感器元件采集聚合物的连续测量值,

其中, 所述方法包括, 当聚合物已部分移位穿过所述纳米孔时, 通过衍生拟合至模型的量度来分析在所述聚合物的部分移位期间, 采集自所述聚合物的一系列所述测量值, 所述模型处理所述测量值作为不同的可能类型的一系列k聚体状态的观察值, 并且包括: 转换加权, 对于所述一系列k聚体状态中的连续的k聚体状态之间的每个转换, 用于在k聚体状态的可能的类型之间的可能转换; 以及发射加权, 对于每种类型的k聚体状态, 其表示观察给定的所述k聚体的测量值的机率, 以及

响应于所述拟合的量度, 操作所述生化分析系统以排斥所述聚合物并采集来自另外的聚合物的测量值。

23. 一种用于分析包含聚合物单元的序列的聚合物的生化分析系统, 其中, 所述生化分析系统包括至少一个包含纳米孔的传感器元件, 并且所述生化分析系统可操作为在所述聚合物移位穿过所述传感器元件的所述纳米孔期间, 从传感器元件采集聚合物的连续测量值,

其中, 所述生化分析系统布置为当聚合物已部分移位穿过所述纳米孔时, 通过衍生拟合至模型的量度来分析在所述聚合物的部分移位期间, 采集自所述聚合物的一系列所述测量值, 所述模型处理所述测量值作为不同的可能类型的一系列k聚体状态的观察值, 并且包

括:转换加权,对于在所述一系列k聚体状态中的连续的k聚体状态之间的每个转换,用于在所述可能类型的k聚体状态之间的可能的转换;以及发射加权,对于每种类型的k聚体状态,其表示观察给定的所述k聚体的测量值的机率,以及

响应于所述拟合的量度,所述生化分析系统布置为排斥所述聚合物并采集来自另外的聚合物的测量值。

聚合物的分析

技术领域

[0001] 本发明的第一至第三方面涉及使用生化分析系统分析聚合物,该生化分析系统包括包含纳米孔的至少一个传感器元件。本发明的第四方面涉及包含聚合物单元的聚合物的一系列测量值与聚合物单元的参考序列之间的对齐映射的估算。在所有方面,聚合物可以是,例如但不限于多核苷酸,其中聚合物单元是核苷酸。

背景技术

[0002] 存在多种类型的生化分析系统,其提供用于确定序列的聚合物单元的测量值。例如,但是在没有限制的情况下,一类测量系统使用纳米孔。使用纳米孔的生化分析系统是近来发展较多的对象。通常,在聚合物移位(translocation,易位,移动)穿过纳米孔期间,由包括纳米孔的传感器元件采集聚合物的连续测量值。系统的一些特性取决于在纳米孔中的聚合物单元,并且采集该特性的测量值。使用纳米孔的这类测量系统具有相当大的前景,特别是在多核苷酸,如DNA或RNA测序领域。

[0003] 使用纳米孔的这种生化分析系统可以提供聚合物的长的连续读取,例如在多核苷酸的情况下,从数百至数十万个(以及可能更多)核苷酸的范围。以这种方式搜集的数据包括测量值,如离子电流的测量值,其中序列通过纳米孔的灵敏部分的每次移位导致在测量特性上的轻微的变化。

发明内容

[0004] 当使用纳米孔的这类生化分析系统可以提供显著的优点时,还期望其增加分析速度。本发明的第一和第二方面涉及提供这种增加。

[0005] 根据本发明的第一方面,提供了控制用于分析聚合物的生化分析系统的方法,所述聚合物包含聚合物单元的序列,其中,生化分析系统包括包含纳米孔的至少一个传感器元件,并且生化分析系统可操作为,在聚合物移位穿过传感器元件的纳米孔期间,由传感器元件采集聚合物的连续测量值,

[0006] 其中,方法包括当聚合物部分移位穿过纳米孔时,使用来源于聚合物单元的至少一种参考序列的参考数据分析,在聚合物的部分移位期间采集的聚合物的一系列测量值,以提供部分移位的聚合物的聚合物单元的序列与至少一种参考序列之间的相似性的量度,以及

[0007] 响应于相似性的量度,操作生化分析系统以排斥聚合物并采集另外的聚合物的测量值。

[0008] 这种方法涉及分析,当聚合物部分移位穿过纳米孔(即在聚合物移位穿过纳米孔期间)时,采集自聚合物的测量值。特别地,使用来源于聚合物单元的至少一种参考序列的参考数据,分析在部分移位期间采集的聚合物的一系列测量值。该分析提供了部分移位的聚合物的聚合物单元的序列与至少一种参考序列之间的相似性的量度。响应于该相似性的量度,如果参考序列的相似性指示不需要进一步分析聚合物,例如因为测量的聚合物不是

感兴趣的,可以进行排斥聚合物以采集另外的聚合物的测量值。

[0009] 排斥聚合物允许在不完成初始测量的聚合物的测量的情况下进行另外的聚合物的测量。这提供了采集测量值中的时间节省,因为在“不工作时 (on-the-fly)” (即进行聚合物的测量期间) 进行操作。在典型的施加中,这种时间节省可以是显著的,因为使用纳米孔的生化分析系统可以提供聚合物的长的连续读取,而分析可以在早期阶段,在这种读取中鉴定,对当前测量的聚合物,不需要进一步的测量。

[0010] 例如在聚合物是多核苷酸的典型施加中,在100%准确度下进行的测序将允许在测量约30个核苷酸之后进行初步确定。因此,考虑实际可实现的准确度,可以在测量几百个核苷酸、典型的250个核苷酸之后进行确定。将这种与能够对长度从数百到数万 (还可能更多) 核苷酸的范围的序列进行测量的生化分析系统进行比较。

[0011] 该方法对于结果潜在地提供了显著更快的时间,其中仅对确定感兴趣的那些聚合物进行继续的测量并排斥确定不感兴趣的那些。这种减少废弃数据获取的量的优点对于要求大量数据获取的施加是特别显著的。产生的时间节省对于本身是有用的,或者可以用于,例如得到更大的覆盖度,并因此通过可获得的时间和资源可以另外得到更高的测序准确度。

[0012] 提供在部分移位的聚合物的聚合物单元的序列与至少一种参考序列之间的相似性的量度的分析,可以本身使用用于比较测量值与参考的已知技术。然而,与本方法相反,这种已知的技术典型地在完成移位之后进行测量。

[0013] 方法可以应用于各种各样的应用。根据应用,相似性的量度可以表示与参考序列的整体或与参考序列的一部分的相似性。

[0014] 根据本发明的第二方面,提供了控制用于分析聚合物的生化分析系统的方法,所述聚合物包含聚合物单元的序列,其中,生化分析系统包括包含纳米孔的至少一个传感器元件,并且生化分析系统可操作为在聚合物移位穿过传感器元件的纳米孔期间由传感器元件采集聚合物的连续测量值,

[0015] 其中,方法包括当聚合物部分移位穿过纳米孔时,通过衍生与拟合的量度分析在聚合物的部分移位期间采集自聚合物的一系列测量值,所述模型处理测量值作为不同可能类型的一系列k聚体 (k-mer) 状态的观察值 (observation, 观测值),模型包括:转换加权,相对于在一系列k聚体状态中的连续的k聚体状态之间的每个转换,用于在可能类型的k聚体状态之间的可能转换;以及发射加权,相对于每种类型的k聚体状态表示观察给定k聚体的测量值的机率,以及

[0016] 响应于拟合的量度,操作生化分析系统以排斥聚合物并采集另外的聚合物的测量值。

[0017] 这种方法涉及分析当聚合物部分移位穿过纳米孔、即在聚合物移位穿过纳米孔期间时采集自聚合物的测量值。特别地,使用来源于聚合物单元的至少一种参考序列的参考数据分析,在部分移位期间采集自聚合物的一系列测量值。该分析提供模型的拟合量度。响应于这种拟合量度,如果如通过模型确定的拟合的量度表示测量具有不良的质量,使得不需要另外的移位和测量,则可以采取行动将聚合物排斥并采集另外的聚合物的测量值。

[0018] 排斥聚合物允许在不完成初始测量的聚合物的测量的情况下采集另外的聚合物的测量值。这提供了采集测量值中的时间节省,因为该操作在“不工作 (on-the-fly)”、即采

集聚合物的测量值期间进行。在典型的施加中,这种时间节省可以是显著的,因为使用纳米孔的生化分析系统可以提供聚合物的长的连续读取,尽管分析可以在早期识别测量值具有不良的质量。

[0019] 本发明的第一和第二方面是相同的,除了生化分析系统操作为排斥聚合物并采集另外的聚合物的测量值的基础。因此,可以将从属权利要求2至20中阐述的根据本发明的第一方面的可选的特征进行必要的变更应用于本发明的第二方面。同样,方法的所有以下特征同样施加于根据本发明的第一或第二方面的方法。

[0020] 排斥聚合物可以以不同的方式发生。

[0021] 在第一途径中,至少一个传感器元件可操作为排出移位穿过纳米孔的聚合物。在该情况下,可以通过操作传感器元件来由纳米孔排出聚合物以及在纳米孔中接收另外的聚合物,来进行操作生化分析系统以排斥聚合物并采集另外的聚合物的测量值的步骤。

[0022] 在第二途径中,生化分析系统包括阵列传感器元件并操作为由多路复用(multiplexed,多路传输)方式选择的传感器元件采集聚合物的连续测量值。在该情况下,操作生化分析系统以排斥聚合物并采集另外的聚合物的测量值的步骤,可以包括采集生化分析系统以停止由当前选择的传感器元件采集测量值并开始由新选择的传感器元件采集测量值。

[0023] 可以组合使用这两种途径。

[0024] 本发明的第三方面涉及可以使用纳米孔进行的特定形式的生化分析的施加。

[0025] 根据本发明的第三方面,提供了分类聚合物的方法,所述聚合物各自包含聚合物单元的序列,该方法使用包括以下各项的系统:包括包含含有聚合物的样品的样品室、与样品室密封的收集室、和包含在样品室和收集室之间连通的纳米孔的传感器元件,

[0026] 该方法包括引起连续的聚合物从样品室开始移位穿过纳米孔,以及在每个聚合物的移位期间:

[0027] 由传感器元件采集聚合物的连续测量值;

[0028] 使用来源于聚合物单元的至少一种参考序列的参考数据,分析在聚合物的部分移位期间,采集自聚合物的一系列测量值来提供部分移位的聚合物的聚合物单元的序列与至少一种参考序列之间的相似性的量度,

[0029] 根据相似性的量度,选择性完成聚合物到收集室的移位或排出聚合物返回到样品室中。

[0030] 因此,该方法利用相似性的量度,其通过分析在部分移位期间采集自聚合物的一系列测量值提供。该分析可以本身使用已知的用于比较,测量值与参考的技术。然而,将相似性的量度用于确定是否收集聚合物。如果这样,那么完成聚合物到收集室中的移位。另外,将聚合物排出回到样品室中。以这种方式,将选择的聚合物收集到收集室中。例如,在完成聚合物从样品的移位之后,或可替换地,在聚合物从样品的移位期间,例如,通过提供系统(具有适用于其的流体系统)可以回收收集的聚合物。

[0031] 该方法可以应用于各种各样的应用。例如,该方法可以应用于多核苷酸,例如病毒基因组或质粒的聚合物。病毒基因组典型地具有10-15kB(千碱基)级别的长度,以及质粒典型地具有4kB级别的长度。在这种实例中,不需要将多核苷酸片段化并可以整个收集。可以以任何方式使用收集的病毒基因组或质粒,例如用于转染细胞。

[0032] 来源于参考数据的聚合物单元的参考序列可以是期望的序列。在这种情况下,响应于表示部分移位的聚合物是期望的序列的相似性的量度,进行选择性完成聚合物移位到收集室中的步骤。然而,这不是必需的。在一些施加中,来源于参考数据的聚合物单元的参考序列可以是不期望的序列。在这种情况下,响应于表示部分移位的聚合物并非不期望的序列的相似性的量度,进行选择性完成聚合物移位到收集室中的步骤。

[0033] 取决于应用,相似性的量度可以表示与参考序列的整体或与参考序列的一部分的相似性。

[0034] 该系统可以包括多个收集室,以及有关每个收集室,包括提供样品室与各个收集室之间的连通的纳米孔的传感器元件。这允许相对于平行的多个纳米孔进行所述方法。与提供加速分选方法的能力一样,其可以允许将不同的聚合物收集到不同的收集室中。为了实现此目的,相应地选择用于收集的参考数据和标准。在一个实例中,可以使用关于不同的纳米孔的不同的参考数据进行该方法。在另一实施例中,可以关于不同的纳米孔使用相同的参考数据进行该方法,但是关于不同的纳米孔用对相似性的量度的不同依赖性进行选择性地完成聚合物到收集室中的移位的所述步骤。

[0035] 根据本发明的进一步的方面,提供了生化分析系统,该生化分析系统进行与本发明的第一、第二或第三方面的方法的那些类似的方法。

[0036] 本发明的第四方面涉及包含聚合物单元的聚合物的一系列测量值与聚合物单元的参考序列之间的对齐(alignment)。

[0037] 一些类型的测量系统采集取决于k聚体的聚合物的测量值,k聚体是聚合物的k个聚合物单元,其中k是整数。通过定义,下文中将k个聚合物的组称为k聚体。一般而言,k可以采用值1,在这样的情况下,k聚体是单个聚合物单元,或者其可以是复数(复数整数,plural integer)。根据聚合物的本质,给定的每个聚合物单元可以是不同的类型。例如,在聚合物是多核苷酸的情况下,聚合物单元是核苷酸,并且不同类型是包含不同的核酸碱基(如胞嘧啶、鸟嘌呤等)的核苷酸。因此,对应于k聚体的每个聚合物单元的不同类型的不同组合,给定的每种k聚体也可以具有不同的类型。

[0038] 对于由测量值估算聚合物单元,在实际类型的测量系统中,难以提供取决于单个聚合物单元的测量值。反而,每次测量的值取决于k聚体,其中k是复数。在概念上,这可以被认为是具有大于被测量的聚合物单元的“钝性读取头”的测量系统。在这种情况下,待解析的不同的k聚体的数目增加到k的乘方。当测量值取决于大量的聚合物单元(较大的k值)时,可能难以解析取自不同类型的k聚体的测量值,因为它们提供重叠的信号分布,特别是当考虑测量系统中的杂音和/或伪迹(artefacts)时。这有损于估算聚合物单元的基本序列。

[0039] 当k是复数时,可能结合来自重叠的k聚体的多个测量值(各自部分地取决于相同的聚合物单元)的信息,以得到在聚合物单元的水平下解析的单个值。举例来说,W0-2013/041878公开了,利用关于作为一系列的不同的可能的类型的k聚体的观察值处理测量值的模型由与聚合物有关的测量值中的至少一种系列估算在聚合物中的聚合物单元的序列的方法。该模型包括:转换加权,关于一系列k聚体状态中的连续的k聚体状态之间的每个转换,用于在可能类型的k聚体状态之间的可能的转换;以及发射加权,关于每种类型的k聚体状态表示观察的给定k聚体的测量值的机率(chance)。该模型可以是,例如隐马尔可夫模型(Hidden Markov Model (HMM))。这种模型可以通过在考虑,由聚合物单元的序列产生的一

系列测量值的模型所预测的可能性时考虑到多个测量值来改善估算的准确度。

[0040] 在多种情况下,期望估算包含聚合物单元的聚合物的一系列测量值与聚合物单元的参考序列之间的对齐映射。这种对齐映射的估算可以用于各种应用,如对参考进行比较以提供样品中的聚合物的存在、不存在或程度的识别或检测,例如来提供诊断。可能范围的特定应用是大量的,并可以应用于检测具有DNA序列的任何分析物。

[0041] 现有的技术涉及初始估算已经测量的聚合物单元的序列,然后通过比较聚合物单元的一致性(identity,统一性,特性)估算与聚合物单元的参考序列的对齐映射。已经开发了多种快速的对齐算法用于施加至聚合物单元是核苷酸(在文献中经常称为碱基)的情况。快速对齐算法的实例是BLAST(基本局部序列对齐检索工具(Basic Local Alignment Search Tool))、FASTA和HMMER、和它们的衍生物。快速对齐算法典型地寻找高度类似的较小区域,这是相对迅速的过程,然后延伸到相似性低的较大区域,这是缓慢的过程。这种算法已经应用施加于以下情形,其中它们通过在最小的时间帧中提供关于测量的聚合物是否匹配参考的相似性得分来表示聚合物的一致性。在这些类型的技术中,直接比较估算序列和参考序列中的聚合物单元的一致性。当提及作为碱基的聚合物单元时,与作为“测量值间隔”的测量值之间的比较相比,可以认为该技术涉及“碱基间隔”的比较。

[0042] 然而,这种技术在估算对齐映射上具有有限的准确度,或换言之,具有有限的判别能力。这是因为估算聚合物单元的序列的初始步骤内在地引起聚合物单元(关于存在于测量值本身中)的一致性的信息损失。

[0043] 期望提供估算对齐映射的方法,其与这种现有的技术相比提供增加的准确度。

[0044] 根据本发明的第四方面,提供了估算以下各项之间的对齐映射的方法:(a)包含聚合物单元的聚合物的一系列测量值,其中测量值取决于k聚体,k聚体是聚合物的k个聚合物单元,其中k是整数,和(b)聚合物单元的参考序列。

[0045] 该方法使用参考模型,该参考模型处理测量值,作为对应于聚合物单元的参考序列的k聚体状态的参考系列的观察值,其中,参考模型包括:

[0046] 用于k聚体状态的参考系列中的k聚体状态之间的转换的转换加权;以及

[0047] 关于每个k聚体状态,当观察k聚体状态时用于观察的不同测量值的发射加权;以及

[0048] 方法包括应用参考模型至一系列测量值以衍生一系列测量值和对应于聚合物单元的参考序列的k聚体状态的参考系列之间的对齐映射的估算值。

[0049] 该方法因此关于参考序列使用参考模型。参考模型作为对应于聚合物单元的参考序列的参考系列的k聚体状态处理测量值,并包括用于参考系列的k聚体状态中的k聚体状态之间的转换的转换加权;以及关于每个k聚体状态,当观察k聚体状态时,用于观察的不同测量值的发射加权。它们可以是,但不限于HMM。结果,与以上讨论的涉及初始估算聚合物单元的序列,然后通过比较聚合物单元的一致性估算至聚合物单元的参考序列的对齐映射的已知技术相比,该方法可以改善对齐方法的评估准确度。这是由于以下原因。

[0050] 一般地说,参考模型的用途与W0-2013/041878中公开的估算聚合物的序列的模型类似,例如使用类似形式的转换加权和发射加权,以及对模型应用相同的数学处理。然而,参考模型本身不同于W0-2013/041878中公开的模型,W0-2013/041878中公开的模型是测量系统的类属模型,其中,每种k聚体状态一般而言可以具有k聚体状态的可能类型中的任一

种。因此,对于k聚体状态的可能类型之间的各种可能转换,关于一系列k聚体状态中的连续的k聚体状态之间的每种转换提供了转换加权。相反,用于本方法的参考模型是对应于聚合物单元的参考序列的参考系列的k聚体状态的模型。因此,提供了转换加权,用于参考系列的k聚体状态中的k聚体状态之间的转换。

[0051] 这种相似性是指本发明的方法可以利用W0-2013/041878中公开的模型的乘方。将关于聚合物单元的一致性的信息(存在于取决于重叠k聚体的测量值中)用于报告生成物。由于参考模型本身的不同本质,施加参考模型可以提供一系列测量值和对应于聚合物单元的参考序列的参考系列的k聚体状态之间的对齐映射,并且因此提供聚合物单元的一系列测量值和参考序列之间的对齐映射。

[0052] 在一些实施中,对于系列中的每个测量值,衍生的对齐映射的估算值可以包括参考系列的k聚体状态中的映射的k聚体状态的离散估算值。作为其中模型是HMM的实例,可以通过使用维特比算法以衍生对齐映射的估算值来实现它。

[0053] 在其他实施方式中,对于系列中的每个测量值,衍生的对齐映射的估算值可以包括关于参考系列的k聚体状态中的不同映射的k聚体状态的加权。作为其中模型是HMM的实例,可以通过使用前向-后向算法衍生对齐映射的估算值来实现它。

[0054] 可选地,方法可以进一步包括衍生得分(表示对齐映射的估算值正确的似然性)。该得分提供测量的聚合物和聚合物单元的参考序列之间的相似性的量度。通过提供与参考序列相比测量的聚合物的一致性的信息,其可用于各种各样的应用。

[0055] 在一些情况下,可以直接应用模型来衍生该得分。其一个实例是模型是HMM并应用维特比算法。

[0056] 在其他情况中,其中衍生的对齐映射的估算值可以包括,针对系列中的每个测量值的关于参考系列的k聚体状态中的不同映射的k聚体状态的加权,可以由那些加权本身衍生该得分。

[0057] 参考模型的来源可以根据应用进行改变。

[0058] 在一些应用中,可以预存储之前由聚合物单元的参考序列或由聚合物的参考序列采集的测量值产生的参考模型。

[0059] 在其他应用中,在进行方法时可以例如,如下生成参考模型。

[0060] 在第一实例中,可以由聚合物单元的参考序列生成参考模型。这可用于,例如由资料库或早期实验已知参考序列的应用。

[0061] 在这种情况下,参考模型的产生可以使用关于一组可能类型的k聚体状态的存储发射加权进行。有利地,这允许仅基于涉及用于可能类型的k聚体状态的发射加权的存储数据来产生用于聚合物单元的任何参考序列的参考模型。

[0062] 例如,可以通过包括以下各项的过程来产生参考模型:衍生对应于接收的聚合物的参考序列的一系列k聚体状态;以及通过产生衍生的k聚体状态系列中的k聚体状态之间的转换的转换加权,以及通过根据k聚体状态的类型由存储的发射加权选择用于衍生系列中的每种k聚体状态的发射加权来产生参考模型。

[0063] 在第二实例中,可以由包含聚合物单元的参考序列的聚合物的一系列参考测量值来产生参考模型。这可以用于,例如其中同时测量聚合物单元的参考序列与靶标聚合物的应用。特别地,在该实例中,不要求本身已知参考序列的聚合物单元的一致性。

[0064] 例如,参考模型可以通过使用另外的模型的方法来产生,该另外的模型处理一系列参考测量值作为不同可能类型的进一步系列的k聚体状态的观察值,其中,另外的模型包括:关于在进一步系列的k聚体状态中的连续的k聚体状态之间的每个转换,用于在可能类型的k聚体状态之间的可能的转换的转换加权;以及关于每种类型的k聚体状态,当k聚体状态是该类型时,用于观察的不同测量值的发射加权。这种另外的模型本身可以是W0-2013/041878中公开的模型类型。在这种情况下,可以通过包括以下各项的过程来产生参考模型:通过将另外的模型施加于一系列参考测量值产生k聚体状态的参考系列的估算值;以及通过产生产生的k聚体状态的参考系列的估算值中的k聚体状态之间的转换的转换加权和通过根据k聚体状态的类型由进一步模型的加权选择用于产生的参考系列的估算值中的每种k聚体状态的发射加权来产生参考模型。

[0065] 模型的产生可以是模型训练的较大框架中的一部分,模型训练检查由观察大量收集的k聚体状态系列衍生的大量收集的参考测量值以找到数学模型的未知参数,如发射和转换加权。典型地,当模型包括潜在的(隐藏的)变量时,可以将期望-最大化(expectation-maximisation,EM)算法用于寻找极大似然估算值。在HMM的特定情况下,可以使用Baum-Welch算法。这种算法是迭代的:进行初始猜测用于模型参数,通过检查一组训练测量值施加更新。将产生的HMM施加于第二截然不同的测量值组将产生改善的结果(假设可以通过与训练数据相同的模型描述第二组)。

[0066] 根据本发明的进一步方面,提供了能够实现根据本发明的第四方面的方法的电子计算机程序,或实现根据本发明的第四方面的方法的分析系统。

附图说明

[0067] 为了更好地理解,现通过非限制性实例参照附图来描述本发明的实施方式,其中:

[0068] 图1是生化分析系统的示意图;

[0069] 图2是系统的传感器设备的剖面图;

[0070] 图3是传感器设备的传感器元件的示意图;

[0071] 图4是通过测量系统随时间测量的事件信号的图;

[0072] 图5是在第一布置中的系统的电子电路的框图;

[0073] 图6是在第二布置中的系统的电子电路的框图;

[0074] 图7是控制生化分析系统来分析聚合物的方法的流程图;

[0075] 图8是状态检测步骤的流程图;

[0076] 图9是状态检测步骤的实例的详细流程图;

[0077] 图10是经历状态检测步骤的一系列原始测量和获得的一系列测量值的图;

