



(12) 发明专利

(10) 授权公告号 CN 101802220 B

(45) 授权公告日 2013. 07. 31

(21) 申请号 200880100371. X

(22) 申请日 2008. 07. 25

(30) 优先权数据

60/962, 036 2007. 07. 26 US

(85) PCT申请进入国家阶段日

2010. 01. 25

(86) PCT申请的申请数据

PCT/US2008/009064 2008. 07. 25

(87) PCT申请的公布数据

W02009/017678 EN 2009. 02. 05

(73) 专利权人 加利福尼亚太平洋生物科学股份
有限公司

地址 美国加利福尼亚州

(72) 发明人 F·克里斯蒂安 S·特纳

(74) 专利代理机构 上海专利商标事务所有限公
司 31100

代理人 韦东

(51) Int. Cl.

C12Q 1/68 (2006. 01)

C07H 21/02 (2006. 01)

(56) 对比文件

US 5268267 A, 1993. 12. 07, 说明书第 4 栏最
后一段, 第 5 栏第 1 段, 第 18 栏, 图 2.

US 6312913 B1, 2001. 11. 06, 摘要, 权利要求
4, 实施例 1.

W0 9907896 A2, 1999. 02. 18, 摘要, 说明书第
6 页, 权利要求 1, 3.

W0 03074734 A2, 2003. 09. 12, 摘要, 权利要
求 2.

审查员 奚静

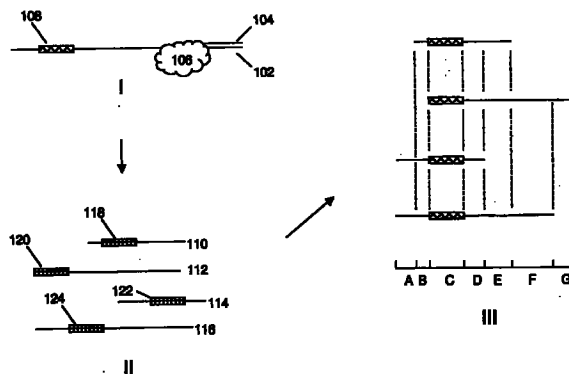
权利要求书2页 说明书10页 附图4页

(54) 发明名称

分子冗余测序法

(57) 摘要

公开了方法、系统和组合物, 其中靶核酸含有
置于其中的登记序列, 该登记序列用于鉴别该靶
核酸的测定序列的数量或相对位置。特别优选的
方面包括环形模板核酸序列内的登记序列, 该环
形模板核酸序列相应地被用于掺入测序法中用于
鉴定该模板的序列, 该方法依赖于模板依赖性聚
合酶介导的引物延伸。



1. 一种鉴别靶核酸的已确定序列的序列前后关系的方法,该方法包括:
在单个靶核酸分子的选定位置提供已知的登记序列;
测定包括该登记序列在内的靶核酸分子的至少一部分序列;
根据该登记序列在该部分序列中的相对位置鉴定该测定步骤中测定的靶核酸分子的该部分序列的序列前后关系。
2. 如权利要求 1 所述的方法,其特征在于,所述测定步骤包括对所述单个核酸分子内的相同核苷酸序列测序多次,且其中相对于该相同核苷酸序列在选定位置提供所述登记序列。
3. 如权利要求 1 所述的方法,其特征在于,所述靶核酸分子包括环形核酸。
4. 如权利要求 2 所述的方法,其特征在于,在所述靶核酸分子内多次提供所述相同核苷酸序列。
5. 如权利要求 3 所述的方法,其特征在于,所述测定步骤包括测定所述靶核酸分子的所述部分序列,其中所述登记序列被测定至少 2 次,所述鉴定序列前后关系的步骤包括确定该完整环形核酸至少已被测序一次。
6. 如权利要求 1 所述的方法,其特征在于,所述靶核酸分子包括线性核酸。
7. 如权利要求 6 所述的方法,其特征在于,在所述靶核酸分子内的多个位点提供所述登记序列。
8. 如权利要求 6 所述的方法,其特征在于,在靠近所述靶核酸分子的一个末端提供所述登记序列。
9. 如权利要求 1 所述的方法,其特征在于,所述测定靶核酸分子的至少部分序列的步骤包括通过在聚合酶介导的模板依赖性引物延伸反应中掺入碱基而测定该序列。
10. 如权利要求 9 所述的方法,其特征在于,所述测定靶核酸分子的至少部分序列的步骤包括通过掺入方法的实时测序。
11. 如权利要求 1 所述的方法,其特征在于,所述登记序列包括 5 个到 100 个碱基。
12. 如权利要求 1 所述的方法,其特征在于,所述登记序列包含 5 个到 50 个碱基。
13. 如权利要求 1 所述的方法,其特征在于,所述登记序列包含 10 个到 20 个碱基。
14. 一种对完整的靶核酸进行测序的方法,其特征在于,所述方法包括:
在该靶核酸序列内提供已知的登记序列;
环化该靶核酸序列;和
采用不破坏该靶核酸序列的测序方法对单个靶核酸分子进行测序,直到该登记序列至少被测序两次。
15. 如权利要求 14 所述的方法,其特征在于,所述测序方法包括在模板导向的聚合酶介导的引物延伸反应中掺入后鉴别碱基,其中,所述靶核酸序列含有该模板。
16. 一种确定核酸序列的方法,其特征在于,所述方法包括:
提供单个模板核酸分子,该模板核酸分子的一部分具有至少第一核苷酸序列和在相对于该第一核苷酸序列的选定位置的登记序列;
对所述单个模板核酸分子进行测序以使第一核苷酸序列和登记序列被测序多次;
至少部分基于测序步骤所鉴别的登记序列比对测序步骤所得的序列信息;和
确定比对步骤所得的共有序列,以确定所述第一核苷酸序列的核酸序列。