[0078] 图11是控制生化分析系统来分析聚合物的替换方法的流程图;

[0079] 图12是控制生化分析系统以分类聚合物的方法的流程图;

[0080] 图13至图16是用于分析不同形式的参考数据的不同方法的流程图;

[0081] 图17是参考系列的k聚体状态的实例的状态图;

[0082] 图18是举例说明了k聚体状态之间的可能类型的转换的参考系列的k聚体状态的状态图;

[0083] 图19是用于产生参考模型的第一过程的流程图;

- [0084] 图20是用于产生参考模型的第二过程的流程图;以及
- [0085] 图21是估算对齐映射的方法的流程图;以及
- [0086] 图22是对齐映射的框图。

具体实施方式

- [0087] 在描述的实施方式中可以使用多种核苷酸和氨基酸序列。特别地:
- [0088] SEQ ID NO:1是核苷酸序列,其编码孔MS- (B1) 8 (=MS- (D90N/D91N/D93N/D118R/D134R/E139K) 8) ;
- [0089] SEQ ID NO:2是氨基酸序列,其编码孔MS- (B1) 8 (=MS- (D90N/D91N/D93N/D118R/D134R/E139K) 8) ;
- [0090] SEQ ID NO:3是核苷酸序列,其编码孔MS- (B2) 8 (=MS- (L88N/D90N/D91N/D93N/D118R/D134R/E139K) 8) ;
- [0091] SEQ ID NO:4是氨基酸序列,其编码孔MS- (B2) 8 (=MS- (L88N/D90N/D91N/D93N/D118R/D134R/E139K) 8)。除突变L88N之外,B2的氨基酸序列与B1的氨基酸序列相同;
- [0092] SEQ ID NO:5是用于野生型大肠杆菌核酸外切酶I (WT EcoExo I) 的序列,优选的多核苷酸处理酶(handling enzyme);
- [0093] SEQ ID NO:6是用于大肠杆菌核酸外切酶III的序列,优选多核苷酸处理酶;
- [0094] SEQ ID NO:7是用于嗜热菌RecJ的序列,优选多核苷酸处理酶;
- [0095] SEQ ID NO:8是用于λ噬菌体核酸外切酶的序列,优选多核苷酸处理酶;和
- [0096] SEQ ID NO:9是用于Phi29DNA聚合酶的序列,优选多核苷酸处理酶。
- [0097] 以下描述的各种特征是实例而不是限制性的。同样,描述的特征不必一起应用,并且可以以任何组合应用。
- [0098] 首先描述可以应用本发明的聚合物的性质(nature,本质)。
- [0099] 聚合物包含聚合物单元的序列。根据聚合物的性质,给定的每个聚合物单元可以是不同的类型(或种类(identity,一致性))。
- [0100] 聚合物可以是多核苷酸(或核酸)、多肽如蛋白、多糖、或任何其它聚合物。聚合物可以是天然的或合成的。聚合物单元可以是核苷酸。核苷酸可以是包含不同的核酸碱基的不同类型。
- [0101] 多核苷酸可以是本领域中已知的脱氧核糖核酸(DNA)、核糖核酸(RNA)、cDNA或合成核酸,如肽核酸(PNA)、甘油核酸(GNA)、苏糖核酸(TNA)、锁定的核酸(LNA)或具有核苷酸侧链的其他合成聚合物。多核苷酸可以是单链、双链、或者包含单链和双链区。通常,cDNA、RNA、GNA、TNA或LNA是单链。
- [0102] 核苷酸可以是以任何类型。核苷酸可以是天然存在的或人工的。核苷酸典型地包含核酸碱基(在本文中可以将其简称为“碱基”)、糖和至少一个磷酸酯基团。核酸碱基典型地是杂环的。合适的核酸碱基包括嘌呤和嘧啶以及更具体的腺嘌呤、鸟嘌呤、胸腺嘧啶、尿嘧啶和胞嘧啶。糖典型地是戊糖。合适的糖包括,但不限于核糖和脱氧核糖。核苷酸典型地是核糖核苷酸或脱氧核糖核苷酸。核苷酸典型地包含单磷酸酯、二磷酸酯或三磷酸酯。
- [0103] 核苷酸可以包括破损的碱基或表观遗传碱基。核苷酸可以被标记或修饰以充当具有明显信号的标记物。这种技术可以用来识别不存在的碱基,例如,在多核苷酸中的无碱基

单元或间隔。

[0104] 当考虑修饰或损伤DNA(或类似系统)的测量时,具有特别用途的是其中考虑互补数据的方法。提供的另外的信息使得在较大数目的基本状态之间进行区别。

[0105] 聚合物也可以是除多核苷酸之外的一类聚合物,其一些非限制性实例如下。

[0106] 聚合物可以是多肽,在这种情况下,聚合物单元可以是天然存在的或合成的氨基酸。

[0107] 聚合物可以是多糖,在这种情况下聚合物单元可以是单糖。

[0108] 特别地,在生化分析系统1包含纳米孔以及聚合物包含多核苷酸时,多核苷酸可以是长的,例如至少5kB(千碱基)即至少5,000核苷酸、或至少30kB(千碱基),即至少30,000核苷酸。

[0109] 在本文中,术语‘k聚体’是指k个聚合物单元的基团,其中k是正整数,包括k是1的情况,其中k聚体是单个聚合物单元。在一些情况下,提及k聚体(其中k是复数),该k聚体是k聚体的亚组,一般来说不包括k是1的情况。

[0110] 因此,对应于k聚体的每个聚合物单元的不同类型的不同组合,给定的每个k聚体也可以具有不同的类型。

[0111] 图1示出了用于分析聚合物的生化分析系统1,其也可以用于分类聚合物。转至图1,生化分析系统1包括连接至电子电路4的传感器设备2,该电子电路4进而连接至数据处理器6。

[0112] 将首先描述一些实例,其中传感器设备2包括各自包括生物纳米孔的传感器元件的阵列。

[0113] 以第一种形式,传感器设备2可以具有如图2中的截面所示的构造,其包括主体20,其中形成各自是具有布置其中的传感器电极22的凹部的凹槽(well,孔)21的阵列。设置大量的凹槽21来优化系统1的数据收集率。一般而言,可以存在任意数目的凹槽21,典型地256或1024个,但是在图2中仅示出了少数的凹槽21。主体20被盖23覆盖,盖23在主体20上延伸并且是中空的以限定每个凹槽21向其开放的样品室24。共用电极25设置在样品室24内。以该第一种形式,传感器设备2可以是在W0-2009/077734中进一步详细描述装置,可以将其教导应用于生化分析系统1,并且通过引用将其结合于此。

[0114] 以第二种形式,传感器设备2可以具有在W0-2014/064443中详细描述的构造,可以将其教导应用于生化分析系统1,并且通过引用将其结合于此。以该第二种形式,传感器设备2具有通常与第一种形式类似的构造,包括通常与凹槽21类似的隔室的阵列,但是它们具有更复杂的构造且其每个包括传感器电极22。

[0115] 为了有助于从收集室收集样品,可以将传感器设备布置为使得收集室21可以从下面的各个电极22拆下以暴露包含在其中的样品。在英国专利申请号1418512.8中更详细地描述了这种设备构造。

[0116] 将传感器设备2制备为形成传感器元件30的阵列,图3中示意性示出了其中的一个。通过形成在传感器设备2的第一种形式中横穿各个凹槽21或在传感器设备2的第二种形式中横穿各个隔室的膜31,然后将孔32嵌入到膜31中制造每个传感器元件30。膜31从样品室24密封各个凹槽21。膜31可以由两亲分子,如脂质制造。

[0117] 孔32是生物纳米孔。孔32以已知的方式连通样品室24和凹槽21。

[0118] 对于传感器设备2的第一种形式,使用W0-2009/077734中详细描述的技术和材料,或对于传感器设备2的第二种形式使用W0-2009/077734中详细描述的技术和材料可以进行这种制备。

[0119] 每个传感器元件30能够操作为在聚合物33移位穿过孔32期间使用关于每个传感器元件30的传感器电极22和共用电极25采集聚合物的电测量值。聚合物33移位穿过孔32产生可以观察到的并可以总体称为“事件”的测量特性的特性信号。

[0120] 在该实例中,孔是生物孔,其可以具有以下特性。

[0121] 生物孔可以是跨膜蛋白孔。用于本文所描述的方法的跨膜蛋白孔可以源自 β 桶状孔或 α 螺旋束孔。 β 桶状孔包括由 β 链形成的桶或通道。合适的 β 桶状孔包括,但不限于 α -毒素如 α -溶血素、炭疽毒素和杀白细胞素,以及细菌的外膜蛋白/孔蛋白如耻垢分枝杆菌孔蛋白(Msp)例如MspA、外膜孔蛋白F(OmpF)、外膜孔蛋白G(OmpG)、外膜磷脂酶A以及奈瑟球菌属(Neisseria)自转运脂蛋白(NalP)。 α -螺旋束孔包括由 α -螺旋形成的桶或通道。合适的 α -螺旋束孔包括,但不限于内膜蛋白和外膜蛋白,如WZA和ClyA毒素。跨膜孔可以源自Msp或源自 α -溶血素(α -HL)。

[0122] 合适的跨膜蛋白孔可以源自Msp,优选地源自MspA。这样的孔是低聚的并且通常包含源自Msp的7、8、9或10个单体。孔可以是源自包含相同单体的Msp的均-低聚体孔。可替换地,孔可以是源自Msp的杂-低聚体孔,其包含至少一种不同于其它单体的单体。该孔还可以包含一种或多种构建体,其包含源自Msp的两种或更多种共价连接单体。W0-2012/107778中描述了合适的孔。孔可以源自MspA或其同系物或旁系同源物(paralog)。

[0123] 生物孔可以是天然存在的孔或可以是突变体孔。在以下各项中描述了典型的孔: Stoddart D et al., Proc Natl Acad Sci, 12;106(19):7702-7, Stoddart D et al., Angew Chem Int Ed Engl. 2010;49(3):556-9, Stoddart D et al., Nano Lett. 2010 Sep 8;10(9):3633-7, Butler TZ et al., Proc Natl Acad Sci 2008;105(52):20647-52和W0-2012/107778。

[0124] 生物孔可以是MS-(B1)8。编码B1和B1的氨基酸序列的核苷酸序列是Seq ID:1和Seq ID:2。

[0125] 生物孔更优选是MS-(B2)8。除突变L88N之外,B2的氨基酸序列与B1的氨基酸序列相同。编码B2的核苷酸序列和B2的氨基酸序列是Seq ID:3和Seq ID:4。

[0126] 可以将生物孔嵌入到膜,如两亲性层,例如脂双层中。两亲性层是由具有亲水性和亲脂性的两亲分子如磷脂形成的层。两亲性层可以是单层或双层。两亲性层可以是如由(Gonzalez-Perez et al., Langmuir, 2009, 25, 10447-10450)或由作为W02014/064444公开的PCT/GB2013/052767公开的共-嵌段聚合物。可替换地,可以将生物孔插入到固态层中。

[0127] 孔32是纳米孔的一个实例。更通常地,传感器设备2可以具有包括至少一个传感器元件30的任何形式,该传感器元件能够操作为在聚合物移位穿过纳米孔期间,采集聚合物的测量值。

[0128] 纳米孔典型地是具有纳米级别的尺寸的孔,其允许聚合物通过。取决于移位穿过孔的聚合物单元的特性可以进行测量。特性可以与聚合物和纳米孔之间的相互作用相关。聚合物的相互作用可以发生在纳米孔的狭窄区(constricted region)。生化分析系统1测量特性,产生取决于聚合物的聚合物单元的测量值。

[0129] 可替代地,纳米孔可以是固态孔,其包含形成于固态层中的孔。在这种情况下,其可以具有以下特性。

[0130] 这种固态层典型地不具有生物来源。换言之,固态层一般不是源自或分离自生物环境,如生物体或细胞、或生物可利用结构的合成制造的形式。固态层可以由有机和无机材料两者形成,其包括,但不限于微电子材料,绝缘材料,如 Si_3N_4 、 Al_2O_3 和 SiO_2 ,有机和无机聚合物如聚酰胺、塑料,如Teflon®或弹性体,如双组分加成固化硅橡胶,以及玻璃。固态层可以由石墨烯形成。在W0-2009/035647和W0-2011/046706中公开了合适的石墨烯。

[0131] 当固态孔是固态层中的孔穴时,孔穴可以是化学或另外改性的以增强其作为纳米孔的特性。

[0132] 可以与另外的组件一起来使用固态孔,其中该另外的元件提供聚合物的替代的或附加的测量值,如隧道电极(Ivanov AP et al., Nano Lett. 2011 Jan 12; 11(1): 279-85),或者场效应晶体管(FET)装置(W0-2005/124888)。可以通过已知方法包括,例如在W0 00/79257中描述的那些来形成固态孔。

[0133] 在如图1所示的生化分析系统1的实例中,测量值是电测量值,特别是流动通过孔32的离子电流的电流测量值。一般而言,这些和其它电测量可以使用,如描述于Stoddart D et al., Proc Natl Acad Sci, 12; 106(19): 7702-7, Lieberman KR et al., J Am Chem Soc. 2010; 132(50): 17961-72和W0-2000/28312中的标准单通道记录装置来进行。可替代地,电测量可以使用,例如在W0-2009/077734和W0-2011/067559中所描述的多通道系统来进行。

[0134] 为了允许当聚合物移位穿过纳米孔32时采集测量值,可以通过聚合物结合部分来控制移位速率。典型地,借助于或针对施加场,该部分可以使聚合物移位穿过孔32。该部分可以是分子马达,其使用例如在该部分是酶的情况下的酶活性,或可以作为分子分子刹车(molecular brake)。在聚合物是多核苷酸的情况下,提出了多种方法来控制移位速率,包括使用多核苷酸结合酶。用于控制多核苷酸的移位速率的合适的酶包括,但不限于聚合酶、解旋酶、核酸外切酶、单链和双链结合蛋白和拓扑异构酶,如促旋酶。对于其它聚合物类型,可以使用与该聚合物类型相互作用的部分。聚合物相互作用部分可以是W0-2010/086603、W0-2012/107778和Lieberman KR et al., J Am Chem Soc. 2010; 132(50): 17961-72)中公开的任一种和针对电压门控方案(Luan B et al., Phys Rev Lett. 2010; 104(23): 238103)所公开的任一种。

[0135] 可以以多种方式使用聚合物结合部分以控制聚合物运动。利用或针对施加的场,该部分可以使聚合物移位穿过孔32。该部分可以用作分子马达,其使用,例如在该部分是酶的情况下的酶促活性,或作为分子刹车。可以通过控制聚合物通过孔的移位的分子棘轮(molecular ratchet)来控制聚合物的移位。分子棘轮可以是聚合物结合蛋白。

[0136] 对于多核苷酸,多核苷酸结合蛋白优选是多核苷酸处理酶。多核苷酸处理酶是一种多肽,其能够与多核苷酸相互作用并改性多核苷酸的至少一种特性。酶可以通过对其切割以形成单个核苷酸或核苷酸的较短链,如二或三核苷酸来改性多核苷酸。酶可以通过定向它或移位它到特定位置来改性多核苷酸。多核苷酸处理酶并不需要显示酶促活性,只要它能够结合靶多核苷酸并控制它通过孔的移位。例如,可以改性酶以除去它的酶活性,或者,可以在防止它作为酶的条件下进行使用。下文更详细地讨论了这样的条件。

[0137] 多核苷酸处理酶可以源自溶核酶(nucleolytic enzyme)。用于构建酶的多核苷酸处理酶更优选地源自酶分类(EC)组3.1.11、3.1.13、3.1.14、3.1.15、3.1.16、3.1.21、3.1.22、3.1.25、3.1.26、3.1.27、3.1.30和3.1.31中的任一种中的成员。酶可以是在W0-2010/086603中公开的那些中的任一种。

[0138] 优选的酶是聚合酶、核酸外切酶、解旋酶和拓扑异构酶,如促旋酶。合适的酶包括,但不限于来自大肠杆菌的核酸外切酶I(Seq ID:5)、来自大肠杆菌的核酸外切酶III酶(Seq ID:6)、来自嗜热菌(*T.thermophilus*)的RecJ(Seq ID:7)和噬菌体 λ 核酸外切酶(Seq ID:8)以及它们的变体。包含Seq ID:8中所示的序列或其变体的三个亚单元相互作用以形成三聚体核酸外切酶。酶优选地源自Phi29DNA聚合酶。源自Phi29聚合酶的酶包括在Seq ID:9或其变体中示出的序列。

[0139] Seq IDs:5、6、7、8或9的变体是这样的酶,其具有从Seq IDs:5、6、7、8或9中的氨基酸序列变化,并保持多核苷酸结合能力的氨基酸序列。该变体可以包括改性,其促进多核苷酸的结合和/或促进它的高盐浓度和/或室温下的活性。

[0140] 对于Seq IDs:5、6、7、8或9的氨基酸序列的整个长度,基于氨基酸一致性,变体将优选地至少50%与上述序列同源。更优选地,对于整个序列,基于氨基酸一致性,变体多肽可以是至少55%、至少60%、至少65%、至少70%、至少75%、至少80%、至少85%、至少90%以及更优选地至少95%、97%或99%与Seq IDs:5、6、7、8或9的氨基酸序列同源。对于200或更多、例如230、250、270或280或更多个相邻氨基酸的一段序列,可以存在至少80%、例如至少85%、90%或95%的氨基酸一致性(“硬同源性(hard homology)”)。如上所述确定同源性。变体可以以以上与参比SEQ ID NO:2讨论的任何方式,不同于野生型序列。如以上所讨论的,酶可以共价附接至孔。

[0141] 用于单链DNA测序的合适的策略是利用或针对施加电位,顺式至反式和反式至顺式地使DNA移位穿过孔32。用于链测序的最有利的机制是在施加电位下单链DNA通过孔32的受控移位。逐渐地或持续地作用于双链DNA的核酸外切酶,可以用于孔的顺侧以在施加电位下使剩余的单链进料通过或在反向电位下用于反侧。同样,还可以以类似的方式来使用解旋双链DNA的螺旋酶。对于要求针对施加的电位而链移位的测序应用也是可能的,但DNA必须在反向或没有电位下首先被酶“捕获”。在结合以后通过切换回来的电位,链将顺式至反式通过孔并通过电流而被保持在伸展构造。单链DNA核酸外切酶或单链DNA依赖性聚合酶可以作为分子马达以针对施加电位反式至顺式以受控分步方式,通过孔将最近移位的单链拉回。可替代地,单链DNA依赖性聚合酶可以充当分子刹车,其放慢多核苷酸通过孔的移位。可以将W0-2012/107778或W0-2012/033524中描述的任何部分、技术或酶用于控制聚合物运动。

[0142] 一般而言,当测量值是流动通过孔32的离子流(ion flow)的电流测量值时,离子流可以通常是DC离子流,但是原则上可替换地使用AC电流(即在施加AC电压下,流动的AC电流的幅度)。

[0143] 生化分析系统1可以通过上述的纳米孔采集除了离子流的电流测量值之外的类型的电测量值。

[0144] 其他可能的电子测量值包括:电流测量值、阻抗测量值、隧道效应测量值(例如如Ivanov AP et al., Nano Lett. 2011 Jan 12; 11(1): 279-85中所公开的)以及电场效应晶体

管(FET)测量值(例如,如在W02005/124888中所公开的)。

[0145] 作为电测量值的替代,生化分析系统1可以采集光学测量值。J. Am. Chem. Soc. 2009, 131 1652-1653公开了合适的光学方法,其涉及荧光的测量。

[0146] 测量系统8可以通过上述纳米孔,采集除了离子流的电流测量值之外的类型的电测量值。可能的电测量值包括:电流测量值、阻抗测量值、隧道效应测量值(例如,如在Ivanov AP et al., Nano Lett. 2011 Jan 12; 11(1): 279-85中所公开的)以及电场效应晶体管(FET)测量值(例如如在W02005/124888中所公开的)。

[0147] 光学测量可以与电测量结合(Soni GV et al., Rev Sci Instrum. 2010 Jan; 81(1): 014301)。

[0148] 生化分析系统1可以同时采集不同性质的测量值。测量值可以具有不同的性质,因为它们是不同的物理特性的测量值,其可以是上述的那些中的任一种。可替代地,测量值可以具有不同的性质,因为它们是在不同条件下的相同物理特性的测量值,例如在不同偏压下的电测量值如电流测量值。

[0149] 由多种类型的传感器设备2作为一系列原始测量值11输出的信号的典型形式是“嘈杂的阶梯波”,但是不限于这种信号类型。对于使用包含纳米孔的一类测量系统8得到的离子流测量值的情况,图4示出了具有这种形式的一系列原始测量值11的实例。

[0150] 典型地,由生化分析系统1采集的每个测量值取决于k聚体,k聚体是聚合物单元的各个序列的个k聚合物单元,其中k是正整数。尽管理想地,测量值将取决于单个聚合物单元(即其中k是1),但是对于多种典型类型的生化分析系统1,每个测量值取决于多个聚合物单元的k聚体(即其中k是复数)。即,每个测量值取决于k聚体中的每个聚合物单元的序列,其中k是复数。

[0151] 在通过生化分析系统1采集的一系列测量值中,多个测量值的连续组取决于相同的k聚体。每个组中的多个测量值具有恒定值,经历以下讨论的一些变化,并因此形成一系列原始测量值中的“水平”。这种水平典型地可以通过取决于相同的k聚体(或相同类型的连续的k聚体)的测量值形成,并且因此对应于生化分析系统1的常态。

[0152] 信号在一组水平(其可以是较大组)之间移动。鉴于仪表的采样速率和在信号上的噪声,可以认为,在水平之间的转换是瞬间的,因此可以通过理想化的步跟踪(step trace)来近似信号。

[0153] 对应于每个状态的测量值在事件的时间标度上是恒定的,但对于大多数类型的生化分析系统1而言,将在短时间范围内经历变化。变化可能起因于测量噪声,例如产生自电路和信号处理,尤其是在电生理学的特殊情况下产生自放大器。由于待测量的特性的幅度小,这种测量噪声是不可避免的。变化也可以来自,在生化分析系统1的基本物理或生物系统中的内在变化或扩展。大部分类型的生化分析系统1将经历这种到更大或更小的程度内在的变化。对于任何给定类型的生化分析系统1,两种变化来源可以起作用或这些噪声源中的一个可以是主要的。

[0154] 另外,典型地,不存在组中的测量值的数目的现有知识,其不可预测地变化。

[0155] 上述两种变化因素以及测量的数目的知识的缺乏可以使得难以区分一些组,例如在组较短和/或两个连续组的测量值的水平彼此接近的情况下。

[0156] 由于发生在生化分析系统1中的物理或生物过程,一系列的原始测量值可以采用

这种形式。因此,在某些情况下,每组测量值可以被称为“状态”。

[0157] 例如,在一些类型的生化分析系统1中,由聚合物通过孔32的移位组成的事件可以以棘轮方式发生。在棘轮移动的每个步骤期间,在通过跨越孔32的给定电压下,流过纳米孔的离子流是恒定的,并经历以上讨论的变化。因此,每组测量值与棘轮移动的步骤相关。每个步骤对应于一种状态,其中聚合物处于相对于纳米孔32的相应位置。虽然在状态期间,在精确位置方面可以存在一些变化,但在状态之间,存在聚合物的大规模的移位。根据生化分析系统1的性质,由于在纳米孔中的结合事件,所述状态可以发生。

[0158] 单个状态的持续时间可以取决于多种因素,如跨越孔施加的电位,用来对聚合物作棘轮的酶的类型,而不论聚合物通过存在的酶、pH、盐浓度和三磷酸核苷的类型被推动或拉动通过孔。状态的持续时间可以典型地在0.5ms至3s之间变化,这取决于生化分析系统1,以及对于任何给定纳米孔系统,具有在状态之间的一些随机变化。对于任何给定的生化分析系统1,可以实验性地确定持续时间的预期分布。

[0159] 可以实验性地检查给定的生化分析系统1提供测量值的程度,该测量值取决于k聚体和k聚体的尺寸。W0-2013/041878中公开了对于此的可能途径。

[0160] 返回到生化分析系统1,其可以通过上述的纳米孔采集除离子流的电流测量值之外的类型的电测量值。

[0161] 其他可能的电测量值包括:电流测量值、阻抗测量值、隧道效应测量值(例如,如在Ivanov AP et al., Nano Lett. 2011 Jan 12; 11(1): 279-85中所公开的)以及电场效应晶体管(FET)测量值(例如,如在W02005/124888中所公开的)。

[0162] 返回到图1,现在将讨论电子电路4的布置。电子电路4连接到关于每个传感器元件30的传感器电极22并连接到共用电极25。电子电路4可以具有如W0 2011/067559中所描述的整体布置。如下布置电子电路4以控制跨越每个传感器元件30的偏压的施加,并由每个传感器元件30采集测量值。

[0163] 在图5示出了用于电子电路4的第一布置,其示出了关于单个传感器元件30的组件,该组件对于每一个传感器元件30是复制的。在该第一布置中,电子电路4包括各自连接到传感器元件30的传感器电极22的检测通道40和偏压控制电路41。

[0164] 检测通道40由传感器电极22采集测量值。检测通道40布置为放大来自传感器电极22的电信号。因此将检测通道40设计为在足够的分辨率下,放大非常小的电流以检测由感兴趣的相互作用引起的特性变化。还检测通道40设计为具有足够高的带宽以提供检测每个这种相互作用所需的时间分辨率。这些限制条件要求灵敏的且因此昂贵的组件。具体地,检测通道40可以布置为,如在W0-2010/122293或W0 2011/067559(参考其中的每个,并通过引用将其结合于本文中)中所详细描述。

[0165] 偏压控制电路41向传感器电极22供应偏压,用于相对于检测通道40的输入来偏置传感器电极22。

[0166] 在正常操作期间,选择通过偏压控制电路41供应的偏压以使聚合物能够移位穿过孔32。这种偏压可以典型地高达-200mV的水平。

[0167] 还可以将由偏压控制电路41供应的偏压选择为使得其足以排出从孔32中的移位。通过引起偏压控制电路41来供应这种偏压,传感器元件30可操作为排出正在移位穿过孔32的聚合物。为了保证可靠的排出,偏压典型地是反向偏压,但不总是必须的。当施加这种这

种偏压时,将到达检测电路40的输入设计为保持在恒定的偏置电位,甚至当呈现负电流时(具有与正常电流类似的量值,典型地是-50pA值-100pA的量级)。

[0168] 用于图5中举例说明的电子电路4的第一布置,要求单独的检测通道40用于每个传感器元件30,这实现起来是昂贵的。图6示出了减少检测通道40的数目的用于电子电路4的第二布置。

[0169] 在该布置中,阵列中的传感器元件30的数目大于检测通道40的数目,并且生物化学传感系统可操作为通过以多路复用方式、特别是以电测量多路复用方式选择的传感器元件采集聚合物的测量值。这通过在传感器元件30的传感器电极23与检测通道40之间提供开关布置42实现。图6示出具有四个传感器单元(cell)30和两个检测通道40的简化的实例,但是传感器单元30和检测通道40的数目可以更大,典型地大更多。例如,对于一些施加,传感器设备2可以包括总共4096个传感器元件30和1024个检测通道40。

[0170] 开关布置42可以,如在W0-2010/122293中详细描述地进行布置。例如,开关布置42可以包括多个各自连接到一组的N个传感器元件30的1至N个多路复用器,并可以包括适当的硬件,如门锁以选择开关的状态。

[0171] 因此,通过切换开关布置42,可以使生化分析系统1操作为通过以电测量多路复用方式选择的传感器元件30采集聚合物的测量值。

[0172] 开关布置42可以,如在W0-2010/122293中所描述的方式进行控制,以便将检测通道40选择性连接到各个传感器元件30,该各个传感器元件30,基于由检测通道40输出的放大的电信号,具有可接收质量的性能,但是另外,如以下进一步描述地控制开关布置。

[0173] 如在第一布置中,该第二布置还包括关于每个传感器元件30的偏压控制电路41。

[0174] 尽管在该实例中,以电测量多路复用方式选择传感器元件30,但是可以将其他类型的生化分析系统1配置为,以空间多路复用方式在传感器元件之间切换,例如通过用于进行电测量的探针的移动,或者通过控制用于由不同的传感器元件30的不同空间位置采集光学测量值的光学系统。

[0175] 连接到电子电路4的数据处理器5布置如下。数据处理器5可以是运行适当的程序的计算机装置,该程序可以通过专用的硬件设备进行,或可以通过它们的任意组合进行。使用的计算机装置可以是任何类型的计算机系统,但典型地具有常规构造。可以用任何合适的编程语言来写计算机程序。可以将计算机程序存储于计算机可读存储介质中,上述计算机可读存储介质可以具有任何类型,例如:记录介质,其可插入计算系统的驱动器以及其可以以磁、光或光磁方式来存储信息;计算机系统的固定记录介质,如硬盘驱动器;或计算机存储器。数据处理器5可以包括插入到计算机,如台式机或膝上型计算机(laptop)的电路板。可以以常规的方式将数据处理器5使用的数据存储于它们的存储器10中。

[0176] 数据处理器5控制电子电路3的操作。与控制检测通道41的操作一样,数据处理器控制偏压控制电路41,并控制开关布置31的切换。数据处理器5还接收并处理来自每个检测通道40的一系列测量值。如以下进一步描述的,数据处理器5存储和分析一系列的测量值。

[0177] 数据处理器5控制偏压控制电路41以施加足以使聚合物移位穿过传感器元件30的孔32的偏压。生化传感器元件41的这种操作使得由不同的传感器元件30收集一系列测量值,其可以通过数据处理器5或通过另一数据处理单元分析,来估算聚合物中的聚合物单元的序列,例如使用,如在W0-2013/041878中所描述的技术。可以收集并组合来自不同的传感

器元件30的数据。

[0178] 数据处理器5接收并分析通过传感器设备2采集并由电子电路4供应的一系列测量值11。数据处理器5还可以向电子电路5提供控制信号,例如来选择横跨传感器设备2中的生物孔1施加的电压。可以在任何合适的连接、例如在数据处理器5和传感器设备2物理定位在一起的情况下的直接连接、或在数据处理器5和传感器设备2物理彼此远离的情况下的任何类型的网络连接施加一系列的原始测量值11。

[0179] 现在将描述图7所示的控制生化分析系统1来分析聚合物的方法。该方法是根据本发明的第一方面并以通过排斥不需要进一步分析的聚合物增加分析速度的方式进行。该方法在数据处理器5中实现。关于采集一系列测量值的每个传感器元件30、即用于电子电路4的第一布置中的每一个传感器元件30、以及在用于电子电路4的第二布置中通过开关布置42连接到检测通道40的每个传感器元件30平行实施该方法。

[0180] 在步骤C1中,通过控制偏压控制电路30来施加横跨传感器元件30的足以使聚合物能够移位的孔32的偏压操作生化分析系统1。基于来自检测通道40的输出信号,检测移位并开始采集测量值。随着时间采集一系列测量值。

[0181] 在一些情况下,以下步骤在通过传感器设备2采集的一系列原始测量值11、即上述类型的一系列测量值上操作,上述类型的一系列测量值包括在没有预先知道任何基团的测量数的情况下取决于相同k聚体的多个测量值的连续组。

[0182] 在其它情况下,如图8所示,使用状态检测步骤SD预处理原始测量值11以衍生代替原始测量值用于以下步骤的一系列测量值12。

[0183] 在这种状态检测步骤SD中,处理一系列原始测量值11以识别原始测量值的连续组并衍生一系列测量值12,其由关于每个识别组的预定数目的测量值组成。因此,关于测量的聚合物单元的每个序列衍生一系列测量值12。状态检测步骤SD的目的是将一系列原始测量值减少到与每个k聚体有关的预定数目的测量值来简化随后的分析。例如噪声阶梯波信号,如图4所示,可以被减少到其中与每个状态有关的单个测量值可以是平均电流的状态。这种状态可以被称为水平。

[0184] 图9示出了如下寻找一系列原始测量值11的衍生物中的短期升高的这种状态检测步骤SD的实例。

[0185] 在步骤SD-1中,差分一系列原始测量值11以衍生其衍生物。

[0186] 在步骤SD-2中,使步骤SD-1的衍生物经受低通过滤来抑制高频噪声,步骤SD-1中的差异趋于放大。

[0187] 在步骤SD-3中,来自步骤SD-2的过滤衍生物被阈值化以检测在测量值的组之间的转换点,从而识别原始测量值的组。

[0188] 在步骤SD-4中,由在步骤SD-3中识别的原始测量值的每个组衍生预定数目的测量值。从步骤SD-4输出的测量值形成一系列测量值12。

[0189] 测量值的预定数目可以是一个或多个。

[0190] 以最简单的途径,由原始测量值的每个组衍生单个测量值,例如每个确定组中的原始测量值的平均值、中值、标准偏差或数目。

[0191] 以其他途径,由每个组衍生不同性质的测量值的预定复数,例如每个确定组中的原始测量值的平均值、中值、标准偏差或数目中的任意两个或更多个。在该情况下,根据相

同k聚体采集不同性质的测量值的预定复数,因为它们都是原始测量值的相同组中的不同量度。

[0192] 状态检测步骤SD可以使用图9所示的那些中的不同的方法。例如,图9所示的方法的常用简化是使用滑动窗口分析,其比较数据的两个相邻窗口的平均值。然后可以基于平均差异来直接设定阈值,或可以基于在两个窗口中数据点的方差(例如,通过计算史蒂特氏t统计量(Student's t-statistic))来设定阈值。这些方法的独特优势在于可以施加它们,而没有施加关于数据的多种假设。

[0193] 可以存储与测量水平相关的其它信息用于以后的分析。这样的信息可以包括但不限于:信号的变化;不对称信息;观察的置信度;组的长度。

[0194] 举例来说,图10举例说明通过实验确定的通过移动窗口t测试减少的一系列原始测量值11。特别地,图10示出了作为细线的一系列原始测量值11。示出在状态检测以后的水平作为暗线重叠。

[0195] 在聚合物部分移位穿过纳米孔时,即在移位期间执行步骤C2。这时,收集在部分移位期间由聚合物采集的一系列测量值用于分析,在本文中将其成为测量值的“组块(chunk)”。在采集了预定数目的测量值使得测量值的组块具有一定大小之后可以进行步骤C2,或可以可替代地在预定量的时间之后进行步骤C2。在前一种情况下,可以通过在运行开始时初始化的参数限定测量值的组块的大小,但是动态地改变使得测量值的组块的大小改变。

[0196] 在步骤C3中,分析在步骤C2中收集的测量值的组块。该分析使用参考数据50。如以下更详细地讨论的,由聚合物单元的至少一个参考序列衍生参考数据50。步骤C3中进行的分析提供(a)已经采集了测量值的部分移位的聚合物的聚合物单元的序列和(b)一种参考序列之间的相似性的测量。用于进行该分析的多种技术是可以的,以下描述了其一些实例。

[0197] 相似性的量度可以表示与参考序列的整体或与参考序列的一部分的相似性,这取决于施加。可以相应地选择在衍生相似性的量度的步骤C3中施加的技术,例如整体或局部方法。

[0198] 另外,相似性的量度可以表示多种不同量度的相似性,条件是其概括地提供序列是如何类似的量度。以下阐述了可以以不同的方式由序列确定的相似性的具体量度的一些实例。

[0199] 在步骤C4中,响应于步骤C3中确定的相似性的量度作出判定(a)排斥测量的聚合物,(b)需要另外的测量值来做判定,或(c)继续采集测量值直到聚合物最后。

[0200] 如果在步骤C4中作出的判定是(a)将测量的聚合物排斥,那么方法进行至步骤C5,其中,控制生化分析系统1以排斥聚合物使得可以由另外的聚合物采集测量值。

[0201] 如下,在电子电路4的第一和第二布置之间不同地进行步骤C5。

[0202] 在电子电路4的第一布置的情况下,然后在步骤C5中,控制偏压控制电路30来施加横跨传感器元件30的孔32的足以排出当前移位的聚合物的偏压。其排出聚合物并从而使孔32可用于接收另外的聚合物。在步骤C5中的这种排出之后,方法返回到步骤C1,所以控制偏压控制电路30来施加横跨传感器元件30的孔32的足以使另外的聚合物移位穿过孔32的偏压。

[0203] 在电子电路4的第二布置的情况下,然后在步骤C5中,通过控制开关布置42来断开

当前连接到传感器元件30的检测通道40以及将检测通道40选择性连接到不同的传感器元件30引起分析生化分析系统1停止由当前选择的传感器元件30采集测量值。同时,在步骤C5中,控制偏压控制电路30来施加横跨传感器元件30的孔32的足以排出当前移位穿过当前选择的传感器元件30的聚合物的偏压,使得传感器元件30可用于接收未来的进一步聚合物。

[0204] 方法然后返回到步骤C1,将其施加于新选择的传感器元件30使得生化分析系统1开始由其采集测量值。

[0205] 如果在步骤C4中作出的判定是(b)需要另外的测量值来作出判定,那么方法回到步骤C2。因此,继续采集移位聚合物的测量值直到接下来在步骤C2中控制测量值的组块和在步骤C3中分析。当再次进行步骤C2时,收集的测量值的组块可以仅仅是隔离分析的新的测量值,或者可以是与之前的测量值组块结合的新的测量值。

[0206] 如果在步骤C4中作出的判定是(c)继续采集测量值直到聚合物的最后,那么在不重复步骤C2和C3的情况下方法进行至步骤C6使得不分析数据的另外的组块。在步骤C6中,传感器元件1继续操作使得继续采集测量值直到聚合物的最后。此后,方法回到步骤C1,使得可以分析另外的聚合物。

[0207] 如通过相似性的量度表示的,相似性的程度,即用作步骤C4中的判定的基础可以根据施加和参考序列的性质而改变。因此,如果判定是响应于相似性的量度,那么一般而言对用于作出不同判定的相似性的程度没有限制。

[0208] 关于相似性的量度的依赖性可能如何改变的一些实例如下。

[0209] 在其中聚合物单元的参考序列是不期望的序列,并且在步骤C4中响应于指出部分移位的聚合物是不期望的序列的相似性的量度作出排斥聚合物的判定的施加中,可以将相对高程度的相似性用作排斥聚合物的基础。类似地,在施加的背景下,相似度可以根据参考序列的性质改变。在旨在区分类似序列时,可以要求将较高的相似度用作用于排斥的基础。

[0210] 相反,在衍生参考数据50的聚合物单元的参考序列是靶标并在在步骤C4中响应于表示部分移位的聚合物不是靶标的相似性的量度作出排斥聚合物的判定的施加中,可以将相对低的相似度用作排斥聚合物的基础。

[0211] 作为另一实例,如果施加是确定来自已知细菌的已知基因是否存在于多种细菌的样品中,则如果基因具有横跨不同菌株的保守序列,确定多核苷酸是否具有与靶标相同的序列所需的相似度将高于如果序列不是保守的。

[0212] 类似地,在本发明的一些实施方式中,相似性的量度将等同于聚合物与靶标聚合物的一致性程度,而在其他实施方式中,相似性的量度将等同于聚合物与靶标聚合物相同的概率。

[0213] 作为用于排斥的基础所需的相似度还可以根据可能的时间节省而改变,可能的时间节省本身取决于以下所描述的施加。可接收的假阳性率可以取决于时间节省。例如,在排斥不期望的聚合物的可能的时间节省相对高时,排斥作为靶标的增加比例的聚合物是可接收的,条件是存在排斥不期望的聚合物的总体时间节省。

[0214] 现在回到图7的方法,如果在采集聚合物的测量值期间的任何点,检测到不再采集测量值,表示已经达到聚合物的末端,那么方法立即回到步骤C1,使得可以分析另外的聚合物。在如此采集整个聚合物的测量值之后,可以如W0-2013/041878中公开的分析那些测量值例如以衍生聚合物单元的序列的估算值。

[0215] 参考数据50的来源可以根据施加改变。可以由聚合物单元的参考序列或由从聚合物单元的参考序列采集的测量值产生参考数据50。

[0216] 在一些施加中,可以预存储之前产生的参考数据50。在其他施加中,在进行方法时产生参考数据50。

[0217] 可以关于聚合物单元的单个参考序列或聚合物单元的多个参考序列提供参考数据50。在后一种情况下,关于每个序列进行任一步骤C3或另外将多个参考序列中的一个选择用于步骤C3。在后一种情况下,可以基于多种标准,根据施加作出选择。例如,参考数据50可以施加于不同类型的生化分析系统1(例如不同的纳米孔)和/或外界条件,在这样的情况下基于实际使用的生化分析系统1的类型和/或实际的外界条件选择以下所描述的参考模型70。

[0218] 图7所示的方法可以根据施加改变。例如,在一些变体中,步骤C4中的判定绝不会是(c)继续采集测量值直到聚合物的末端,使得方法方法重复收集和分析测量值的组块直到聚合物的末端。

[0219] 在另一变体中,在步骤C3中,代替使用参考数据50和确定相似性的量度,在步骤C4中排斥聚合物的判定可以是基于一系列测量值的其他分析,一般而言是基于测量值的组块的任何分析。

[0220] 在一种可能中,步骤C3可以分析测量值的组块是否质量不够,例如具有超过阈值的噪声水平、具有错误的比例,或聚合物被损坏的特性。

[0221] 基于该分析作出步骤C4中的判定,从而基于内部质量控制检查排斥聚合物。这仍涉及基于测量值的组块、即在部分移位期间由聚合物采集的一系列测量值作出排斥聚合物的判定,所以与引起阻断的排出聚合物相反,在排斥聚合物的情况下,聚合物不再移位,所以没有采集依赖k聚体的测量值。

[0222] 在另一可能性中,其中方法是根据本发明的第二方面,如图11所示修改方法。该方法与图7的方法相同,除了修改了步骤C3。在步骤C3中,代替使用由聚合物单元的至少一个参考序列衍生的参考数据50和确定相似性的量度,使用作为不同可能类型的一系列k聚体状态的观察值处理测量值,并包括以下的一般模型60:转换加权61,关于在一系列k聚体状态中的连续的k聚体状态之间的每个转换,用于在可能类型的k聚体状态之间的可能转换;以及发射加权62,关于每种类型的k聚体状态,其表示观察给定k聚体的测量值的机率。修改步骤C3以包括拟合参考模型60的量度。

[0223] 一般模型60可以是W0-2013/041878中描述的类型。模型的细节参考W0-2013/041878。参考图13,以下进一步描述了一般模型60。衍生拟合的量度,例如作为由k聚体状态的最类似序列观察到的测量值的似然性。这种拟合的量度表示测量值的质量。

[0224] 当以这种方式修改步骤C3时,基于所述拟合的量度作出步骤C4中的判定,从而基于内部质量控制检查排斥聚合物。

[0225] 因此,如果与聚合物单元的参考序列的相似性表示不需要进一步分析聚合物或如果由聚合物采集的测量值具有通过模型确定的不良的质量使得不批准另外的移位和测量,则该方法引起聚合物被排斥。模型表示数据不足够良好的程度取决于模型本身的复杂性。例如,更复杂的模型可以具有可以解决一些可引起排斥的条件的参数。

[0226] 可以引起排斥的条件可以包括例如:流进不可接收的信号;高噪声;非模型行为;

不规则的系统误差如温度波动;和/或由于电-物理系统的误差。

[0227] 例如,一种可能是聚合物或其他碎片已经容纳于纳米孔中,产生缓慢变化的、相当静态的电流。模型通常期望数据上良好分离的(时间上分段地恒定)步骤,所以这种测量将具有不良的拟合至模型的量度。

[0228] 第二种可能是瞬态噪声,例如另外紧密的组的步骤之间的电流的大变化。如果这种噪声高频出现,数据可能对于实际的目的具有很少的用途。由于不期望的测量值的频率高,拟合至模型的量度将是低的。

[0229] 这些“误差”可以以非瞬时的方式出现。确实,经常观察到关于临近部分,测量部分在它们的平均电流上出现偏移。对此的可能解释是,孔的形态和聚合物分子的变化。无论原因如何,这种行为没有被模型捕获,所以对于实际目的来说数据具有很少的用途。

[0230] 通过增加模型的复杂性可以将这种误差的影响减轻到某种程度。然而,这并非是人令人期望的,并可以导致模型化数据和解码聚合物序列的计算费用升高。

[0231] 由于排斥了这种聚合物链,仅衍生模型转换加权和发射加权的具有强同源性的那些聚合物序列给出具有良好的拟合至模型的量度的测量值。

[0232] 在完成采集整个聚合物的测量值之后,可以如WO-2013/041878中公开的分析那些测量值例如以衍生聚合物单元的序列的估算值。

[0233] 可以独立或组合施加图7和11的可替换方法,在这样的情况下,可以同时(例如平行进行两种方法的步骤C3,以及共同进行其他步骤)或连续(例如在图7的方法之前进行图11的方法)施加它们。

[0234] 现在将描述图12所示的控制生化分析系统1来分类聚合物的方法。该方法是根据本发明的第三方面。在这种情况下,样品室24包含含有可以是不同类型的聚合物的样品,以及凹槽21充当收集分类的聚合物的收集室。

[0235] 该方法在数据处理器5中实现。关于平行的多个传感器元件30、例如用于电子电路4的第一布置中的每一个传感器元件30、以及在用于电子电路4的第二布置中通过开关布置42连接到检测通道40的每个传感器元件30平行进行该方法。

[0236] 在步骤D1中,通过控制偏压控制电路30来施加横跨传感器元件30的孔32的足以使聚合物能够移位的偏压操作生化分析系统1。这引起聚合物开始移位穿过纳米孔并在移位期间进行以下步骤。基于来自检测通道40的输出信号,检测移位并开始采集测量值。随时间由传感器元件30采集聚合物的一系列测量值。

[0237] 在一些情况下,以下步骤在通过传感器设备2采集的一系列原始测量值11、即上述类型的一系列测量值上操作,上述类型的一系列测量值包括在没有预先知道任何组的测量值数的情况下,取决于相同k聚体的多个测量值的连续组。

[0238] 在其它情况下,使用状态检测步骤SD预处理原始测量值11以衍生代替原始测量值,用于以下步骤的一系列测量值12。可以以与参考图8和图9,在以上描述的步骤C1相同的方式进行状态检测状态SD。

[0239] 在聚合物部分移位穿过纳米孔时,即在移位期间进行步骤D2。这时,收集在部分移位期间,由聚合物采集的一系列测量值用于分析,在本文中将其称为测量值的“组块(chunk)”。在采集了预定数目的测量值使得测量值的组块具有一定大小之后可以进行步骤D2,或可以可替代地在预定量的时间之后进行步骤D2。在前一种情况下,可以通过在运行开

始时初始化的参数限定测量值的组块的大小,但是动态地改变使得测量值的组块的大小改变。

[0240] 在步骤D3中,分析在步骤D2中收集的测量值的组块。该分析使用参考数据50。如以下更详细地讨论的,由聚合物单元的至少一种参考序列衍生参考数据50。步骤D3中进行的分析提供 (a) 已经采集了测量值的部分移位的聚合物的聚合物单元的序列和 (b) 一种参考序列之间的相似性的量度。用于进行该分析的多种技术是可以的,以下描述了其一些实例。

[0241] 相似性的量度可以表示与参考序列的整体或与参考序列的一部分的相似性,这取决于施加。可以相应地选择在衍生相似性的量度的步骤D3中施加的技术,例如整体或局部方法。

[0242] 另外,相似性的量度可以表示多种不同量度的相似性,条件是其概括地提供序列是如何类似的量度。以下阐述了可以以不同的方式由序列确定的相似性的具体量度的一些实例。

[0243] 在步骤D4中,根据步骤D3中确定的相似性的量度作出任一判定,(a) 需要另外的测量值来作出判定,(b) 完成聚合物到凹槽21中的移位,或(c) 将测量的聚合物排出返回到样品室24。如果在步骤D4中作出的判定是(a) 需要另外的测量值来作出判定,那么方法回到步骤D2。因此,继续采集移位聚合物的测量值直到接下来在步骤D2中收集测量值的组块并在步骤D3中分析。当再次进行步骤D2时收集的测量值的组块可以仅仅是隔离分析的新的测量值或可以是与之前的测量值组块结合的新的测量值。

[0244] 如果在步骤D4中作出的判定是(b) 完成聚合物到凹槽21中的移位,那么在不重复步骤D2和D3的情况下方法进行至步骤D6,使得不进行测量值的进一步分析。

[0245] 在步骤D6中,完成聚合物到凹槽21中的移位。结果,将聚合物收集到凹槽21中。

[0246] 可以通过横跨传感器元件30的孔32施加使聚合物能够移位的相同偏压进行步骤D6。

[0247] 可替代地,在步骤D6中,可以改变偏压来以增加的速率进行聚合物的剩余移位来减少移位花费的时间。这是有利的,因为其增加分类过程的总体速率。增加移位速率是可接收的,因为不再需要分析聚合物。典型地,偏压的变化可以是升高的。在典型的系统中,增加可以是显著的。例如,在一个实施方式中,移位速度可以从约30碱基/秒升高到约10,000碱基/秒。改变移位速度的可能性可以取决于传感器元件的配置。例如,在将聚合物结合部分例如酶用于控制移位时,这可以取决于使用的聚合物结合部分。有利地,可以选自可以控制速率的聚合物结合部分。