17. 如权利要求 16 所述的方法,其特征在于,所述模板核酸是环形核酸,所述对该第一核苷酸序列测序多次的步骤包括对该环形核酸测序多次。

18. 如权利要求 16 所述的方法,其特征在于,所述第一核苷酸序列在所述模板核酸分子中存在多次。

19. 如权利要求 16 所述的方法,其特征在于,所述第一核苷酸序列被测序至少 5 次。

20. 如权利要求 16 所述的方法,其特征在于,所述第一核苷酸序列被测序至少 10 次。

21. 如权利要求 16 所述的方法,其特征在于,所述第一核苷酸序列被测序至少 20 次。

22. 如权利要求 16 所述的方法,其特征在于,采用计算机实施所述比对步骤。

分子冗余测序法

[0001] 相关申请的交叉参考

[0002] 本申请要求 2007 年 7 月 26 日提交的临时专利申请 60/962,036 作为优先权,出于所有目的,该申请的全部公开内容以引用的方式纳入本文。

[0003] 联邦资助研究声明

[0004] 不适用。

[0005] 发明背景

[0006] 遗传分析是生物学研究中的关键手段,且很快就成为药理学甚至医学诊断领域中不可或缺的手段。已使用各种新旧技术实施这类遗传分析,尤其是用于遗传物质的较大片段的核苷酸序列分析的鉴定。

[0007] 但是,总的来说,原始的遗传序列数据在总分析中一样重要,就好比写的小说中使用的字母串一样。虽然字母的顺序是重要的,但是是字词、句子、段落和章节中的内容传递最有用的信息。类似地,虽然纯粹的核苷酸序列信息对于遗传分析而言是非常重要的,但是密码子、基因、基因簇、染色体和全基因组中的序列信息传递更多数量的信息。

[0008] 除了序列前后关系(context)外,大多数的常规测序技术基于成群核酸的分析,因此,这些技术从核酸混合物的大量分析中获得共有序列。虽然这种方法能有效获得全部的共有序列,但是它忽略了分子间的差异,而这些差异对于各种不同应用可能尤为重要。相反,单分子测序方法则可能出现大量共有序列方法(bulkconsensus method)中不明显的不准确性。

[0009] 本发明总的涉及提供单个核酸分子上的冗余序列信息的方法和系统,所述方法和系统可用于提高精确测定以及在测序方法中确定序列前后关系信息(contextinformation)。下文将更详细地描述本发明的这些和其它方面。

发明内容

[0010] 本发明总的涉及用于重复分析单个核酸分子的方法和系统,以提供这些单个分子的高度准确的序列信息。具体而言,本发明提供改进的方法、系统和组合物,它们可用于实施单种聚合物尤其是核酸聚合物的冗余序列分析。

[0011] 第一方面,本发明提供鉴别靶核酸的确定序列的序列前后关系的方法。该方法包括提供靶核酸序列中选定位置的已知登记序列(registration sequence),测定包括该登记序列在内的所述靶核酸的至少一部分序列,和由该序列的该部分中的该登记序列的相对位置鉴别在该测定步骤中测定的靶核酸的该部分序列的序列前后关系。

[0012] 另一方面,本发明提供了给完整的靶核酸测序的方法,该方法包括提供靶核酸序列内的已知登记序列。将该靶核酸环化(circularize),使用不破坏该靶核酸序列的测序方法对该靶核酸测序,直到该登记序列至少已被测序两次。