[0248] 在步骤D6期间,可以继续操作传感器元件1使得继续采集测量值直到聚合物的末端,但是这是可选的,因为不需要确定剩余的序列。

[0249] 在步骤D6之后,方法回到步骤D1,使得可以移位另外的聚合物。

[0250] 如果在步骤D4中作出的判定是(c) 排出聚合物,那么方法进行至步骤D5,其中控制生化分析系统1来将测量的聚合物排出回到样品室24,使得可以采集另外的聚合物的测量值。

[0251] 在步骤D5中,控制偏压控制电路30来横跨传感器元件30的孔32应用足以排出当前移位的聚合物的偏压。其排出聚合物并从而使孔32可用于接收另外的聚合物。在步骤D5中的这种排出之后,方法返回到步骤D1,所以控制偏压控制电路30来应用横跨传感器元件30

的孔32的足以使另外的聚合物移位穿过孔32的偏压。

[0252] 在回到步骤D1时,方法重复。方法的重复性能引起聚合物从样品室24连续移位和处理。

[0253] 因此,方法利用通过分析由在部分移位期间采集自聚合物的一系列测量值提供的相似性的量度,作为是否将连续的聚合物收集在凹槽21中的基础。以这种方式,分类样品室24中的样品的聚合物,且将期望的聚合物选择性收集在凹槽21中。

[0254] 可以回收收集的聚合物。通过从样品室24除去样品,然后回收凹槽21的聚合物,其可以在重复运行方法之后进行。可替代地,例如通过提供具有从凹槽21提取聚合物的流体系统的生化分析系统1,其可以在样品的聚合物的移位期间进行。

[0255] 该方法可以应用于各种各样的应用。例如,该方法可以应用于是多核苷酸、例如病毒基因组或质粒的聚合物。病毒基因组典型地具有10-15kB(千碱基)级别的长度以及质粒典型地具有4kB级别的长度。在这种实施例中,不需要将多核苷酸片段化并可以整个收集。可以以任何方式使用收集的病毒基因组或质粒例如来转染细胞。转染是将DNA引入到细胞核中的过程,并且是重要工具用于探究基因功能和基因表达的调节的研究,从而促进基础细胞科研、药物新发现和靶标验证的前进。还可以转染RNA和蛋白质。

[0256] 如通过相似性的量度指出的,相似度,即用作步骤D4中的判定的基础可以根据应用和参考序列的性质而改变。因此,如果判定是依赖相似性的量度,则一般而言,对用于作出不同判定的相似度没有限制。

[0257] 关于相似性的量度的依赖性可能如何改变的一些实例如下。

[0258] 在多种应用中,衍生参考数据50的聚合物单元的参考序列是期望的序列。在那种情况下,在步骤D4中,作出完成移位的判定来响应表示部分移位的聚合物是期望的序列的相似性的量度,可以将相对高的相似度用作完成移位的基础。

[0259] 然而,这不是必需的。在一些应用中,聚合物单元的参考序列是不期望的序列。在那种情况下,在步骤D4中,作出完成移位的判定来响应表示部分移位的聚合物不是不期望的序列的相似性的量度。

[0260] 类似地,在应用的背景下,相似度可以根据参考序列的性质改变。在旨在区分类似序列时,可以要求将较高的相似度作用于排斥的基础。

[0261] 关于每个传感器元件30在步骤D4中使用相同的参考数据50和相同的标准进行该方法。在那种情况下,每个凹槽21平行收集相同的聚合物。

[0262] 可替代地,可以进行方法来将不同的聚合物收集到不同的凹槽21中。在这种情况下,进行差分分类。在其一个实例中,将不同的参考数据50用于不同的传感器元件30。在其另一个实例中,将相同的参考数据50用于不同的传感器元件30,但是在对不同的传感器元件的相似性量度的不同依赖性下进行步骤D4。

[0263] 根据应用可以改变图7、11和12所示的方法。

[0264] 根据应用可以使用聚合物单元的各种不同类型的参考序列。在没有限制的情况下,在聚合物是多核苷酸时,聚合物单元的参考序列可以包含比较其测量值的一种或多种参考基因组或,一种或多种基因组的感兴趣区域。

[0265] 参考数据50的来源可以根据应用改变。可以由聚合物单元的参考序列或由从聚合物单元的参考序列采集的测量值产生参考数据。

[0266] 在一些应用中,可以预存储之前产生的参考数据50。在其他应用中,在进行方法时产生参考数据50。

[0267] 可以关于聚合物单元的单个参考序列或聚合物单元的多个参考序列提供参考数据50。在后一种情况下,关于每个序列进行任一步骤D3,或者另外地将多个参考序列中的一个选择用于步骤D3。在后一种情况下,可以基于多种标准,根据应用作出选择。例如,参考数据50可以应用于不同类型的生化分析系统1(例如不同的纳米孔)和/或外界条件,在这样的情况下,基于实际使用的生化分析系统1的类型和/或实际的外界条件选择以下所描述的参考模型70。

[0268] 上述的生化分析系统1是包括每个包含纳米孔的传感器元件阵列的生化分析系统的一个实例。然而,通常可以将上述方法应用于可操作为可能不使用纳米孔的采集聚合物的连续测量值的任何生化分析系统。

[0269] 不包含纳米孔的这种生化分析系统的一个实例是扫描探针显微镜,其可以是原子力显微镜(AFM)、扫描隧道显微镜(STM)或扫描显微镜的另一形式。在这种情况下,生化分析系统可以可操作为采集以空间多路复用方式选择的聚合物的连续测量值。例如,可以将聚合物设置在不同空间位置的基板上,并可以通过扫描探针显微镜的移动提供空间多路复用。

[0270] 在读数是AFM的情况下,相比于单个聚合物单元的尺寸,AFM尖端的分辨率可以是较不精细的。因此,测量值可以是多个聚合物单元的函数。AFM尖端可以被功能化,从而以替代方式与聚合物单元相互作用至如同它未被功能化。可以以接触模式、非接触模式、轻敲模式或任何其它模式来操作AFM。

[0271] 在读数是STM的情况下,相比于单个聚合物单元的尺寸,测量值的分辨率可以是较不精细的,使得测量值是多个聚合物单元的函数。可以常规地或以任何其它模式操作STM或进行光谱测量(STS)。

[0272] 现在将讨论用于上述任何方法的参考数据50的形式。参考数据50可以采取以不同方式由聚合物单元的参考序列衍生的多种形式。在步骤C4或D4中进行的提供相似性的量度的分析取决于参考数据50的形式。现在将描述一些非限制性的实例。

[0273] 在第一实例中,参考数据50表示至少一种参考序列的聚合物单元的一致性。在那种情况下,步骤C4或D4包括如下图13所示的过程。

[0274] 在步骤C4a-1中,分析测量值的组块63以提供部分移位的聚合物的聚合物单元的序列的聚合物单元的一致性的估算值64。一般而言,可以使用用于分析由生化分析系统采集的测量值的任何方法进行步骤C4a-1。

[0275] 可以特别使用在W0-2013/041878中详细描述的方法进行步骤C4a-1,通过引用将其结合于本文中。参考W0-2013/041878中的方法的细节,但是如下给出概要。

[0276] 该方法参考包括关于一系列对应于测量值组块63的k聚体状态的转换加权61和发射加权62的一般模型60。

[0277] 关于一系列k聚体状态中的连续k聚体状态之间的每种转换提供转换加权61。可以将每种转换视为从起点k聚体状态至终点k聚体状态。转换加权61表示可能类型的k聚体状态之间的可能转换的相对加权,其是从任何类型的起点k聚体状态至任何类型的终点k聚体状态。一般而言,这包括用于相同类型的两个k聚体状态之间的转换的加权。

[0278] 关于每种类型的k聚体状态提供发射加权62。发射加权62是当k聚体状态是该类型时用于观察的不同测量值的加权。概念上,可以认为发射加权62表示观察该k聚体状态的测量值的给定值的几率,但是它们不需要是概率。

[0279] 概念上,可以认为转换加权61表示可能转换的几率,虽然它们不需要是概率(probabilities)。因此,转换加权61考虑在测量值取决于其不同k聚体状态之间转换的k聚体状态的几率,其可以或多或少可能取决于起点和终点k聚体状态的类型。

[0280] 通过举例而非限制性地,模型可以是HMM,其中转换加权61和发射加权62是概率(probabilities)。

[0281] 步骤C4a-1使用参考模型60来衍生部分移位的聚合物的聚合物单元的序列的聚合物单元的一致性的估算值64。这可以使用可应用于参考模型60的性质的已知技术进行。典型地,这种技术基于通过参考模型60预测的测量值的似然性衍生由k聚体状态的序列观察的估算值64。如W0-2013/041878中所描述的,这种技术可以在一系列原始测量值11或一系列测量值12上进行。

[0282] 这种方法还可以提供测量值与模型的拟合的量度,例如表示由k聚体状态的最可能序列观察的参考模型60预测的测量值的似然性的质量得分。典型地衍生这种量度,因为它们用于衍生估算值64。

[0283] 作为实施例,在一般模型是HMM的情况下,分析技术可以用于解决HMM的已知的算法,例如在本领域熟知的维特比算法。在那种情况下,基于通过一般模型预测的通过k聚体状态的总序列产生的似然性来衍生估算值64。

[0284] 作为另一实施例,在一般模型60是HMM的情况下,分析技术可以是2005年1月4日提交的,归档在Cornell大学中的Casadio大学的生物系的Fariselli et al.,“The posterior-Viterbi:a new decoding algorithm for hidden Markov models”中公开的类型。在该方法中,得到后矩阵(表示由每种k聚体状态观察到的测量值的概率)和一致路径(其中相邻k聚体状态偏向重叠的路径),而不是简单地选择每个事件的最可能的k聚体。实质上,这使得恢复由Viterbi算法的应用直接到的相同的信息。

[0285] 给出的以上描述是依据一般模型60,其是HMM,其中转换加权61和发射加权62是概率,以及方法使用指的是一般模型60的概率技术。然而,可替代地可能的是一般模型60使用一种框架,其中转换加权61和/或发射加权62不是概率,但以某种其它方式表示转换或测量的几率。在这种情况下,方法可以使用分析技术而不是概率技术,其是基于由聚合物单元的序列产生的一系列测量值的一般模型60所预测的似然性。分析技术可以明确使用似然函数,但是一般而言这不是必需的。

[0286] 在步骤C4a-2中,将估算值64与参考数据50比较以提供相似性的量度65。这种比较可以使用用于比较聚合物单元的两个序列的任何已知的技术,典型地是衍生聚合物单元之间的对齐映射的对齐算法,连同用于对齐映射的准确度(因此是相似性的量度65)的得分。可以使用任何数目的可获得的快速对齐算法,如Smith-Waterman对齐算法、BLAST或它们的衍生物、或k聚体计数技术。

[0287] 该形式的参考数据50的这种实施例具有用于衍生相似性的量度65的过程迅速的优点,但是其他形式的参考数据是可能的。

[0288] 在第二实施例中,参考数据50表示通过生化分析系统1采集的实际或模拟的测量

值。在那种情况下,步骤C4或D4包括图14所示的过程,其简单地包括比较测量值的组块63(在这种情况下采集自一系列原始测量值11)与参考数据50以衍生相似性的量度65的步骤C4b。可以进行任何合适的比较,例如使用距离函数来提供两个系列的测量值之间的距离的量度作为相似性的量度65。

[0289] 在第三实施例中,参考数据50表示时序特征的特征向量,其表示由生化分析系统1采集的测量值的特性。可以如W0-2013/121224中详细描述地衍生这种特征向量,参考并通过引用将其结合于本文中。在那种情况下,步骤C4或D4包括如下进行的图15所示的过程。

[0290] 在步骤C4c-1中,分析测量值的组块63,其在该情况下是由一系列原始测量值11采集的,以衍生表示测量值的特性的时序特征的特征向量66。

[0291] 在步骤C4c-2中,将特征向量66与参考数据50比较以衍生相似性的量度65。可以使用W0-2013/121224中详细描述的方法进行比较。

[0292] 在第四实施例中,参考数据50表示参考模型70。在那种情况下,步骤C4或D4包括图16所示的过程,其包括将模型拟合至一系列测量值的组块63来提供作为参考模型70至测量值的组块63的拟合的相似性的量度65的步骤C4d。测量值的组块63可以是一系列原始测量值11或一系列测量值12。

[0293] 可以如下进行C4d步骤。

[0294] 参考模型70是在生化分析系统1中的聚合物单元的参考序列的模型。参考模型70处理测量值作为对应于聚合物单元的参考序列的参考系列的k聚体状态的观察值的。参考模型70的k聚体状态可以模型化测量值取决于其的实际k聚体,但是在数学上这不是必需的,所以k聚体状态可以是实际的k聚体的抽象概念。因此,不同类型的k聚体状态可以对应于存在于聚合物单元的参考序列中的不同类型的k聚体。

[0295] 可以将参考模型70考虑为上述和W0-2013/041878中的类型的一般模型60的改编,以模型化当测量参考序列时具体得到的测量值。因此,参考模型70处理测量值作为对应于聚合物单元的参考序列的参考系列的k聚体状态73的观察值。因而,参考模型70具有与一般模型60相同的形式,特别是包括现在将描述的转换加权71和发射加权72。

[0296] 转换加权71表示参考系列的k聚体状态73之间的转换。那些k聚体状态73对应于聚合物单元的参考序列。因此,参考系列中的连续的k聚体状态73对应于k聚合物单元的连续的重叠组。因而,存在于参考系列的k聚体状态73和参考序列的聚合物单元之间的内在映射。类似地,每种k聚体状态73具有对应于k聚合物单元的组中的每种聚合物单元的不同类型的组合的类型。

[0297] 参考图17的状态图对其进行举例说明,图17示出了估算的k聚体状态73的参考系列中的三种连续的k聚体状态73的实施例。在该实施例中,k是3,并且聚合物单元的参考序列包含标记为A、A、C、G、T的连续的聚合物单元(但是当然k聚体状态73的那些具体类型不受限制)。因此,对应于那些聚合物单元的参考系列的连续的k聚体状态73是类型AAC、ACG、CGT,其对应于聚合物单元的测量序列AACGT。

[0298] 图18的状态图示出了由转换加权71表示的参考系列的k聚体状态73之间的转变。在该实施例中,状态可以仅向前地允许参考系列的k聚体状态73行进通过(但是一般而言另外可以允许向后行进)。如下示出了三种不同类型的转换74、75和76。

[0299] 由参考系列中的每种给定的k聚体状态73,允许到下一k聚体状态73的转换74。这

模型化由参考序列的聚合物单元的连续k聚体采集的一系列测量值12的连续测量值的似然性。在预处理测量值63的组块以识别测量值的连续组,以及衍生由相对于每个识别组预定数目的测量值组成的一系列过程测量值用于进一步分析的情况下,转换加权71表示具有相对高的似然性的这种转换74。

[0300] 由参考系列中的每种给定的k聚体状态73,允许至相同的k聚体状态的转换75。这模型化由参考序列的聚合物单元的相同k聚体采集的一系列测量值12的连续测量值的似然性。可以将其称为“逗留(stay)”。在预处理测量值的组块63以识别测量值的连续组和衍生由预定数目的测量值(关于每个识别组)组成的一系列过程测量值用于进一步分析的情况下,转换加权71表示与转换74相比具有相对高的似然性的这种转换75。

[0301] 由参考系列中的每种给定的k聚体状态73,允许越过下一k聚体状态73的至随后的k聚体状态73的转换76。这模型化采集自下一k聚体状态的无测量值的似然性,使得由聚合物单元的参考序列的k聚体采集的一系列测量值12中的连续测量值是分离的。可以将其称为“跳跃(skip)”。在预处理测量值的组块63以识别测量值的连续组和衍生由预定数目的测量值(关于每个识别组)组成的一系列过程测量值用于进一步分析的情况下,转换加权71表示与转换74相比具有相对高的似然性的这种转换76。

[0302] 可以以与上述的一般模型31中的用于跳跃和逗留的转换加权61相同的方式得到表示用于跳跃和逗留的转换75和76关于表示转换74的转换加权71的水平。

[0303] 在替换实施方式中,没有预处理测量值的组块63来识别测量值的连续组和衍生一系列处理的测量值,使得对测量值的组块63本身进行另外的分析,那么转换加权71类似,但是改写为增加表示跳跃的转换75的似然性以表示由相同k聚体采集的连续测量值的似然性。用于转换75的转换加权71的水平取决于由任何给定k聚体采集的预期的测量值的数目并可以由针对使用的特定的生化分析系统1的实验确定。

[0304] 关于每种k聚体状态,提供发射加权72。发射加权72是当观察k聚体状态时用于观察的不同测量值的加权。发射加权72因此取决于所讨论的k聚体状态的类型。特别地,用于任何给定类型的k聚体状态的发射加权72与用于上述的一般模型60中的那些类型的k聚体状态的发射加权62相同。

[0305] 除了参考模型70替换一般模型60,使用与以上参考图13所描述的相同的技术进行步骤C4d,将模型拟合至一系列的测量值的组块63以提供与参考模型70至测量值的组块63的拟合的相似性的量度65。

[0306] 由于参考模型70的形式,特别是k聚体状态73的参考系列之间的转换的表示,应用模型内在地衍生测量值的组块63和k聚体状态73的参考系列之间的对齐映射的估算值。对其的理解可以如下。由于一般模型60表示k聚体状态的可能类型之间的转换,所以应用该模型提供尤其观察每个测量值的k聚体状态的类型的估算值。由于参考模型70表示k聚体状态73的参考系列之间的转换,所以应用该参考模型70反而估算由其观察每个测量值的参考序列的k聚体状态73,其是一系列测量值和参考系列的k聚体状态73之间的对齐映射。

[0307] 另外,算法衍生对齐映射的准确度的得分,例如表示对齐映射的估算值正确的似然性,例如因为算法基于这种用于模型中不同路径的得分衍生对齐映射。因此,针对对齐映射的准确度的这种得分因此是相似性的量度65。

[0308] 作为一个实施例,在参考模型70是HMM以及应用的分析技术是上述的维特比算法

的情况下,那么得分简单地是通过参考模型70预测的与对齐映射的衍生估算值有关的似然性。

[0309] 作为另一个实施例,在一般模型60是HMM的情况下,分析技术可以是上述的Fariselli等人公开的类型。其再次衍生是相似性65的量度的得分。

[0310] 可以如下由聚合物单元的参考序列或由从聚合物单元的参考序列采集的测量值产生参考模型70。

[0311] 可以如下通过图19所示的过程由聚合物单元80的参考序列产生参考模型70。这可用于由资料库或早期实验已知参考序列的应用。表示聚合物单元80的参考序列的输入数据可以已经存储在数据处理器5中或可以输入其中。

[0312] 该过程使用存储的发射加权81,其包括关于一组可能类型的k聚体状态类型-1至类型-n的发射加权e1至en。有利地,这允许仅基于用于可能类型的k聚体状态的发射加权81来产生用于聚合物单元80的任何参考序列的参考模型。

[0313] 该过程如下进行。

[0314] 在步骤P1中,接收聚合物单元80的参考序列并由其产生k聚体状态73的参考序列。这是简单的过程以针对参考序列中的每种k聚体状态,基于k聚体状态73对应于其的聚合物单元80的类型的组合73,建立那些k聚体状态73的类型。

[0315] 在步骤P2中,如下产生参考模型。

[0316] 针对在步骤P1中衍生的k聚体状态73的参考系列之间的转换衍生转换加权71。转换加权71采用上述的关于k聚体状态73的参考系列定义的形式。

[0317] 在步骤P1中,通过根据k聚体状态73的类型选择存储的发射加权81,针对一系列k聚体状态73中的每种k聚体状态73衍生发射加权72。例如,如果给定的k聚体状态73是类型类型-4,那么选择发射加权e4。

[0318] 如下通过图20所示的过程由从聚合物单元的参考序列采集的一系列参考测量值93产生参考模型70。这可用于,例如其中同时测量聚合物单元的参考序列与靶标聚合物的应用。特别地,在该实施例中,不要求本身已知参考序列的聚合物单元的一致性。可以通过生化分析系统1由包含参考序列的聚合物单元的聚合物采集一系列参考测量值93。

[0319] 该过程使用另外的模型90,其作为不同可能类似的进一步系列k聚体状态的观察值处理一系列参考测量值。这种另外的模型90是用于采集一系列参考测量值93的生化分析系统1的模型并可以与上述的一般模型例如W0-2013/041878中公开的类型60相同。因此,另外的模型包括关于在进一步系列的k聚体状态中的连续的k聚体状态之间的每个转换的转换加权91,其是用于可能类型的k聚体状态之间的可能的转换的转换加权91;和关于每种类型的k聚体状态的发射加权92,当k聚体状态是该类型时是用于观察的不同测量值的发射加权92。

[0320] 如下进行该过程。

[0321] 在步骤Q1中,将另外的模型90应用于一系列参考测量值93来作为离散估算的k聚体状态估算参考系列的k聚体状态73。这可以使用上述的技术进行。

[0322] 在步骤Q2中,如下产生参考模型70。

[0323] 针对在步骤D1中衍生的k聚体状态73的参考系列之间的转换衍生转换加权71。转换加权71采用上述关于k聚体状态73的参考系列定义的形式。

[0324] 在步骤Q1中,通过根据k聚体状态73的类型由另外的模型50的加权选择发射加权,针对一系列k聚体状态73中的每种k聚体状态73衍生发射加权72。因此,用于参考模型中的k聚体状态73的每种类型的发射加权与进一步模型50中的该类型的k聚体状态73的发射加权相同。

[0325] 现在将描述图7示出的方法的、以及更通常根据本发明的第一方面的多种应用的实施例,解释聚合物单元的参考序列的性质、步骤C4中判定的基础和可能的时间节省的表示。在以下实施例中,聚合物是多核苷酸并且假设测量第一250个核苷酸之后与参考序列比较将足以确定(a)其是否涉及该参考序列和(b)其关于总序列的位置。然而,其可以比该数目多或者少。确定所需的聚合物单元的数目将不必须是固定的。典型地,将在连续基础上连续进行测量直到作出这种确定。

[0326] 对于应用类型中的每种,可能存在图7所示的方法的稍微不同的用途。还可以使用应用类型的混合物。还可以随着运行行进动态调节步骤C3中进行的分析和/或步骤C4中的判定的基础。例如,可以不存在初始应用的判定逻辑,然后当建立了足够的数据来作出判定之后将逻辑用于运行。可替代地,判定逻辑可以在运行期间改变。

[0327] 在第一类应用中,衍生参考数据50的聚合物单元的参考序列是不期望的序列,以及在步骤C4中,响应于表示部分移位的聚合物是不期望的序列的相似性的量度作出排斥聚合物的判定。

[0328] 该第一类应用具有多种可能的用途。例如,可以将这种应用于有机体的基因组的序列不完整部分。如果部分限定了有机体的基因组,但是序列不完整,则可以使用本发明的方法确定序列的不完整部分。在这种实施方式中,参考序列可以是基因组的完整部分的序列。聚合物可以是来自有机体的多核苷酸的片段。如果相似性的量度表示聚合物是参考序列(即基因组的已经限定部分的序列),则排斥聚合物并可以通过纳米孔接收新的聚合物。可以重复其直到与参考序列不类似的聚合物部分移位穿过纳米孔,这种聚合物将对应于基因组的之前未限定的部分且可以保留在纳米孔中以及被全部测序。该方法允许迅速测序基因组的未定义的部分。

[0329] 第一类型的应用也可以有利地用于测序来自包含人DNA的聚合物样品的聚合物。人DNA的测序具有与其有关的伦理问题。因此,能够测序聚合物的样品以及忽视人DNA的序列(例如从人患者提取的样品中的细菌识别)是有用的。在这种情况下,参考序列(不期望的序列)可以是人基因组。可以排斥具有表示它们对应于人基因组的部分的相似性的量度的任何聚合物,同时可以将具有表示它们不对应于人基因组的相似性的量度的聚合物保留在纳米孔中并完整完全测序。因此,这是其中相似性的量度表示与参考序列的一部分的相似性的方法的一个实施例。在本申请中,方法避免测序人DNA,但是允许测序细菌DNA。如果细菌在来自人肠道的样品中,我们假设细菌DNA(其是我们想要测序的DNA或“目标”DNA)是DNA的约5%以及样品中的DNA的95%是人DNA(“脱靶DNA”)。如果我们假设每个片段约250bp(碱基对)序列将足以提供所需的相似性的量度,以及聚合物可以以25碱基/秒的速率移位穿过孔,那么不是靶标DNA的聚合物(即与人DNA参考序列类似的DNA(“脱靶”聚合物)将在被排出之前移位穿过纳米孔约10秒。因此,可以认为其中纳米孔包含脱靶聚合物的相对时间量是 $95\% \times 10 = 9.5$ 。另一方面,假设将DNA片段化为10kB片段,则测序靶标DNA中的一个片段所花费的时间的量将是 $10,000/25$,是400秒。因此,可以将其中纳米孔包含目标聚合物的时间

的相对量认为是 $5\% \times 400$,是20秒。所以可以将其中纳米孔包含目标链的时间的比例认为是其中纳米孔包含目标链的时间/其中纳米孔包含脱靶链的时间+其中纳米孔包含目标链的时间,其是 $20/29.5$ 。另一方面,如果需要以它们的整体测序脱靶链,则其中纳米孔包含脱靶链的时间的相对量将是 $95\% \times 400$,其是380,以及所以可以将纳米孔包含目标链的时间的比例认为是 $20/380$ 。这表示约13.6倍的效率。

[0330] 第一类应用还可以有利地用于测序样品中的污染物。在这种实施方式中,参考序列将是存在于样品中的已知的组分的序列。例如,可以使用其来检测食品如类似牛肉产品的肉制品中的污染物。在这种情况下,参考序列将是来自源自食品的有机体的多核苷酸(例如该有机体的基因组)的序列。参考序列可以是奶牛的基因组的序列。可以排斥具有表示它们对应于奶牛基因组的相似性的量度的样品中的任何聚合物,同时可以将具有表示它们不对应于奶牛基因组的相似性的量度的聚合物保留在纳米孔中并完全测序。这将允许在不需要知道污染物的性质的情况下快速和简单地定义污染物的性质。这相比要求知道怀疑污染物的现有技术的方法如定量PCR是有利的。假设99%的DNA脱靶(肉类DNA)和1%的DNA是目标(例如污染物),那么本发明的方法将比如果纳米孔不能排出不期望的聚合物更有效约29倍。

[0331] 在第二类应用中,衍生参考数据50的聚合物单元的参考序列是靶标,以及在步骤C4中,响应于表示部分移位的聚合物不是靶标的相似性的量度作出排斥聚合物的判定。

[0332] 这种第二类型的应用可以有利地用于测序来自DNA样品的感兴趣的基因。在这种应用中,参考序列是靶标,其可以是多核苷酸的一部分如感兴趣的基因,并且聚合物可以包含来自样品的多核苷酸如DNA的片段。可以排斥具有表示它们与靶标(感兴趣的基因)不类似的相似性的量度的样品中的任何聚合物。可以保留剩余的聚合物并测序。这允许迅速测序感兴趣的基因且关于现有技术是有利的,现有技术要求在测序之前分离感兴趣的靶标基因(例如通过将感兴趣的基因杂交至附接至固体表面的探针)。这种分离技术耗时且当使用本发明的方法时是不需要的。这种应用的一个实施例将是测序人基因组。人基因组包含50Mb(百万碱基)编码序列。能够测序该50Mb而不是剩余的3,000Mb将是理想的。因此,“脱靶”(应当排斥)的DNA的量是3,000Mb。DNA将被片段化至长度约10kB的片段,并因此3,000Mb将表示约300,000个片段。假设每个片段约250bp的序列将足以提供要求的相似性的量度,且聚合物可以以25碱基/秒的速率移位穿过孔,那么与靶标聚合物不类似(“脱靶”人DNA)的聚合物将在被喷射之前移位穿过纳米孔约10秒。由于存在300,000个脱靶片段,所以脱靶片段将以约3,000,000秒/纳米孔保留在孔内(片段数乘以每个片段保留在孔中的时间-约10秒)。与靶标聚合物类似的剩余的50Mb(“目标”)将花费2,000秒(在25碱基/秒下将花费的时间等于 $50,000,000/25$ 或2,000,000秒)。测序描述的50Mb靶标聚合物的总时间是测序脱靶聚合物花费的时间的量 and 测序目标聚合物花费的时间的量的总和,其是 $3,000,000+2,000,000$ 或5,000,000秒/纳米孔。另一方面,如果测序300,000个脱靶片段中的每个的整体,那么这将花费 $3,000,000,000/25$ (在25碱基对/秒的速率下测序3,000Mb)+2,000,000(测序目标聚合物花费的时间),其是122,000,000秒/孔(长超过50倍)来测序一次基因组。

[0333] 该第二类应用还可以有利地用于识别样品(例如来自住院患者)中的细菌是否耐抗生素。此处,参考序列将是靶标,其可以是对应于特定的耐抗生素基因的多核苷酸。可以排斥具有表示与目标耐抗生素基因类似的相似性的量度的样品中的任何聚合物。如果检测

到没有聚合物具有表示它们与耐抗生素基因类似的相似性的量度,则这将表示细菌正在丢失特定的耐抗生素基因。可替代地,如果检测到聚合物确实具有表示它们与耐抗生素的基因类似的相似性的量度,则可以保留它们并测序,且序列用于确定耐抗生素基因是否是功能性的。在这种情况下,脱靶聚合物(细菌的基因组)将是约5000kB,以及目标聚合物(感应区的区域)将是约5kB。作与上述相同的假设,意味着本发明的方法将比如果纳米孔不能喷射不期望的聚合物快约40倍测序DNA。

[0334] 这种第二类型的应用也可以有利地用于测序总细菌mRNA。在这种情况下,期望能够测序mRNA,但是能够忽略rRNA或tRNA的序列。此处,参考序列可以是靶标序列如细菌基因组的注释版本。聚合物可以包含来自细菌的样品的RNA。样品中具有表示它们与靶标细菌基因组不类似的相似性的量度的任何聚合物将与rRNA或tRNA有关,且可以将其排斥。剩余的聚合物将对应于mRNA并可以对其测序以提供总细菌mRNA的序列。在这种情况下,目标聚合物将是mRNA(其是总RNA的约5%),以及脱靶聚合物将是tRNA和rRNA,其是总RNA的约95%。使用与以上定义的那些相同的假设,我们预期测序效率升高约8.4倍。

[0335] 该第二类型的应用也可以有利地用于识别菌株用于表型或SNP(单核苷酸多形性)检测,其中细菌的菌株不是已知的。例如,在这种情况下,聚合物可以是来自细菌样品的多核苷酸的片段。最初,不排斥聚合物(没有使用参考序列)并测序已经移位穿过孔的聚合物,但是当已经得到足够的序列信息来允许用户确定细菌的菌株时,那么选择参考序列。参考序列将对应于感兴趣的靶标区域且将取决于已经定义的细菌的种类。一旦已经定义的参考序列,则保留并完全测序部分移位穿过孔并具有表示它们与参考序列类似的相似性的量度的任何聚合物(感兴趣的靶标部分),同时可以排斥其他聚合物。这将允许检测表型或SNP的存在。

[0336] 类似地,这种第二类应用将可用于癌症的表型。在这种应用中,聚合物可以是由癌症患者得到的多核苷酸的片段。最初,参考序列可以是靶标序列。这些靶标序列可以是多核苷酸如与不同类别的癌症有关的基因的序列。将保留具有与这些靶标序列的相似性的量度的任何聚合物,并排斥其他聚合物。然而,一旦识别了癌症的类别,但是可以细化参考序列使得参考序列现在包含具有与癌症的子类别有关的多核苷酸的序列的靶标。

[0337] 在第三类应用中,衍生参考数据50的聚合物单元的参考序列是已经测量的聚合物单元的序列,以及在步骤C4中,响应于表示部分移位的聚合物是已经测量的聚合物单元的序列的相似性的量度作出排斥聚合物的判定。

[0338] 这类应用可用于使能够准确测序基因组。确定基因组的序列要求进行DNA的多个链的序列,以及为了准确度,将确定该部分DNA的共有序列。因此,应当将对应于该序列的相同部分的聚合物测序足够次以能够定义准确的共有序列。为此,本发明的方法可用于迅速和准确地测序基因组。例如,聚合物可以包含来自将定义基因组的有机体的DNA的样品的DNA。参考序列是已经采集了充分的测量值的DNA的一部分(在这种情况下已经得到充分的序列数据以提供准确的共有序列)。最初,没有排斥序列。然而,一旦计算到已经得到充分的用于基因组的一部分的序列以允许计算准确的共有序列,那么该共有序列变为靶标(参考序列)。可以排斥部分移位穿过孔且具有表示它们与参考序列(已经定义了其准确的共有序列的DNA的部分)类似的相似性的量度的任何聚合物,释放纳米孔以测序还没有收集其充分的信息的基因组的其他部分。

[0339] 在第四类应用中,衍生参考数据50的聚合物单元的参考序列包含多个靶标,以及在步骤C4中,响应于表示部分移位的聚合物是靶标中的一种的相似性的量度作出排斥聚合物的判定。

[0340] 这是可以用于量化靶标聚合物的样品中的每种靶标聚合物的比例的计数法。例如,靶标可以表示不同的聚合物。当聚合物部分移位穿过纳米孔时,可以将具有表示它们与参考序列类似的相似性的量度的任何聚合物分配到一个“桶”并可以量化检测属于每个“桶”的聚合物的数目。在这种实施方式中,一旦得到关于聚合物的充分的信息来确定它是否具有表示其与参考序列中的一种类似的相似性的量度,将排斥聚合物。这种技术的用途的一个实施例是量化污染物。例如,聚合物可以是食品如牛肉产品的样品。在这种情况下,参考序列可以包含具有在奶牛DNA中发现的序列的靶标和具有在马DNA中发现的序列的靶标。可以使用该方法计算与奶牛DNA靶标类似的聚合物的比例和与马DNA类似的聚合物的比例,且这将表示牛肉产品被马肉污染的水平。

[0341] 类似地,如果使用的参考序列包含具有在不同细菌中发现的序列的靶标,则该技术可以用于确定存在于样品如来自受感染患者的样品中的不同细菌的比例。

[0342] 图16所示的方法导致产生对齐映射。可以如下更通常地应用该方法。

[0343] 图21示出了估算(a)包含聚合物单元的聚合物的一系列测量值和(b)聚合物单元的参考序列之间的对齐映射的方法。如下进行该方法。

[0344] 如图21所示,输入该方法的可以是通过由生化分析系统1采集聚合物单元的序列的一系列原始测量值以及使它们经受如上述的预处理衍生的一系列测量值12。作为替换,输入该方法的可以是一系列原始测量值11。

[0345] 该方法使用聚合物单元的参考序列的参考模型70,所述参考模型70存储在数据处理器5的存储器10中。参考模型70采用与上述相同的形式,处理测量值,作为对应于聚合物单元的参考序列的k聚体状态的参考序列的观察值。

[0346] 将参考模型70用于对齐步骤S1。特别地,在对齐步骤S1中,将参考模型70应用于一系列测量值12。以与以上步骤C4d相同的方式进行对齐步骤S1。换言之,除了参考模型70替换一般模型60,通过使用与以上参考图13所描述的相同的技术进行步骤C4d,将模型拟合至一系列的测量值63的组块以提供与参考模型70至测量值63的组块的拟合的相似性65的量度来进行对齐步骤S1。

[0347] 由于参考模型70的形式,特别是k聚体状态73的参考系列之间的转换的表示,应用模型内在地衍生一系列测量值和k聚体状态73的参考系列之间的对齐映射的估算值。对其的理解可以如下。由于一般模型60表示k聚体状态的可能类型之间的转换,所以应用该模型提供由其观察每个测量值的k聚体状态的类型的估算值,即,k聚体状态34的初始系列的估算值和离散估算的k聚体状态35,由k聚体状态的类型观察每个测量值的每个估算值。由于参考模型70表示k聚体状态73的参考系列之间的转换,所以应用该参考模型70反而估算尤其观察每个测量值的参考序列的k聚体状态73,其是一系列测量值和参考系列的k聚体状态73之间的对齐映射。

[0348] 由于参考系列的k聚体状态73和参考序列的聚合物单元之间存在固有的映射,所以k聚体状态73的一系列测量值和参考系列之间的对齐映射还提供聚合物单元的一系列测量值与参考序列之间的对齐映射。

[0349] 图22示出了对齐映射的一个实施例来举例说明其性质。特别地,图22示出了参考序列的聚合物单元p0至p7、参考系列的k聚体状态k1至k6、以及测量值m1至m7之间的对齐映射。通过举例说明,在该实施例中,k是三。水平线表示k聚体状态和测量值之间的对齐,或在短划线的情况下其他系列中的缺口的对齐。因此,固有地,如举例说明的参考序列的聚合物单元p0至p7对其到参考系列的k聚体状态k1至k6。k聚体状态k1对应于并映射到聚合物单元p1至p3等等。至于参考系列的k聚体状态k1至k6和测量值m1至m7之间的映射:k聚体状态k1映射到测量值m1,k聚体状态k2映射到测量值m2,k聚体状态k3映射到一系列测量值中的缺口,k聚体状态k4映射到测量值m3,以及测量值m4和m5映射到一系列k聚体状态中的缺口。

[0350] 取决于应用的方法,可以如下改变对齐映射的估算值13的形式。

[0351] 如上所述,在对齐步骤S1中应用的分析技术可以采用适用于参考模型70的形式的各种形式。例如,在参考模型70是HMM的情况下,分析技术可以是用于解决HMM的已知的算法,例如本领域熟知的前向-后向算法(Forward-Backward algorithm)或维特比算法。一般来说,这样的算法可以避免通过状态的序列的所有可能的路径的似然(可能性)的蛮力计算(force calculation),而是利用基于似然的简化的方法来确定状态序列。

[0352] 通过在对齐步骤S1中应用的一些技术,对齐映射的衍生估算值13对于系列中的每个测量值12包括关于k聚体状态73的参考系列中的不同k聚体状态73的加权。例如,可以通过 $M_{i,j}$ 表示这种对齐映射,其中指数i标示测量值以及指数j标示参考系列中的k聚体状态,从而在存在K个k聚体状态时, $M_{i,1}$ 至 $M_{i,K}$ 的值表示用于关于k聚体状态73的参考系列中的每个k聚体状态73的第i个测量值的加权。在这种情况下,估算值13因为映射到每个测量值不表示单个k聚体状态73,反而提供如此映射到每个测量值的不同可能的k聚体状态73的加权。

[0353] 作为参考模型70是HMM的情况下的一个实施例,当应用的分析技术是上述的前向-后向算法时,衍生的估算值可以是这种类型。在前向-后向算法(Forward-Backward algorithm)中,使用转换和发射加权以前向和后向方向循环计算以给定k聚体状态结束的所有序列的总似然性。结合这些前向和后向概率以及连同数据的总似然一起计算,来自给定的k聚体状态的每个测量的概率。称作后矩阵的这种概率矩阵是对齐映射的估算值13。

[0354] 在这种情况下,在随后的得分步骤S2(其是可选的)中,存在表示对齐映射的估算值13正确的似然性的得分14。这可以使用简单的概率技术,由对齐映射的估算值13衍生,或可替代地,可以作为对齐步骤S1的内在部分衍生。

[0355] 通过在对齐步骤S1中应用的其他技术,对齐映射的衍生估算值13对于系列中的每个测量值包括k聚体状态的参考系列中的k聚体状态的离散估算值。例如,这种对齐映射可以由 M_i 表示,其中指数i标示测量值以及 M_i 可以采用表示K个k聚体状态的值1至K。在这种情况下,估算值13表示映射到每个测量值的单个k聚体状态73。

[0356] 作为在参考模型70是HMM的情况下的一个实施例,当应用的分析技术是上述的维特比算法时,衍生的估算值可以是这种类型,其中分析技术基于通过k聚体状态的参考系列产生的一系列测量值的模型预期的似然性估算k聚体的序列。

[0357] 在衍生的对齐映射的估算值13包括k聚体状态的离散估算值的情况下,算法内在的衍生表示对齐映射的估算值正确的似然性的得分14,因为算法基于针对通过模型的不同路径的得分衍生对齐映射。因此,在这种情况下,不进行单独的得分步骤S2。作为一个

实施例,在参考模型70是HMM以及应用的分析技术是上述的维特比算法的情况下,那么得分简单地是通过参考模型70预测的与对齐映射的衍生估算值13有关的似然性。

[0358] 图21所示的方法具有广泛的应用,其中期望其估算聚合物的一系列测量值和聚合物单元的参考序列之间的对齐映射和/或表示对齐映射准确的似然性的得分。这种对齐映射的评估可以用于各种应用,如比较参考来提供样品中的聚合物的存在、不存在或程度的识别或检测,例如来提供诊断。可能范围的特定应用是大量的并可以应用于检测具有DNA序列的任何分析物。

[0359] 以上实施例涉及单个参考模型70。在多种应用中,可以使用多个参考模型70。如图21所示的方法可以应用为使用每个参考模型70,或可以选择参考模型70中的一个。根据应用,可以基于多种标准进行选择。例如,参考模型70可以应用于不同类型的传感器设备2(例如不同的纳米孔)和/或外界条件,在这样的情况下,基于实际使用的传感器设备2的类型和/或实际的外界条件选择以下所描述的参考模型8。在另一实施例中,可以基于待检测的分析物作出选择,例如特定的G/C富集或是否用实验确定特定的外遗传信息。

[0360] 因此,根据本发明的第四方面,提供了估算以下各项之间的对齐映射的方法:(a) 包含聚合物单元的聚合物的一系列测量值,其中测量值取决于k聚体,k聚体是聚合物的k个聚合物单元,其中k是整数,和(b) 聚合物单元的参考序列;

[0361] 该方法使用参考模型,该参考模型处理作为对应于聚合物单元的参考序列的一系列参考k聚体状态的观察值的参考数据,其中,参考模型包括:

[0362] 用于参考系列的k聚体状态中的k聚体状态之间的转换的转换加权;以及

[0363] 关于每个k聚体状态,当观察k聚体状态时用于观察的不同测量值的发射加权;以及

[0364] 方法包括应用参考模型至一系列测量值以衍生一系列测量值和对应于聚合物单元的参考序列的k聚体状态的参考系列之间的对齐映射的估算值。

[0365] 可以以任意组合将以下特征可选地应用于本发明的第四方面:

[0366] 对于系列中的每个测量值,衍生的对齐映射的估算值可以包括参考系列k聚体状态中的映射的k聚体状态的离散估算值。

[0367] 对于系列中的每个测量值,衍生的对齐映射的估算值可以包括关于参考系列k聚体状态中的不同映射的k聚体状态的加权。

[0368] 方法可以进一步包括衍生表示对齐映射的估算值正确的似然性的得分。

[0369] 方法可以进一步包括通过包括以下各项的过程,使用存储的关于k聚体状态的一组可能类型的发射加权由聚合物单元的参考序列产生参考模型:

[0370] 衍生对应于接收的聚合物的参考序列的一系列k聚体状态;

[0371] 通过产生衍生的k聚体状态系列中的k聚体状态之间的转换的转换加权和通过根据k聚体状态的类型由存储的发射加权选择用于衍生系列中的每种k聚体状态的发射加权来产生参考模型。

[0372] 方法可以进一步包括由包含聚合物单元的参考序列的聚合物的一系列参考测量值产生参考模型。

[0373] 产生参考模型的步骤可以使用另外的模型,该另外的模型作为不同可能类型的进一步系列的k聚体状态的观察值处理一系列参考测量值,其中,另外的模型包括:

[0374] 关于在进一步系列的k聚体状态中的连续的k聚体状态之间的每个转换,用于在可能类型的k聚体状态之间的可能的转换的转换加权;以及

[0375] 关于每种类型的k聚体状态,当k聚体状态是该类型时,用于观察的不同测量值的发射加权。

[0376] 产生参考模型的步骤包括:

[0377] 通过将另外的模型应用于一系列参考测量值产生参考系列的k聚体状态的估算值;以及

[0378] 通过产生参考系列的k聚体状态的估算值中的k聚体状态之间的转换的转换加权和通过根据k聚体状态的类型由进一步模型的加权选择用于产生的参考系列的估算值中的每种k聚体状态的发射加权来产生参考模型。

[0379] 可以预存储参考模型。

[0380] 转换加权和发射加重的一个或两个可以是概率。

[0381] 模型可以是隐马尔可夫模型。

[0382] 整数k可以是复数。

[0383] 测量值可以是在所述聚合物移位穿过纳米孔期间采集的测量值。

[0384] 所述聚合物通过纳米孔的移位可以以棘轮的方式进行。

[0385] 纳米孔可以是生物孔。

[0386] 聚合物可以是多核苷酸,以及聚合物单元可以是核苷酸。

[0387] 单个测量值可以取决于k聚体,或不同性质的预定复数的测量值可以是取决于相同的k聚体。

[0388] 测量值可以包含电流测量值、阻抗测量值、隧道测量值、电场效应晶体管测量值和光学测量值中的一种或多种。

[0389] 可以将参考模型存储在存储器中。

[0390] 在将参考模型应用于一系列测量值的步骤之前,方法可以进一步包括通过以下衍生一系列所述测量值:

[0391] 在先前未知组中的测量值的数目的情况下,由聚合物接收一系列的原始测量值,其中多个原始测量值的一系列原始测量值组取决于相同的k聚体,以及

[0392] 处理一系列原始测量值以识别测量值的连续组以及关于每个识别的组衍生不同类型的单个测量值或多个测量值以形成所述系列的测量值。

[0393] 方法可以进一步包括由聚合物采集一系列的原始测量值。

[0394] 在多个系列的测量值的每个中,在未知组中的测量值的数目的情况下,多个测量值的组可以取决于相同的k聚体。

[0395] 方法可以进一步包括由聚合物采集所述系列的测量值。

[0396] 序列表

[0397] Seq ID 1:MS-(B1)8=MS-(D90N/D91N/D93N/D118R/D134R/E139K)8

[0398] ATGGGTCTGGATAATGAAGTGAAGCTGGTGGACGGTCAAGATCGTACCCTGACGGTGCAACAATGGGATACCTTTCTGAATGGCGTTTTTCCGCTGGATCGTAATCGCCTGACCCGTGAATGGTTTCATTCCGGTCGCGCAAAATATATCGTCGAGGCCCGGGTGCTGACGAATTCGAAGGCACGCTGGAAGTGGGTTATCAGATTGGCTTTCCGTGGTCAC TGGGCGTTGGTATCAACTTCTCGTACACCACGCCGAATATTCTGATCAACAATGGTAACATTACCGCACCGCGTTT

GGCCTGAACAGCGTGATTACGCCGAACCTGTTTCCGGGTGTTAGCATCTCTGCCCCGTCTGGGCAATGGTCCGGGCAT
TCAAGAAGTGGCAACCTTTAGTGTGCGCGTTTCCGGCGCTAAAGGCGGTGTGCGGTGTCTAACGCCACGGTACCG
TTACGGGCGCGGCCGGGTGTCCTGCTGCGTCCGTTTCGCGCGCCTGATTGCCTCTACCGGCGACAGCGTTACGACC
TATGGCGAACCGTGAATATGAACTAA

[0399] Seq ID 2:MS- (B1) 8=MS- (D90N/D91N/D93N/D118R/D134R/E139K) 8

[0400] GLDNELSLVDGQDRTLTVQQWDTFLNGVFPLDRNRLTREWFHSGRAKYIVAGPGADEFEGTLELGYQIG
FPWSLGVGINFSYTPNININNGNITAPPFGLNSVITPNLFPGVISARLNGPGIQEVATFSVRVSGAKGGVAVSN
AHGTVTGAAGGVLLRPFARLIASGTDSVTTYGEPWNMN

[0401] Seq ID 3:MS- (B2) 8=MS- (L88N/D90N/D91N/D93N/D118R/D134R/E139K) 8

[0402] ATGGGTCTGGATAATGAACTGAGCCTGGTGGACGGTCAAGATCGTACCCTGACGGTGCAACAATGGGAT
ACCTTTCTGAATGGCGTTTTTCCGCTGGATCGTAATCGCCTGACCCGTGAATGGTTTCATTCCGGTCGCGCAAATA
TATCGTCGAGGCCCGGGTGTGACGAATTCGAAGGCACGCTGGAACGGGTTATCAGATTGGCTTTCCGTGGTCAC
TGGGCGTTGGTATCAACTTCTCGTACACCACGCCGAATATTAACATCAACAATGGTAACATTACCGCACCGCGTTT
GGCCTGAACAGCGTGATTACGCCGAACCTGTTTCCGGGTGTTAGCATCTCTGCCCCGTCTGGGCAATGGTCCGGGCAT
TCAAGAAGTGGCAACCTTTAGTGTGCGCGTTTCCGGCGCTAAAGGCGGTGTGCGGTGTCTAACGCCACGGTACCG
TTACGGGCGCGGCCGGGTGTCCTGCTGCGTCCGTTTCGCGCGCCTGATTGCCTCTACCGGCGACAGCGTTACGACC
TATGGCGAACCGTGAATATGAACTAA