[0013] 又一方面,本发明提供了测定核酸序列的方法,该方法包括提供模板核酸,该模板核酸的一部分具有至少第一核苷酸序列和相对于该第一核苷酸序列处于选定位置的登记序列。该第一核苷酸序列和登记序列被测序多次,至少部分基于测序步骤所鉴别的该登记

序列比对测序步骤所得的序列信息。然后从比对步骤中确定共有序列,从而确定该第一核苷酸序列的核酸序列。

[0014] 在相关方面,本发明提供用于实施前述方法的组合物。这些组合物通常含有核酸合成复合物,该复合物含有至少第一外源登记序列的第一模板核酸序列、核酸聚合酶、和与该第一模板核酸序列的至少一部分互补的引物序列。该模板核酸序列针对该核酸聚合酶设计,用于在模板核酸序列的相同核苷酸序列上实施多次引物延伸反应。该组合物通常还包含多种类型的核苷酸或核苷酸类似物。

附图说明

[0015] 图 1 示意性阐述本发明方法中在使用测序方法获得的序列信息输出的比对中使用登记序列。

[0016] 图 2 示意性阐述本发明用于环形模板序列获得的序列信息的方法。

[0017] 图 3 示意性阐述本发明方法中共有序列的测定。

[0018] 图 4 示意性阐述使用登记序列的另外一种序列前后关系确定方法。

[0019] 发明详述

[0020] I. 总论

[0021] 本发明总的涉及通过单个核酸分子的冗余测序来测定靶核酸序列信息的改进方法。通过重复测定相同靶序列或其部分的序列,人们可极大地提高对得自该方法的序列信息的置信水平。在单分子测序方法中使用冗余序列分析在出版的美国专利申请 2006-0063264 中有描述,出于所有目的,本文将全文以引用的方式纳入。

[0022] 虽然本文所述的冗余测序方法将会在各种测序方法中有广泛的应用,但是,应理解,在特别优选的方面,这些方法用于单分子测序方法。美国专利 7,033,764、7,052,847、7,056,661 和 7,056,676 描述了单分子(测序)方法的例子,出于所有目的,本文将这些申请的全文以引用的方式纳入。

[0023] 简言之,提供靶核酸或模板核酸并对其设计,这样,单分子模板依赖性测序反应将可在该序列中的相同核苷酸序列上进行多次。模板内相同的核苷酸序列的重复测序通过提供相同碱基序列的冗余序列分析而提高该方法获得的序列信息的置信水平。举例而言,当给定测序方法中存在潜在的序列鉴定或测定相关的误差水平时,对给定序列的单次测序将具有该潜在的错误,这是对所测得的序列的置信水平的限制。换言之,如果测序方法具有 10% 的误差比例,则人们在测定该序列的任何碱基时将仅可获得 90% 的置信水平。但是,通过重复测定相同序列的序列,比较每一次测序的序列信息,人们可系统性地减少该误差水平。因此,对于每一次测序,与多次测序所得的碱基的鉴定相关的误差比例将极大下降。

[0024] 除了降低任何测序方法中固有的不准确性外,单核酸分子的冗余测序在鉴别低拷贝数量环境中的遗传变异(如寻找分子间的变异)也有非常高的价值,所述变异可能无法从较大的核酸样品中鉴定。例如,某些遗传异常可能存在于给定样品的相对小的细胞亚组内,或者可能仅仅是单个细胞内的小亚组的遗传物质。在这些情况下,基于广泛选择遗传物质的采样可能遗漏掉变异材料和正常材料之间的任何不同。

[0025] 根据本申请公开的内容将理解,本发明的方法可依赖于各种实施形态

(configuration), 以实现本文所述的冗余序列测定。这些实施形态包括单模板分子内给定序列的多个拷贝的测序、通过单分子复合物实施的给定模板序列内相同核苷酸组的冗余测序、和使用多种不同的分子复合物实施的一条或多条相同序列的测序, 以及这些测序的任意组合。

[0026] 除了核苷酸序列的冗余测序方法外, 本发明还提供一些有助于这些冗余测序方法的改进, 以及评估和处理从这些和其它相关方法获得的冗余序列信息的方法。

[0027] 除了上文所述, 本发明的一个目的是提供本文所述的组合物, 例如包括核酸合成组合物, 该组合物包括含有登记序列的本发明模板序列、核酸聚合酶、引物序列、以及多种类型的核苷酸和 / 或核苷酸类似物, 所述核苷酸和 / 或核苷酸类似物用于实施引物延伸, 优选的用于核酸测序操作, 如使用可检测的类似物如荧光标记的核苷酸或核苷酸类似物。同样地, 本发明也包括使用这些组合物和实施这些方法的系统。