[0403] Seq ID 4:MS- (B2) 8=MS- (L88N/D90N/D91N/D93N/D118R/D134R/E139K) 8

[0404] GLDNELSLVDGQDRTLTVQQWDTFLNGVFPLDRNRLTREWFHSGRAKYIVAGPGADEFEGTLELGYQIG
FPWSLGVGINFSYTPNININNGNITAPPFGLNSVITPNLFPGVISARLNGPGIQEVATFSVRVSGAKGGVAVSN
AHGTVTGAAGGVLLRPFARLIASGTDSVTTYGEPWNMN

[0405] Seq ID:5 (WT EcoExo I) :

[0406] MMNDGKQQSTFLFHDYETFGTHPALDRPAQFAAIRTDSEFNVIGEPEVFYCKPADDYLPQPGAVLITGI
TPQEARAKGENEAAFAARIHSLFTVPKTCILGYNNVRFDDDEVTRNIFYRNFYDPYAWSQHDNSRWDLDDVMRACYA
LRPEGINWPENDDGLPSFRLEHLTKANGIEHSNAHDAMADVYATIAMAKLVKTRQPRFLDYLFTHRNKHKLMALIDV
PQMKPLVHVSGMFGAWRGNTSWVAPLAWHPENRNAVIMVDLAGDISPILLESDTLRERLYTAKTDLGDNAAPVKL
VHINKCPVLAQANTLRPEDADRLGINRQHCLDNLIKILRENPPVREKVVAIFAEAEPTPSDNVDAQLYNGFFSDADR
AAMKIVLETEPRNLPALDITFVDKRIEKLNFYRARNFPGTLDYAEQQRWLEHRRQVFTPEFLQGYADELQMLVQQY
ADDKEKVALLKALWQYAEIEVSGSGHHHHHH

[0407] Seq ID:6 (大肠杆菌核酸外切酶III) :

[0408] MKFVSFNINGLRARPHQLEAIVEKHQPDVIGLQETKVHDDMFPLEEVAKLGYNVFYHGQKGHYGVALLT
KETPIAVRRGFGPDDEEAQRRIIMAEIPSLLGNTVINGYFPQGESRDHPKFKPAKAQFYQNLQNYLETELKRDNPV
LIMGDMNISPTDLDIGIGEENRKRWLRTGKCSFLPEEREWMDRMSWGLVDTRHANPQTADRFSWFDYRSKGFDDN
RGLRIDLLLASQPLAECCVETGIDYEIRSMEKPSDHAPVWATFRR

[0409] Seq ID:7 (嗜热菌RecJ) :

[0410] MRDRVRWRVLSLPLAQWREVMAALEVGPAAALAYWHRGFRRKEDLPPLALLPLKGLREAAALLEAL
RQGKRIRVHGDYDADGLTGTAIVRGLAALGADVHPFIPHRLEEGYGVLMERVPEHLEASDLFTVDCGITNHAELR
ELLENGVEVIVTDHHTPGKTPSPGLVVHPALTPDLKEKPTGAGVVFLLLWALHERLGLPPLEYADLAAVGTIADVA
PLWGWNRALVKEGLARIPASSWVGLRLLAEAVGYTGKAVEVAFRIAPRINAASRLGEAEKALRLLLTDAAEAQALV

GELHRLNARRQTLEEAMLRKLLPQADPEAKAIVLLDPEGHPGVMGIVASRILEATLRPVFLVAQGKGTVRSLAPISA
VEALRSAEDLLRLRYGGHKEAAGFAMDEALFPAFKARVEAYAARFPDPVREVALLDLLPEPGLLPQVFRELALLEPYG
EGNPEPLFLLFGAPEEARRLGEGRHAFRLKGVRLAWKQGDALPPEVEVAGLLSENAWNGHLAYEVQAVDLRKPE
ALEGGIAPFAYPLPLLEALARARLGEGVYPEDNPEGLDYARKAGFRLLPPEEAGLWLGLPPRPVLGRRVEVALGRE
ARARLSAPPVLHTPEARLKALVHRLLFAYERRHPLGFSEALLAYWEVNRVQEPAGSP

[0411] Seq ID:8 (λ核酸外切酶) :

[0412] MTPDIILQRTGIDVRAVEQGGDAWHKLRGLGITASEVHNVIKPRSGKKWPDMMKMSYFHTLLAEVCTGV
APEVNAKALAWGKQYENDARTLFEFTSGVNVTESPIIYRDESMRTACSPDGLCSDGNGLLEKCPFTSRDFMKFRLGG
FEAIKSAYMAQVQYSMWVTRKNAWYFANYDPRMKREGLHYVVIERDEKYMASFDEIVPEFIEKMDEALAEIGFVFGE
QWR

[0413] Seq ID:9 (Phi29 DNA聚合酶) :

[0414] MKHMPRKMYSCAFETTTKVEDCRVWAYGYMNIEDHSEYKIGNSLDEFMAWVLKVQADLYFHNLFKFDGAF
IINWLERNGFKWSADGLPNTYNTIISRMGQWYMIDICLGKGRKIHTVIYDSLKKLPFPVKKIADFKLTVLKGD
DYHKERPVGKITYPEEYAYIKNDIQIIAEALLIQFKQGLDRMTAGSDSLKGFKDIITTKKFKKVFPTLSLGLDKEVR
YAYRGGFTWLNDRFKEKEIGEGMVFDVNSLYPAQMYSRLLPYGEPIVFEGKYVWDEDYPLHIQHIRCEFELKEGYIP
TIQIKRSRFYKGNEYLKSSGGEIADLWLSNVDLELMKEHYDLYNVEYISGLKFKATTGLFKDFIDKWYIKTTSEGA
IKQLAKLMLNSLYGKFASNPDVTGKVPYLGKENGALGFRLGEEETKDPVYTPMGVFIWARYTTITAAQACYDRIIY
CDTDSIHLTGTEIPDVIKDIVDPKKLGYWAHESTFKRAKYLRQKTYIQDIYMKEVDGKLVEGSPDDYTDIKFSVKCA
GMTDKIKKEVTFFENFKVGFSRKMKPKPVQVPGGVVLVDDTFTIKSGGSAWSHPQFEKGGGSGGGSGGSAWSHPQFEK

- [0001] 序列表
- [0002] <110> 牛津楠路珀尔科技有限公司
- [0003] <120> 聚合物的分析
- [0004] <130> N403966W0
- [0005] <150> GB 1418379.2
- [0006] <151> 2014-10-16
- [0007] <150> GB 1418366.9
- [0008] <151> 2014-10-16
- [0009] <150> GB 1507742.3
- [0010] <151> 2014-05-10
- [0011] <160> 9
- [0012] <170> PatentIn 版本 3.5
- [0013] <210> 1
- [0014] <211> 558
- [0015] <212> DNA
- [0016] <213> 人工序列
- [0017] <220>
- [0018] <223> MS- (B1) 8 = MS- (D90N/D91N/D93N/D118R/D134R/E139K) 8
- [0019] <400> 1
- [0020] atgggtcttg ataatgaact gagcctggtg gacggtcaag atcgtaccct gacggtgcaa 60
- [0021] caatgggata cctttctgaa tggcggttttt cgcgtggatc gtaatcgctt gacccgtgaa 120
- [0022] tggttttcatt ccggtcgcgc aaaatatatc gtcgcaggcc cgggtgctga cgaattcgaa 180
- [0023] ggcacgctgg aactgggtta tcagattggc tttccgtggt cactgggcgt tggatatcaac 240
- [0024] ttctcgtaca ccacgccgaa tattctgata aacaatggta acattaccgc accgccgttt 300
- [0025] ggctgaaca gcgtgattac gccgaacctg tttccgggtg ttagcatctc tgcccgtctg 360
- [0026] ggcaatggtc cgggcattca agaagtggca acctttagtg tgcgcgtttc cggcgctaaa 420
- [0027] ggcggtgtcg cgggtgtctaa cgcccacggt accgttacgg gcgcggccgg cgggtgtcctg 480
- [0028] ctgcgtccgt tcgcgcgcct gattgcctct accggcgaca gcgttacgac ctatggcgaa 540
- [0029] ccgtggaata tgaactaa 558
- [0030] <210> 2
- [0031] <211> 184
- [0032] <212> PRT
- [0033] <213> 人工序列
- [0034] <220>
- [0035] <223> MS- (B1) 8 = MS- (D90N/D91N/D93N/D118R/D134R/E139K) 8
- [0036] <400> 2
- [0037] Gly Leu Asp Asn Glu Leu Ser Leu Val Asp Gly Gln Asp Arg Thr Leu
- [0038] 1 5 10 15

[0039]	Thr Val Gln Gln Trp Asp Thr Phe Leu Asn Gly Val Phe Pro Leu Asp
[0040]	20 25 30
[0041]	Arg Asn Arg Leu Thr Arg Glu Trp Phe His Ser Gly Arg Ala Lys Tyr
[0042]	35 40 45
[0043]	Ile Val Ala Gly Pro Gly Ala Asp Glu Phe Glu Gly Thr Leu Glu Leu
[0044]	50 55 60
[0045]	Gly Tyr Gln Ile Gly Phe Pro Trp Ser Leu Gly Val Gly Ile Asn Phe
[0046]	65 70 75 80
[0047]	Ser Tyr Thr Thr Pro Asn Ile Leu Ile Asn Asn Gly Asn Ile Thr Ala
[0048]	85 90 95
[0049]	Pro Pro Phe Gly Leu Asn Ser Val Ile Thr Pro Asn Leu Phe Pro Gly
[0050]	100 105 110
[0051]	Val Ser Ile Ser Ala Arg Leu Gly Asn Gly Pro Gly Ile Gln Glu Val
[0052]	115 120 125
[0053]	Ala Thr Phe Ser Val Arg Val Ser Gly Ala Lys Gly Gly Val Ala Val
[0054]	130 135 140
[0055]	Ser Asn Ala His Gly Thr Val Thr Gly Ala Ala Gly Gly Val Leu Leu
[0056]	145 150 155 160
[0057]	Arg Pro Phe Ala Arg Leu Ile Ala Ser Thr Gly Asp Ser Val Thr Thr
[0058]	165 170 175
[0059]	Tyr Gly Glu Pro Trp Asn Met Asn
[0060]	180
[0061]	<210> 3
[0062]	<211> 558
[0063]	<212> DNA
[0064]	<213> 人工序列
[0065]	<220>
[0066]	<223> MS- (B2) 8 = MS- (L88N/D90N/D91N/D93N/D118R/D134R/E139K) 8
[0067]	<400> 3
[0068]	atgggtcttg ataatgaact gagcctggtg gacggtcaag atcgtagcct gacggtgcaa 60
[0069]	caatgggata cttttctgaa tggcggtttt ccgctggatc gtaatcgctt gacccgtgaa 120
[0070]	tggttttcatt cgggtcgcgc aaaatatatc gtcgcaggcc cgggtgctga cgaattcgaa 180
[0071]	ggcacgctgg aactgggtta tcagattggc ttcccggtgt cactgggcgt tggtatcaac 240
[0072]	ttctcgtaca ccacgcgaa tattaacatc aacaatggta acattaccgc accgccgttt 300
[0073]	ggcctgaaca gcgtgattac gccgaacctg ttccgggtg ttagcatctc tgcccgtctg 360
[0074]	ggcaatggtc cgggcattca agaagtggca accttttagt tgcgcgtttc cggcgctaaa 420
[0075]	ggcgggtgtc cgggtgtctaa cgcccacggt accgttacgg gcgcggccgg cgggtgtcctg 480
[0076]	ctgcgtccgt tcgcgcgcct gattgcctct accggcgaca gcgttacgac ctatggcgaa 540
[0077]	ccgtggaata tgaactaa 558

44

[0117]	20	25	30
[0118]	Ala Ile Arg Thr Asp Ser Glu Phe Asn Val Ile Gly Glu Pro Glu Val		
[0119]	35	40	45
[0120]	Phe Tyr Cys Lys Pro Ala Asp Asp Tyr Leu Pro Gln Pro Gly Ala Val		
[0121]	50	55	60
[0122]	Leu Ile Thr Gly Ile Thr Pro Gln Glu Ala Arg Ala Lys Gly Glu Asn		
[0123]	65	70	75
[0124]	Glu Ala Ala Phe Ala Ala Arg Ile His Ser Leu Phe Thr Val Pro Lys		
[0125]	85	90	95
[0126]	Thr Cys Ile Leu Gly Tyr Asn Asn Val Arg Phe Asp Asp Glu Val Thr		
[0127]	100	105	110
[0128]	Arg Asn Ile Phe Tyr Arg Asn Phe Tyr Asp Pro Tyr Ala Trp Ser Trp		
[0129]	115	120	125
[0130]	Gln His Asp Asn Ser Arg Trp Asp Leu Leu Asp Val Met Arg Ala Cys		
[0131]	130	135	140
[0132]	Tyr Ala Leu Arg Pro Glu Gly Ile Asn Trp Pro Glu Asn Asp Asp Gly		
[0133]	145	150	155
[0134]	Leu Pro Ser Phe Arg Leu Glu His Leu Thr Lys Ala Asn Gly Ile Glu		
[0135]	165	170	175
[0136]	His Ser Asn Ala His Asp Ala Met Ala Asp Val Tyr Ala Thr Ile Ala		
[0137]	180	185	190
[0138]	Met Ala Lys Leu Val Lys Thr Arg Gln Pro Arg Leu Phe Asp Tyr Leu		
[0139]	195	200	205
[0140]	Phe Thr His Arg Asn Lys His Lys Leu Met Ala Leu Ile Asp Val Pro		
[0141]	210	215	220
[0142]	Gln Met Lys Pro Leu Val His Val Ser Gly Met Phe Gly Ala Trp Arg		
[0143]	225	230	235
[0144]	Gly Asn Thr Ser Trp Val Ala Pro Leu Ala Trp His Pro Glu Asn Arg		
[0145]	245	250	255
[0146]	Asn Ala Val Ile Met Val Asp Leu Ala Gly Asp Ile Ser Pro Leu Leu		
[0147]	260	265	270
[0148]	Glu Leu Asp Ser Asp Thr Leu Arg Glu Arg Leu Tyr Thr Ala Lys Thr		
[0149]	275	280	285
[0150]	Asp Leu Gly Asp Asn Ala Ala Val Pro Val Lys Leu Val His Ile Asn		
[0151]	290	295	300
[0152]	Lys Cys Pro Val Leu Ala Gln Ala Asn Thr Leu Arg Pro Glu Asp Ala		
[0153]	305	310	315
[0154]	Asp Arg Leu Gly Ile Asn Arg Gln His Cys Leu Asp Asn Leu Lys Ile		
[0155]	325	330	335

[0156]	Leu Arg Glu Asn Pro Gln Val Arg Glu Lys Val Val Ala Ile Phe Ala
[0157]	340 345 350
[0158]	Glu Ala Glu Pro Phe Thr Pro Ser Asp Asn Val Asp Ala Gln Leu Tyr
[0159]	355 360 365
[0160]	Asn Gly Phe Phe Ser Asp Ala Asp Arg Ala Ala Met Lys Ile Val Leu
[0161]	370 375 380
[0162]	Glu Thr Glu Pro Arg Asn Leu Pro Ala Leu Asp Ile Thr Phe Val Asp
[0163]	385 390 395 400
[0164]	Lys Arg Ile Glu Lys Leu Leu Phe Asn Tyr Arg Ala Arg Asn Phe Pro
[0165]	405 410 415
[0166]	Gly Thr Leu Asp Tyr Ala Glu Gln Gln Arg Trp Leu Glu His Arg Arg
[0167]	420 425 430
[0168]	Gln Val Phe Thr Pro Glu Phe Leu Gln Gly Tyr Ala Asp Glu Leu Gln
[0169]	435 440 445
[0170]	Met Leu Val Gln Gln Tyr Ala Asp Asp Lys Glu Lys Val Ala Leu Leu
[0171]	450 455 460
[0172]	Lys Ala Leu Trp Gln Tyr Ala Glu Glu Ile Val Ser Gly Ser Gly His
[0173]	465 470 475 480
[0174]	His His His His His
[0175]	485
[0176]	<210> 6
[0177]	<211> 268
[0178]	<212> PRT
[0179]	<213> 大肠杆菌
[0180]	<400> 6
[0181]	Met Lys Phe Val Ser Phe Asn Ile Asn Gly Leu Arg Ala Arg Pro His
[0182]	1 5 10 15
[0183]	Gln Leu Glu Ala Ile Val Glu Lys His Gln Pro Asp Val Ile Gly Leu
[0184]	20 25 30
[0185]	Gln Glu Thr Lys Val His Asp Asp Met Phe Pro Leu Glu Glu Val Ala
[0186]	35 40 45
[0187]	Lys Leu Gly Tyr Asn Val Phe Tyr His Gly Gln Lys Gly His Tyr Gly
[0188]	50 55 60
[0189]	Val Ala Leu Leu Thr Lys Glu Thr Pro Ile Ala Val Arg Arg Gly Phe
[0190]	65 70 75 80
[0191]	Pro Gly Asp Asp Glu Glu Ala Gln Arg Arg Ile Ile Met Ala Glu Ile
[0192]	85 90 95
[0193]	Pro Ser Leu Leu Gly Asn Val Thr Val Ile Asn Gly Tyr Phe Pro Gln
[0194]	100 105 110

[0195]	Gly	Glu	Ser	Arg	Asp	His	Pro	Ile	Lys	Phe	Pro	Ala	Lys	Ala	Gln	Phe
[0196]				115				120					125			
[0197]	Tyr	Gln	Asn	Leu	Gln	Asn	Tyr	Leu	Glu	Thr	Glu	Leu	Lys	Arg	Asp	Asn
[0198]			130					135					140			
[0199]	Pro	Val	Leu	Ile	Met	Gly	Asp	Met	Asn	Ile	Ser	Pro	Thr	Asp	Leu	Asp
[0200]	145						150					155				160
[0201]	Ile	Gly	Ile	Gly	Glu	Glu	Asn	Arg	Lys	Arg	Trp	Leu	Arg	Thr	Gly	Lys
[0202]					165					170					175	
[0203]	Cys	Ser	Phe	Leu	Pro	Glu	Glu	Arg	Glu	Trp	Met	Asp	Arg	Leu	Met	Ser
[0204]				180					185					190		
[0205]	Trp	Gly	Leu	Val	Asp	Thr	Phe	Arg	His	Ala	Asn	Pro	Gln	Thr	Ala	Asp
[0206]			195					200					205			
[0207]	Arg	Phe	Ser	Trp	Phe	Asp	Tyr	Arg	Ser	Lys	Gly	Phe	Asp	Asp	Asn	Arg
[0208]		210					215					220				
[0209]	Gly	Leu	Arg	Ile	Asp	Leu	Leu	Leu	Ala	Ser	Gln	Pro	Leu	Ala	Glu	Cys
[0210]	225					230					235					240
[0211]	Cys	Val	Glu	Thr	Gly	Ile	Asp	Tyr	Glu	Ile	Arg	Ser	Met	Glu	Lys	Pro
[0212]					245					250					255	
[0213]	Ser	Asp	His	Ala	Pro	Val	Trp	Ala	Thr	Phe	Arg	Arg				
[0214]				260					265							
[0215]	<210> 7															
[0216]	<211> 666															
[0217]	<212> PRT															
[0218]	<213> 嗜热菌															
[0219]	<400> 7															
[0220]	Met	Arg	Asp	Arg	Val	Arg	Trp	Arg	Val	Leu	Ser	Leu	Pro	Pro	Leu	Ala
[0221]	1				5					10					15	
[0222]	Gln	Trp	Arg	Glu	Val	Met	Ala	Ala	Leu	Glu	Val	Gly	Pro	Glu	Ala	Ala
[0223]				20					25					30		
[0224]	Leu	Ala	Tyr	Trp	His	Arg	Gly	Phe	Arg	Arg	Lys	Glu	Asp	Leu	Asp	Pro
[0225]			35					40					45			
[0226]	Pro	Leu	Ala	Leu	Leu	Pro	Leu	Lys	Gly	Leu	Arg	Glu	Ala	Ala	Ala	Leu
[0227]		50					55					60				
[0228]	Leu	Glu	Glu	Ala	Leu	Arg	Gln	Gly	Lys	Arg	Ile	Arg	Val	His	Gly	Asp
[0229]	65					70					75					80
[0230]	Tyr	Asp	Ala	Asp	Gly	Leu	Thr	Gly	Thr	Ala	Ile	Leu	Val	Arg	Gly	Leu
[0231]					85					90					95	
[0232]	Ala	Ala	Leu	Gly	Ala	Asp	Val	His	Pro	Phe	Ile	Pro	His	Arg	Leu	Glu
[0233]				100					105					110		

[0234]	Glu Gly Tyr Gly Val Leu Met Glu Arg Val Pro Glu His Leu Glu Ala		
[0235]	115	120	125
[0236]	Ser Asp Leu Phe Leu Thr Val Asp Cys Gly Ile Thr Asn His Ala Glu		
[0237]	130	135	140
[0238]	Leu Arg Glu Leu Leu Glu Asn Gly Val Glu Val Ile Val Thr Asp His		
[0239]	145	150	155
[0240]	His Thr Pro Gly Lys Thr Pro Ser Pro Gly Leu Val Val His Pro Ala		
[0241]	165	170	175
[0242]	Leu Thr Pro Asp Leu Lys Glu Lys Pro Thr Gly Ala Gly Val Val Phe		
[0243]	180	185	190
[0244]	Leu Leu Leu Trp Ala Leu His Glu Arg Leu Gly Leu Pro Pro Pro Leu		
[0245]	195	200	205
[0246]	Glu Tyr Ala Asp Leu Ala Ala Val Gly Thr Ile Ala Asp Val Ala Pro		
[0247]	210	215	220
[0248]	Leu Trp Gly Trp Asn Arg Ala Leu Val Lys Glu Gly Leu Ala Arg Ile		
[0249]	225	230	235
[0250]	Pro Ala Ser Ser Trp Val Gly Leu Arg Leu Leu Ala Glu Ala Val Gly		
[0251]	245	250	255
[0252]	Tyr Thr Gly Lys Ala Val Glu Val Ala Phe Arg Ile Ala Pro Arg Ile		
[0253]	260	265	270
[0254]	Asn Ala Ala Ser Arg Leu Gly Glu Ala Glu Lys Ala Leu Arg Leu Leu		
[0255]	275	280	285
[0256]	Leu Thr Asp Asp Ala Ala Glu Ala Gln Ala Leu Val Gly Glu Leu His		
[0257]	290	295	300
[0258]	Arg Leu Asn Ala Arg Arg Gln Thr Leu Glu Glu Ala Met Leu Arg Lys		
[0259]	305	310	315
[0260]	Leu Leu Pro Gln Ala Asp Pro Glu Ala Lys Ala Ile Val Leu Leu Asp		
[0261]	325	330	335
[0262]	Pro Glu Gly His Pro Gly Val Met Gly Ile Val Ala Ser Arg Ile Leu		
[0263]	340	345	350
[0264]	Glu Ala Thr Leu Arg Pro Val Phe Leu Val Ala Gln Gly Lys Gly Thr		
[0265]	355	360	365
[0266]	Val Arg Ser Leu Ala Pro Ile Ser Ala Val Glu Ala Leu Arg Ser Ala		
[0267]	370	375	380
[0268]	Glu Asp Leu Leu Leu Arg Tyr Gly Gly His Lys Glu Ala Ala Gly Phe		
[0269]	385	390	395
[0270]	Ala Met Asp Glu Ala Leu Phe Pro Ala Phe Lys Ala Arg Val Glu Ala		
[0271]	405	410	415
[0272]	Tyr Ala Ala Arg Phe Pro Asp Pro Val Arg Glu Val Ala Leu Leu Asp		

[0273]	420	425	430
[0274]	Leu Leu Pro Glu Pro Gly Leu Leu Pro Gln Val Phe Arg Glu Leu Ala		
[0275]	435	440	445
[0276]	Leu Leu Glu Pro Tyr Gly Glu Gly Asn Pro Glu Pro Leu Phe Leu Leu		
[0277]	450	455	460
[0278]	Phe Gly Ala Pro Glu Glu Ala Arg Arg Leu Gly Glu Gly Arg His Leu		
[0279]	465	470	475
[0280]	Ala Phe Arg Leu Lys Gly Val Arg Val Leu Ala Trp Lys Gln Gly Asp		
[0281]	485	490	495
[0282]	Leu Ala Leu Pro Pro Glu Val Glu Val Ala Gly Leu Leu Ser Glu Asn		
[0283]	500	505	510
[0284]	Ala Trp Asn Gly His Leu Ala Tyr Glu Val Gln Ala Val Asp Leu Arg		
[0285]	515	520	525
[0286]	Lys Pro Glu Ala Leu Glu Gly Gly Ile Ala Pro Phe Ala Tyr Pro Leu		
[0287]	530	535	540
[0288]	Pro Leu Leu Glu Ala Leu Ala Arg Ala Arg Leu Gly Glu Gly Val Tyr		
[0289]	545	550	555
[0290]	Val Pro Glu Asp Asn Pro Glu Gly Leu Asp Tyr Ala Arg Lys Ala Gly		
[0291]	565	570	575
[0292]	Phe Arg Leu Leu Pro Pro Glu Glu Ala Gly Leu Trp Leu Gly Leu Pro		
[0293]	580	585	590
[0294]	Pro Arg Pro Val Leu Gly Arg Arg Val Glu Val Ala Leu Gly Arg Glu		
[0295]	595	600	605
[0296]	Ala Arg Ala Arg Leu Ser Ala Pro Pro Val Leu His Thr Pro Glu Ala		
[0297]	610	615	620
[0298]	Arg Leu Lys Ala Leu Val His Arg Arg Leu Leu Phe Ala Tyr Glu Arg		
[0299]	625	630	635
[0300]	Arg His Pro Gly Leu Phe Ser Glu Ala Leu Leu Ala Tyr Trp Glu Val		
[0301]	645	650	655
[0302]	Asn Arg Val Gln Glu Pro Ala Gly Ser Pro		
[0303]	660	665	
[0304]	<210> 8		
[0305]	<211> 226		
[0306]	<212> PRT		
[0307]	<213> λ噬菌体		
[0308]	<400> 8		
[0309]	Met Thr Pro Asp Ile Ile Leu Gln Arg Thr Gly Ile Asp Val Arg Ala		
[0310]	1	5	10
[0311]	Val Glu Gln Gly Asp Asp Ala Trp His Lys Leu Arg Leu Gly Val Ile		15

[0312]	20	25	30
[0313]	Thr Ala Ser Glu Val His Asn Val Ile Ala Lys Pro Arg Ser Gly Lys		
[0314]	35	40	45
[0315]	Lys Trp Pro Asp Met Lys Met Ser Tyr Phe His Thr Leu Leu Ala Glu		
[0316]	50	55	60
[0317]	Val Cys Thr Gly Val Ala Pro Glu Val Asn Ala Lys Ala Leu Ala Trp		
[0318]	65	70	75
[0319]	Gly Lys Gln Tyr Glu Asn Asp Ala Arg Thr Leu Phe Glu Phe Thr Ser		
[0320]	85	90	95
[0321]	Gly Val Asn Val Thr Glu Ser Pro Ile Ile Tyr Arg Asp Glu Ser Met		
[0322]	100	105	110
[0323]	Arg Thr Ala Cys Ser Pro Asp Gly Leu Cys Ser Asp Gly Asn Gly Leu		
[0324]	115	120	125
[0325]	Glu Leu Lys Cys Pro Phe Thr Ser Arg Asp Phe Met Lys Phe Arg Leu		
[0326]	130	135	140
[0327]	Gly Gly Phe Glu Ala Ile Lys Ser Ala Tyr Met Ala Gln Val Gln Tyr		
[0328]	145	150	155
[0329]	Ser Met Trp Val Thr Arg Lys Asn Ala Trp Tyr Phe Ala Asn Tyr Asp		
[0330]	165	170	175
[0331]	Pro Arg Met Lys Arg Glu Gly Leu His Tyr Val Val Ile Glu Arg Asp		
[0332]	180	185	190
[0333]	Glu Lys Tyr Met Ala Ser Phe Asp Glu Ile Val Pro Glu Phe Ile Glu		
[0334]	195	200	205
[0335]	Lys Met Asp Glu Ala Leu Ala Glu Ile Gly Phe Val Phe Gly Glu Gln		
[0336]	210	215	220
[0337]	Trp Arg		
[0338]	225		
[0339]	<210> 9		
[0340]	<211> 608		
[0341]	<212> PRT		
[0342]	<213> 噬菌体 phi-29		
[0343]	<400> 9		
[0344]	Met Lys His Met Pro Arg Lys Met Tyr Ser Cys Ala Phe Glu Thr Thr		
[0345]	1	5	10
[0346]	Thr Lys Val Glu Asp Cys Arg Val Trp Ala Tyr Gly Tyr Met Asn Ile		
[0347]	20	25	30
[0348]	Glu Asp His Ser Glu Tyr Lys Ile Gly Asn Ser Leu Asp Glu Phe Met		
[0349]	35	40	45
[0350]	Ala Trp Val Leu Lys Val Gln Ala Asp Leu Tyr Phe His Asn Leu Lys		

[0351]	50	55	60
[0352]	Phe Asp Gly Ala Phe Ile Ile Asn Trp Leu Glu Arg Asn Gly Phe Lys		
[0353]	65	70	75
[0354]	Trp Ser Ala Asp Gly Leu Pro Asn Thr Tyr Asn Thr Ile Ile Ser Arg		
[0355]	85	90	95
[0356]	Met Gly Gln Trp Tyr Met Ile Asp Ile Cys Leu Gly Tyr Lys Gly Lys		
[0357]	100	105	110
[0358]	Arg Lys Ile His Thr Val Ile Tyr Asp Ser Leu Lys Lys Leu Pro Phe		
[0359]	115	120	125
[0360]	Pro Val Lys Lys Ile Ala Lys Asp Phe Lys Leu Thr Val Leu Lys Gly		
[0361]	130	135	140
[0362]	Asp Ile Asp Tyr His Lys Glu Arg Pro Val Gly Tyr Lys Ile Thr Pro		
[0363]	145	150	155
[0364]	Glu Glu Tyr Ala Tyr Ile Lys Asn Asp Ile Gln Ile Ile Ala Glu Ala		
[0365]	165	170	175
[0366]	Leu Leu Ile Gln Phe Lys Gln Gly Leu Asp Arg Met Thr Ala Gly Ser		
[0367]	180	185	190
[0368]	Asp Ser Leu Lys Gly Phe Lys Asp Ile Ile Thr Thr Lys Lys Phe Lys		
[0369]	195	200	205
[0370]	Lys Val Phe Pro Thr Leu Ser Leu Gly Leu Asp Lys Glu Val Arg Tyr		
[0371]	210	215	220
[0372]	Ala Tyr Arg Gly Gly Phe Thr Trp Leu Asn Asp Arg Phe Lys Glu Lys		
[0373]	225	230	235
[0374]	Glu Ile Gly Glu Gly Met Val Phe Asp Val Asn Ser Leu Tyr Pro Ala		
[0375]	245	250	255
[0376]	Gln Met Tyr Ser Arg Leu Leu Pro Tyr Gly Glu Pro Ile Val Phe Glu		
[0377]	260	265	270
[0378]	Gly Lys Tyr Val Trp Asp Glu Asp Tyr Pro Leu His Ile Gln His Ile		
[0379]	275	280	285
[0380]	Arg Cys Glu Phe Glu Leu Lys Glu Gly Tyr Ile Pro Thr Ile Gln Ile		
[0381]	290	295	300
[0382]	Lys Arg Ser Arg Phe Tyr Lys Gly Asn Glu Tyr Leu Lys Ser Ser Gly		
[0383]	305	310	315
[0384]	Gly Glu Ile Ala Asp Leu Trp Leu Ser Asn Val Asp Leu Glu Leu Met		
[0385]	325	330	335
[0386]	Lys Glu His Tyr Asp Leu Tyr Asn Val Glu Tyr Ile Ser Gly Leu Lys		
[0387]	340	345	350
[0388]	Phe Lys Ala Thr Thr Gly Leu Phe Lys Asp Phe Ile Asp Lys Trp Thr		
[0389]	355	360	365

[0390]	Tyr	Ile	Lys	Thr	Thr	Ser	Glu	Gly	Ala	Ile	Lys	Gln	Leu	Ala	Lys	Leu
[0391]		370					375					380				
[0392]	Met	Leu	Asn	Ser	Leu	Tyr	Gly	Lys	Phe	Ala	Ser	Asn	Pro	Asp	Val	Thr
[0393]	385					390					395				400	
[0394]	Gly	Lys	Val	Pro	Tyr	Leu	Lys	Glu	Asn	Gly	Ala	Leu	Gly	Phe	Arg	Leu
[0395]					405					410					415	
[0396]	Gly	Glu	Glu	Glu	Thr	Lys	Asp	Pro	Val	Tyr	Thr	Pro	Met	Gly	Val	Phe
[0397]				420					425					430		
[0398]	Ile	Thr	Ala	Trp	Ala	Arg	Tyr	Thr	Thr	Ile	Thr	Ala	Ala	Gln	Ala	Cys
[0399]			435					440					445			
[0400]	Tyr	Asp	Arg	Ile	Ile	Tyr	Cys	Asp	Thr	Asp	Ser	Ile	His	Leu	Thr	Gly
[0401]		450					455					460				
[0402]	Thr	Glu	Ile	Pro	Asp	Val	Ile	Lys	Asp	Ile	Val	Asp	Pro	Lys	Lys	Leu
[0403]	465					470					475				480	
[0404]	Gly	Tyr	Trp	Ala	His	Glu	Ser	Thr	Phe	Lys	Arg	Ala	Lys	Tyr	Leu	Arg
[0405]				485					490						495	
[0406]	Gln	Lys	Thr	Tyr	Ile	Gln	Asp	Ile	Tyr	Met	Lys	Glu	Val	Asp	Gly	Lys
[0407]			500						505					510		
[0408]	Leu	Val	Glu	Gly	Ser	Pro	Asp	Asp	Tyr	Thr	Asp	Ile	Lys	Phe	Ser	Val
[0409]			515					520					525			
[0410]	Lys	Cys	Ala	Gly	Met	Thr	Asp	Lys	Ile	Lys	Lys	Glu	Val	Thr	Phe	Glu
[0411]		530					535					540				
[0412]	Asn	Phe	Lys	Val	Gly	Phe	Ser	Arg	Lys	Met	Lys	Pro	Lys	Pro	Val	Gln
[0413]	545					550					555				560	
[0414]	Val	Pro	Gly	Gly	Val	Val	Leu	Val	Asp	Asp	Thr	Phe	Thr	Ile	Lys	Ser
[0415]				565					570					575		
[0416]	Gly	Gly	Ser	Ala	Trp	Ser	His	Pro	Gln	Phe	Glu	Lys	Gly	Gly	Gly	Ser
[0417]			580					585					590			
[0418]	Gly	Gly	Gly	Ser	Gly	Gly	Ser	Ala	Trp	Ser	His	Pro	Gln	Phe	Glu	Lys
[0419]			595					600					605			

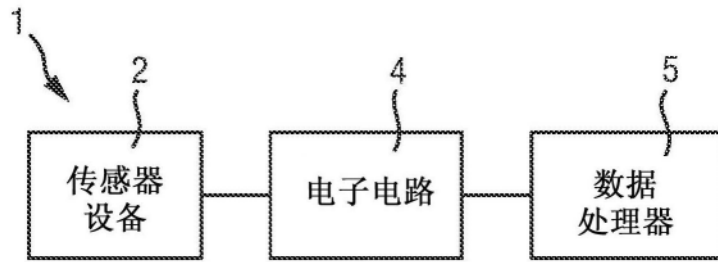


图1

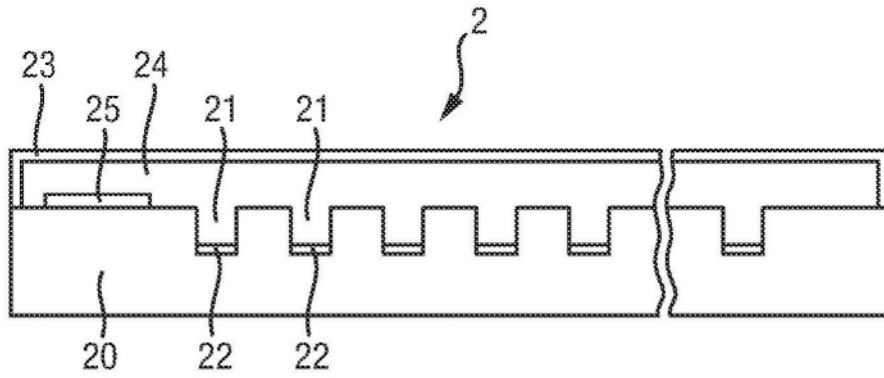


图2

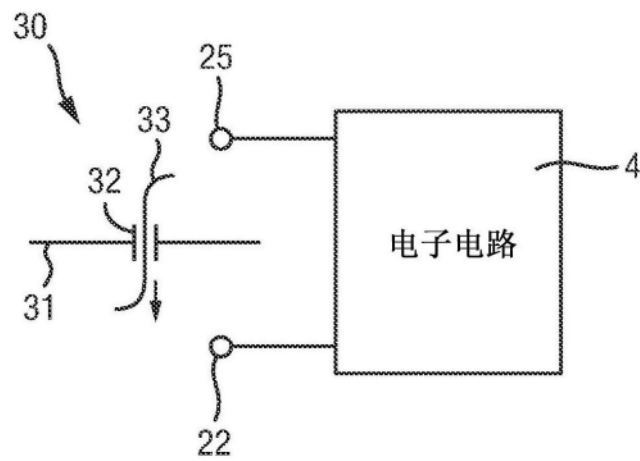


图3

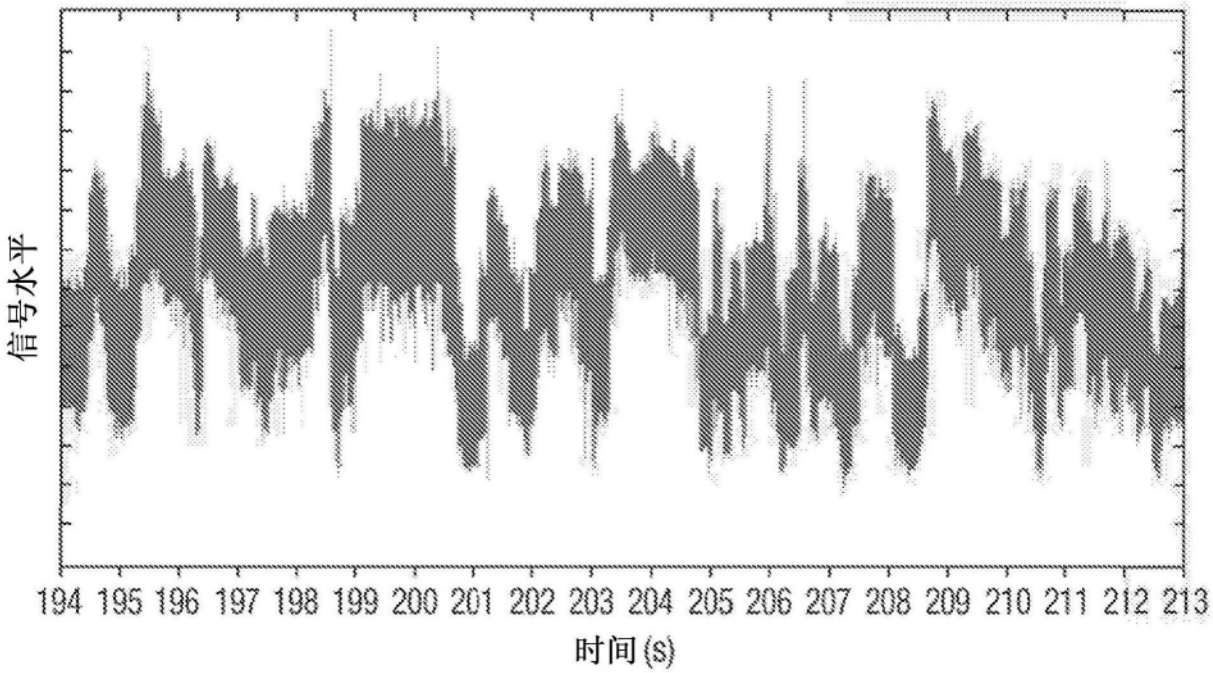


图4

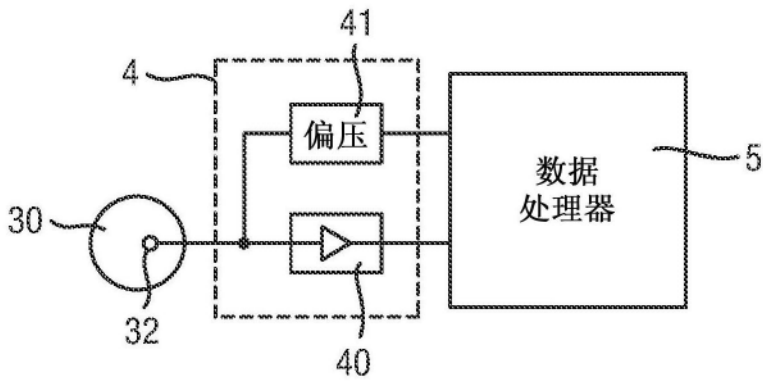


图5

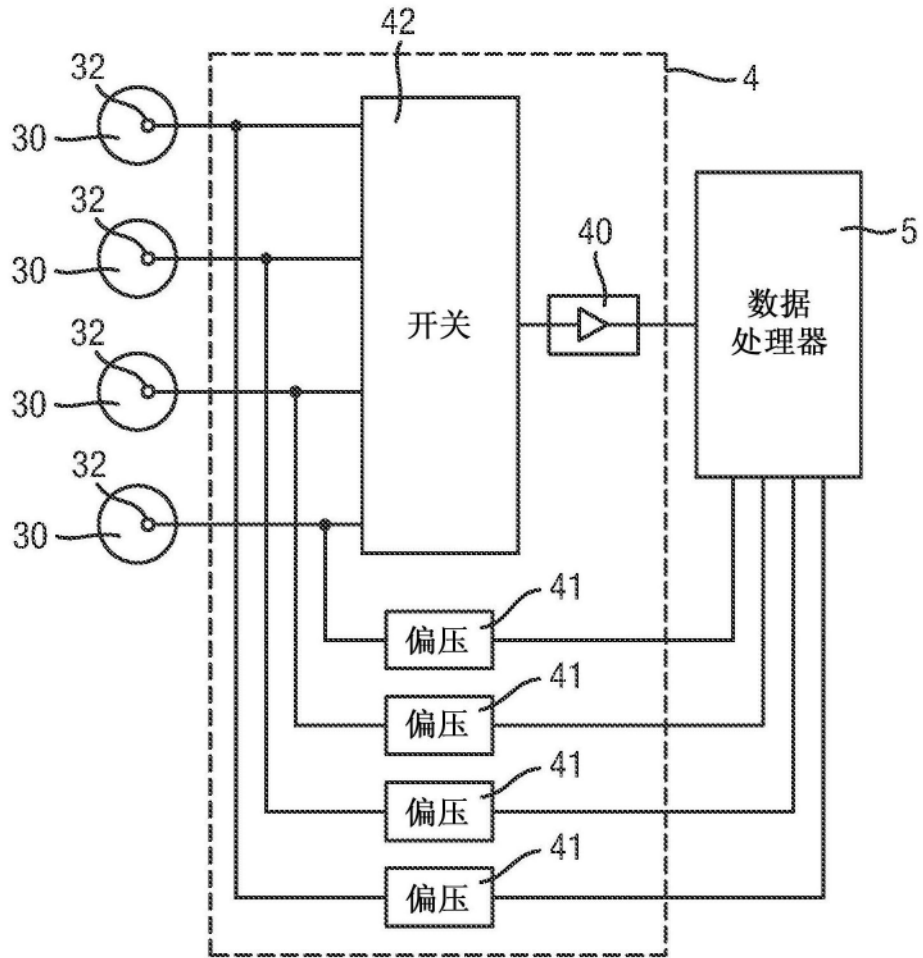


图6

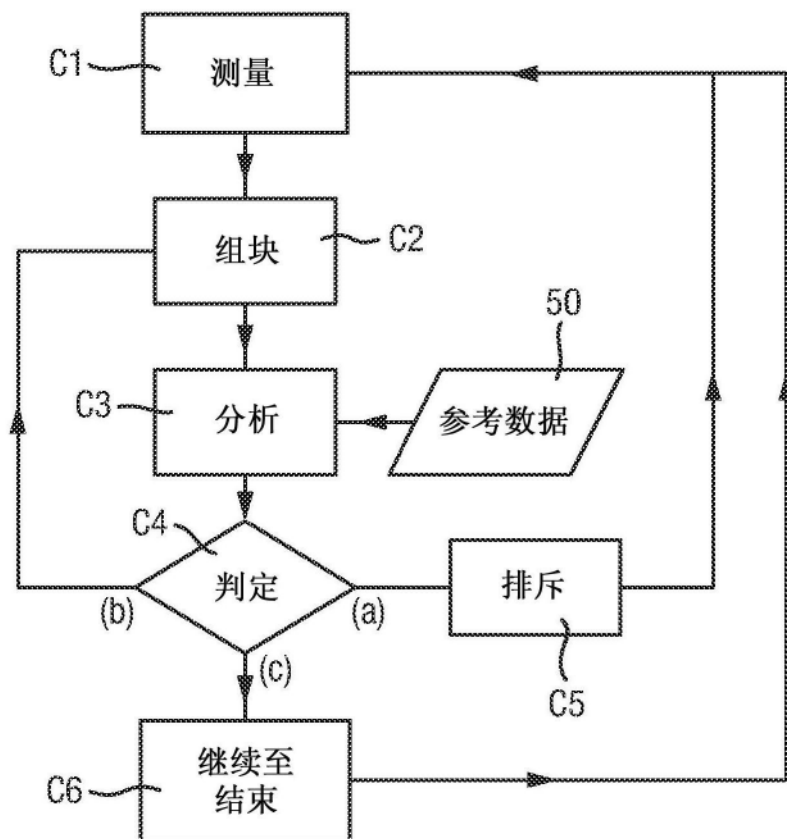


图7

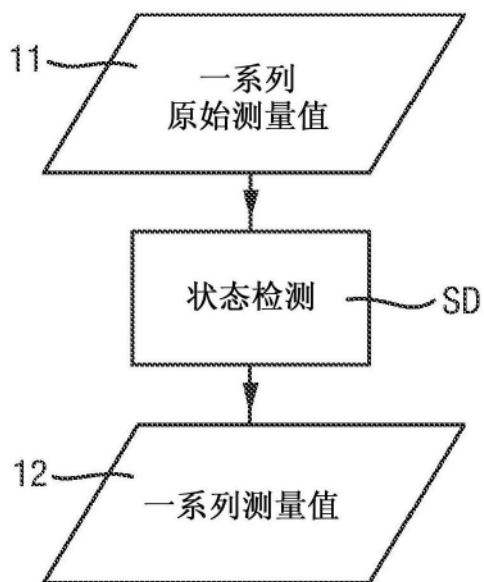


图8

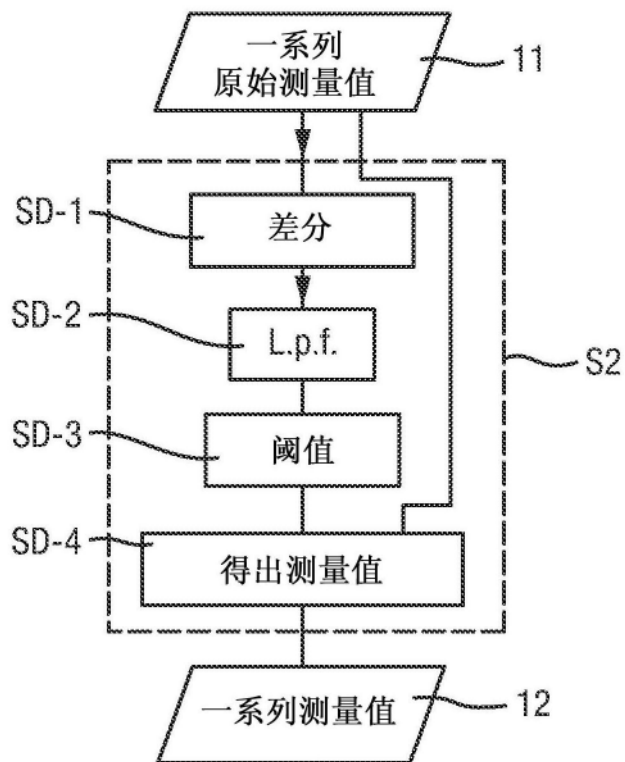


图9

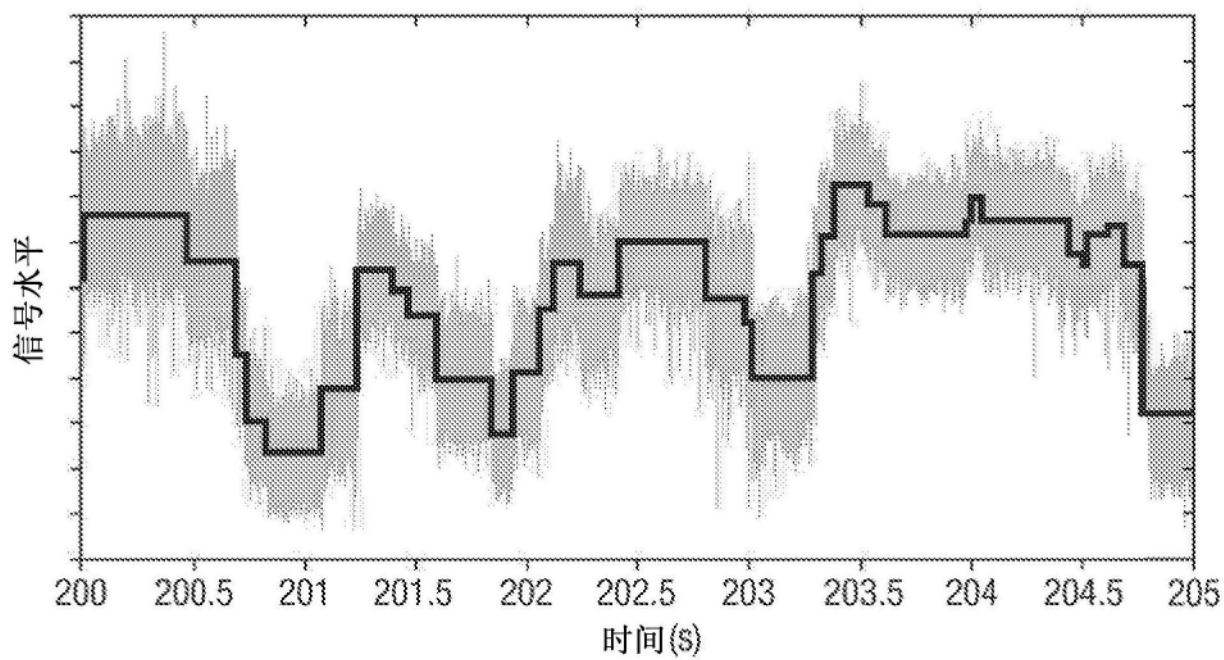


图10

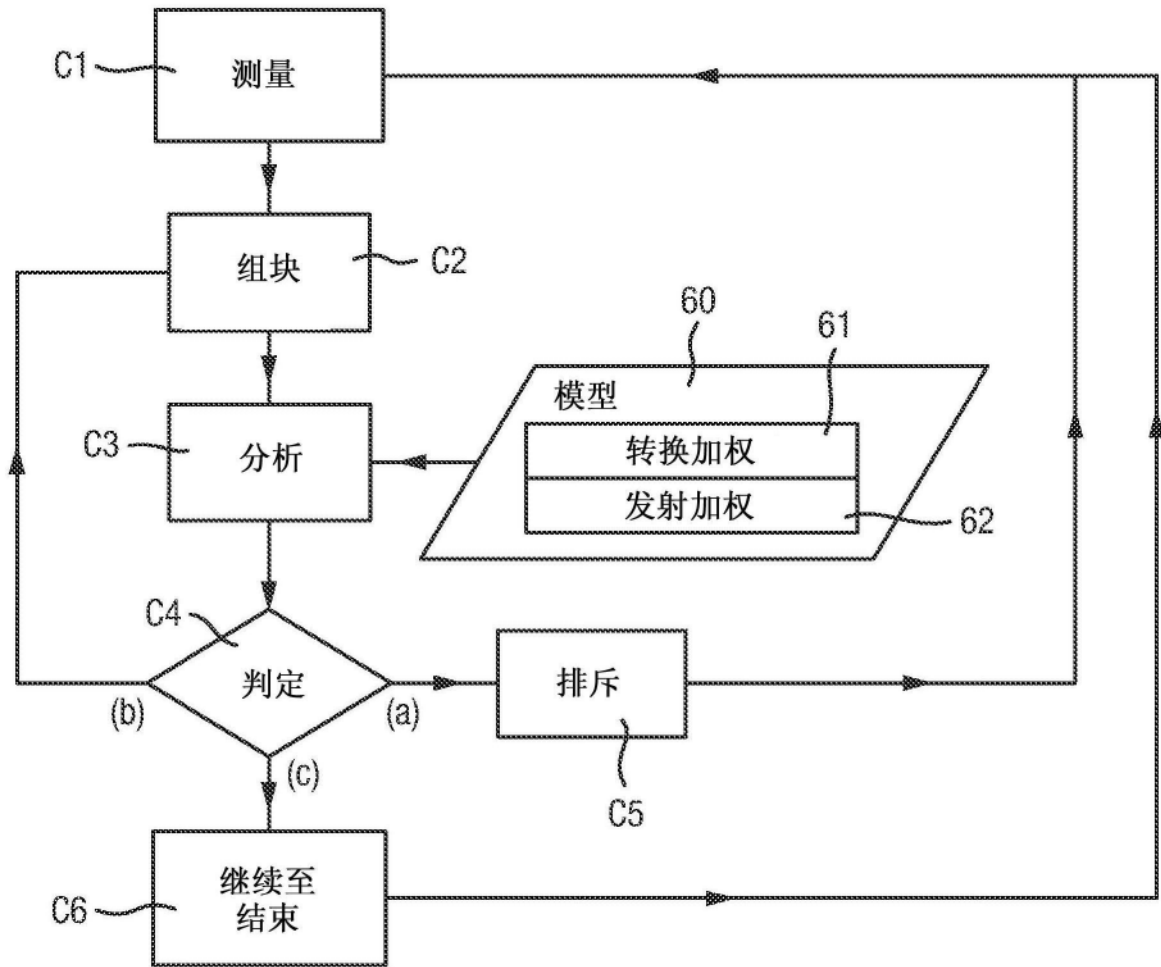


图11

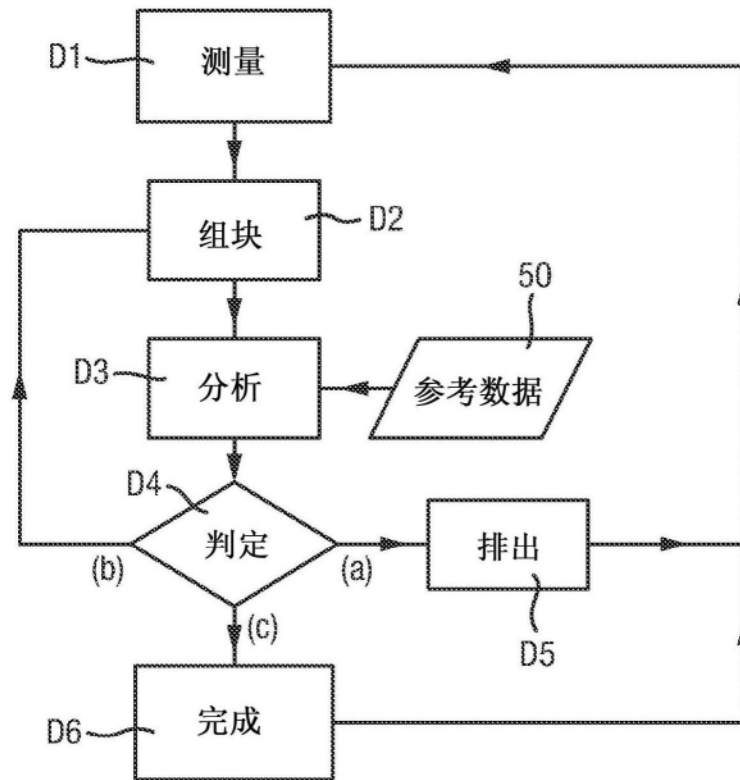


图12

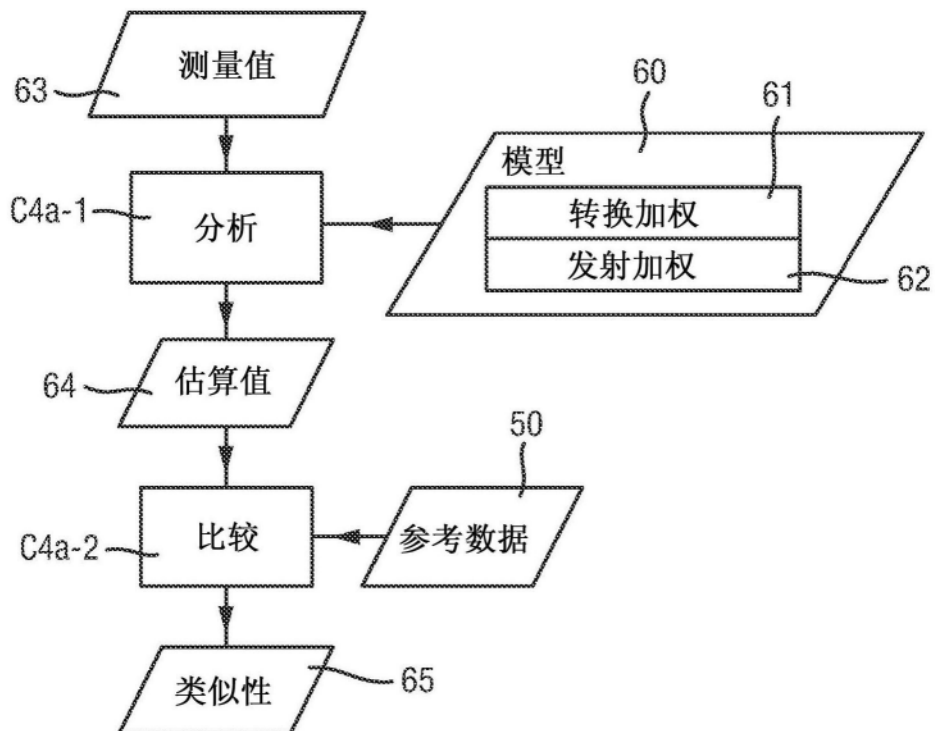


图13

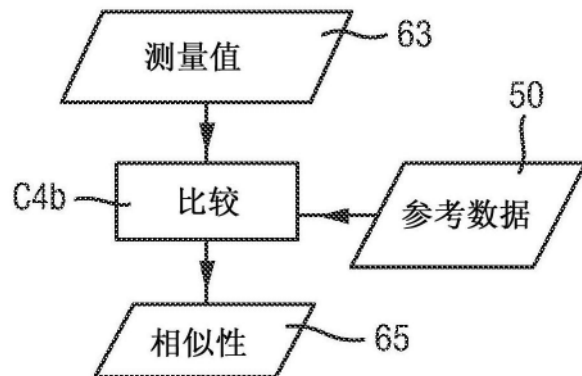


图14

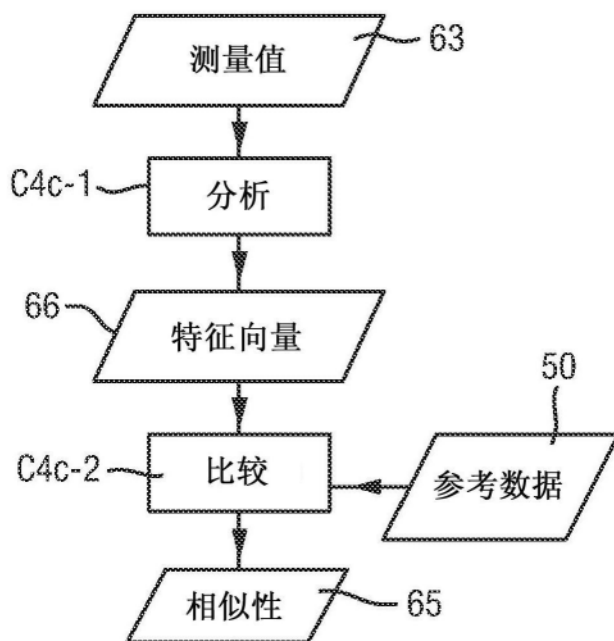


图15

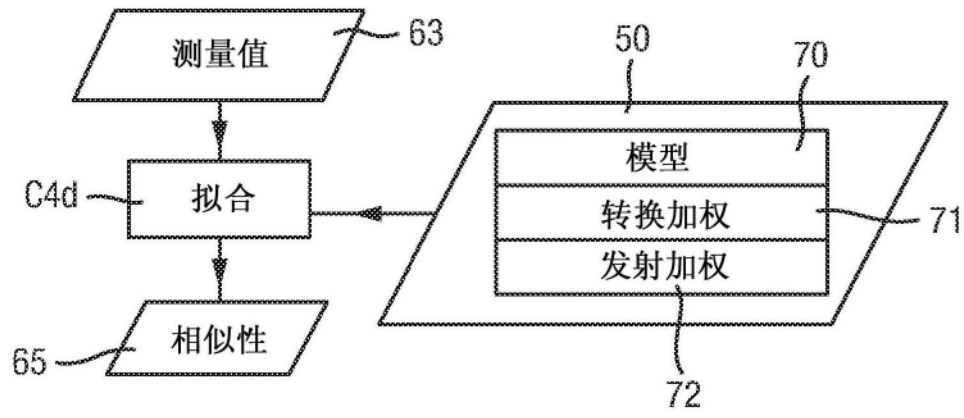


图16

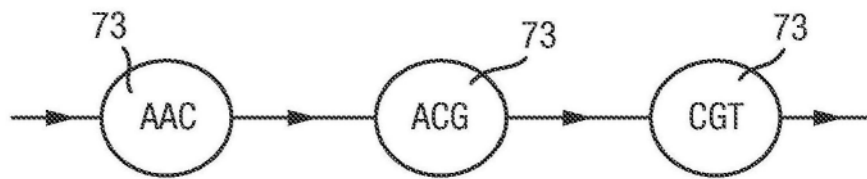


图17

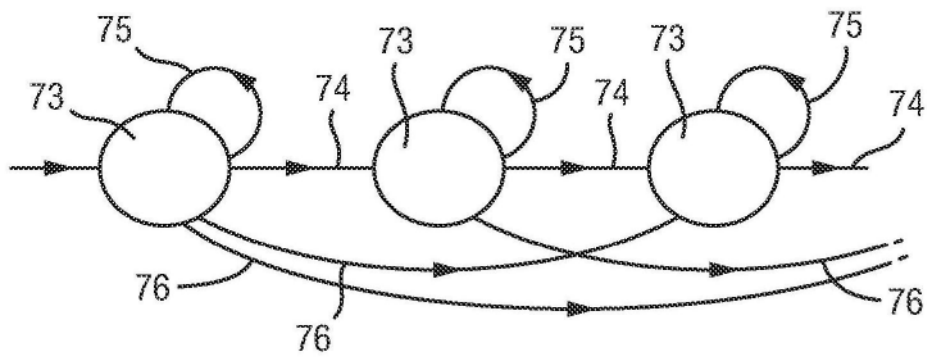


图18

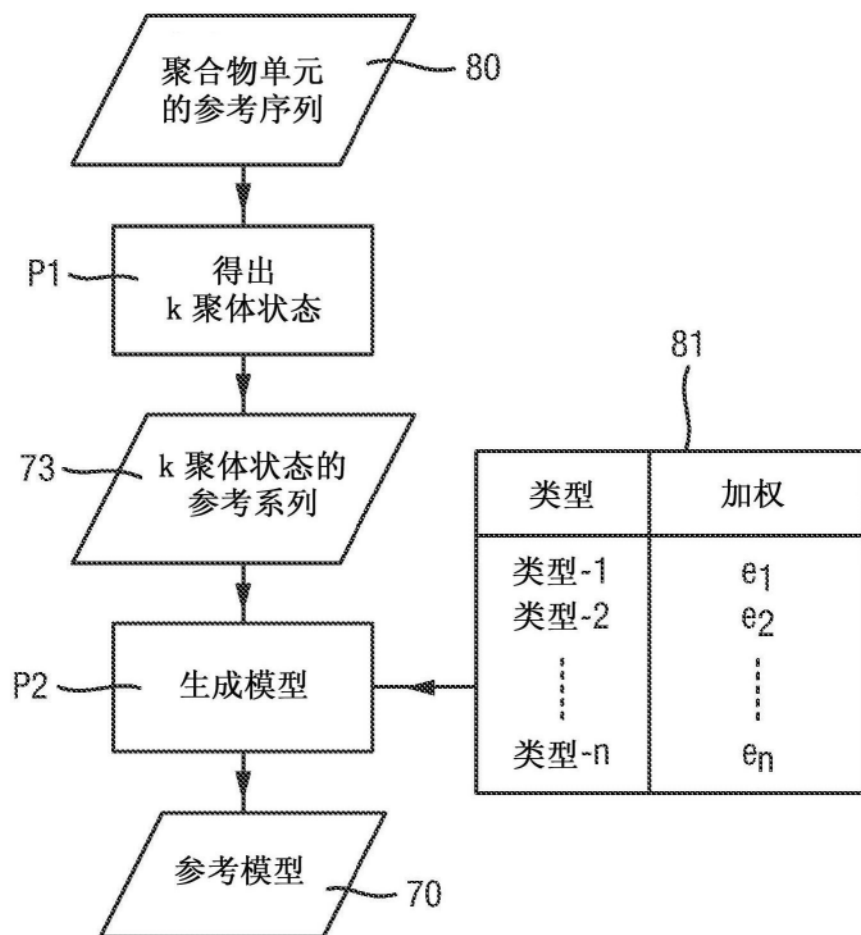


图19

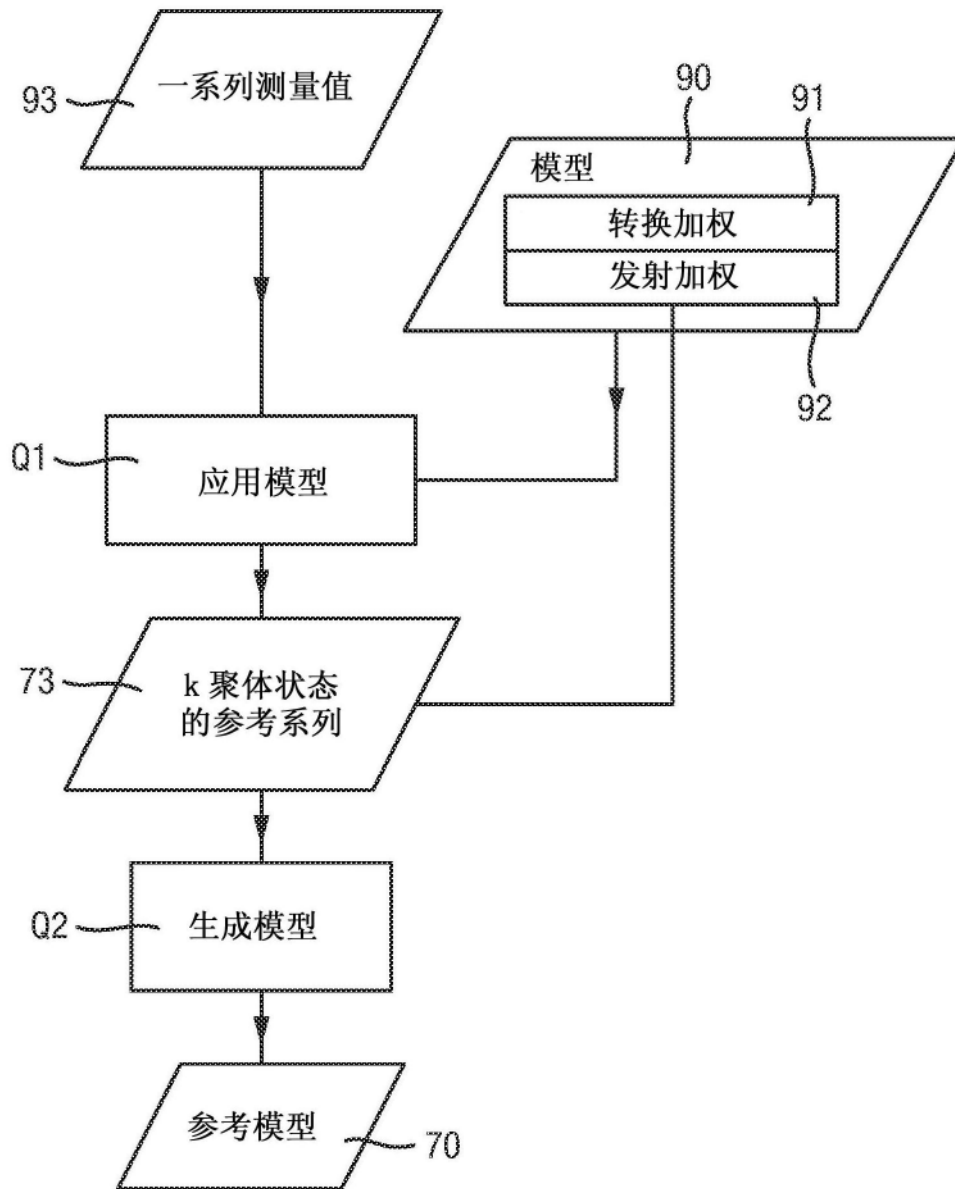


图20

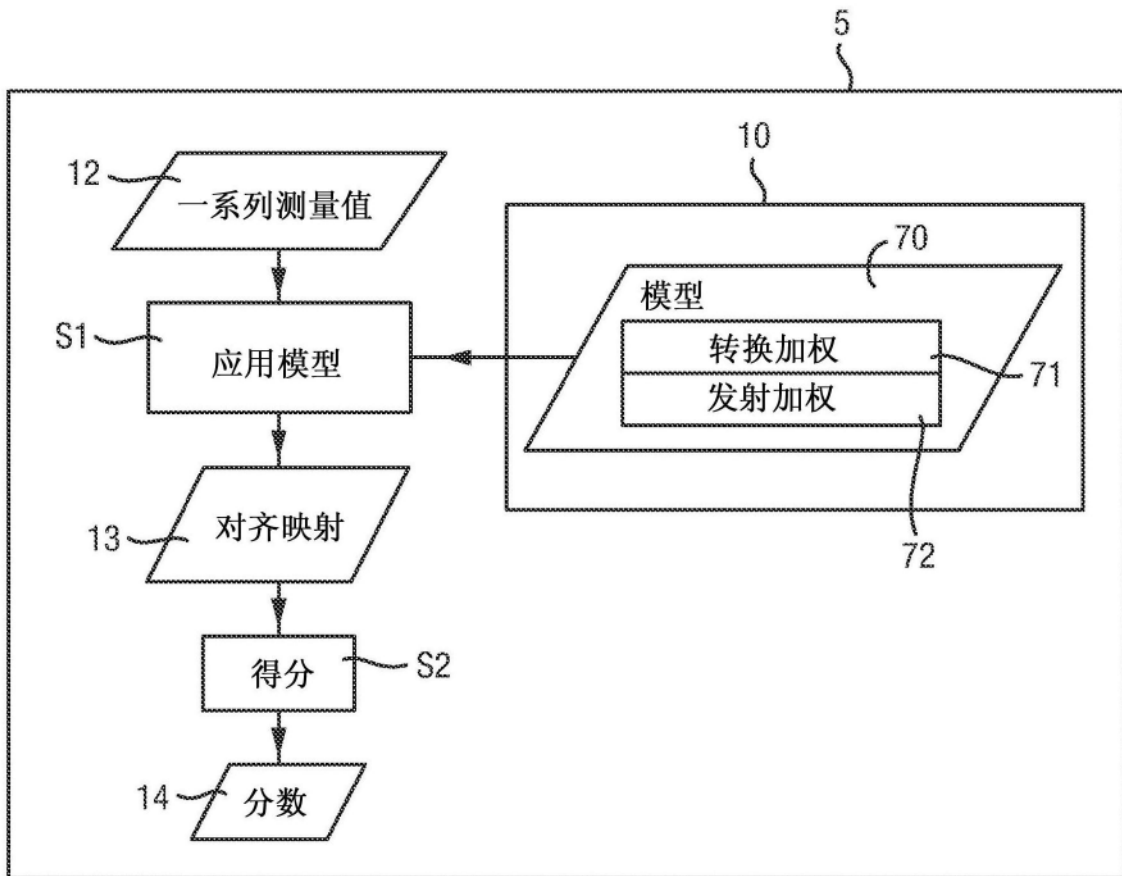


图21

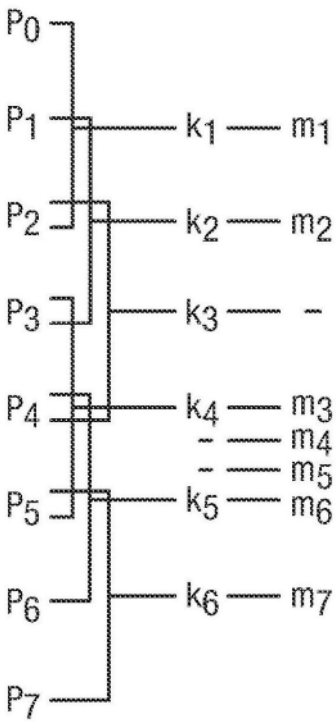


图22