[0028] II. 冗余序列

[0029] 如前文所述, 本发明提供对模板序列分子内给定的核苷酸序列实施多次冗余测序。根据本发明, 对模板分子内的给定序列的这种冗余测序可使用各种不同实施方式。例如, 在至少一个方面, 待冗余测序的核苷酸的序列存在于环形模板分子内, 测序方法围绕着该环形模板重复实施。应理解, 当使用环形模板时, 围绕该环的第一次周转之后新生链的移走将允许连续的或重复的测序。在其最简单的实施方式中, 通过在测序过程中使用取代聚合酶 (displacing polymerase enzyme) 而实现上述“移走”。已描述了各种链取代聚合酶, 这些酶尤其可用于结合入测序中 (见, 例如国际专利申请 WO 2007/075987、WO2007/075873、WO2007/076057, 和它们的美国同族申请, 出于所有目的, 本文将这些文献的全文以引用的方式纳入)。简言之, 在围绕着环形模板完成单次周转后, 聚合酶将移走新合成的新生链, 以继续围绕该模板进行合成。

[0030] 虽然在涉及环形模板时使用链取代聚合酶是本发明的特别优选的实施方式, 但是也可使用大量的其它方法来移走新生链。例如, 可使用特定的核酸外切酶来消化开口的新生链, 以从基础模板除去该新生链, 从而使该合成方法得以继续进行。在任何一种情况中, 当需要对特定分子中的相同核苷酸组重复测序时, 应理解用于测序的方法对于该模板核酸序列是没有破坏性的。举例而言, 在这些方面, 消化性的测序方法不是优选的。类似的, 需要重复洗涤步骤以使掺入的和 / 或标记的核苷酸或终止子核苷酸去保护和 / 或去标记的测序方法往往在少数轮次后降解该模板链, 因而不包括在本发明的使用不破坏序列的工艺的方法中。

[0031] 在其它方面, 可通过在单模板链内提供多个拷贝的靶序列而对相同序列的核苷酸或靶序列测序。如上所述, 模板可以是环形的, 以提供额外层的冗余度, 或者模板可含有线性模板。在使用具有多拷贝的核苷酸序列的线性模板中, 应理解可对相同的核苷酸序列测序而不需要将链移走或除去, 因为在该模板上的单次通过将提供这样的冗余序列读数。相应地, 可从单一来源的模板或靶序列 (例如, 通过滚环过程 (rolling circle process) 环化和复制) 制备线性模板, 以提供这些靶序列的重复模式。

[0032] 在另一方面, 通过在给定引导位点重新启动模板的引导, 然后重复测定该模板相同部分的序列, 这样就可对核苷酸的序列进行重复测序。最简单的重新启动包括洗涤已固定的模板, 以除去先前的聚合反应复合物, 和重新加入新鲜的聚合酶引物混合物以重新

启动合成,从而重新开始测序过程。已固定的模板的洗涤可包括调整反应混合物的温度、pH 和 / 或盐浓度中一种或多种,以使聚合酶从模板上解离。同样,也可使聚合酶合成至 (synthesize off) 线性模板的末端而完成单模板分子的测序。之后,通过调整杂交条件,如增加温度和 / 或盐浓度使新生链从模板解离,从而实现新生链 / 引物链的去除。

[0033] 在还有其它方面,可使用例如处于基底(如显微镜载玻片)上相邻但光学可分辨的部分上或者 ZMW 阵列中的相邻零模式波导上的多个相邻放置的聚合酶来针对相同模板上的不同部分进行引导,从而它们可重复进行并测定该模板的相同部分的序列多次。

[0034] III. 测定序列前后关系

[0035] 如上所述,通过提供给定分子的序列内的各个碱基的共有序列鉴定,由分子冗余测序得到的序列信息的比较或重叠 (overlying) 提供了增加所获得的序列信息的置信的优点。为了有助于比较,各部分序列信息的前后关系也是非常有价值的。为此,本发明提供并使用模板序列内的登记序列,以提供这些前后关系。

[0036] 在本发明的至少一个实施方式中,模板序列具有登记序列,该登记序列指示模板中紧随的序列或之前的序列的前后关系。例如,通过在序列内各相同核苷酸链的起点提供登记序列,可容易地鉴别该序列的起点。在环形模板的情况中,登记序列在该模板上提供了一个设定点,以及在该环形序列上提供一提示该方法已进行到指定时间的指示。具体而言,通过在这种序列测定中加入序列前后关系参数(而实现)。该序列前后关系参数有助于确定全部靶核酸序列的较大内容内的预定序列片段或甚至更大的序列区域的位置。根据本发明,序列前后关系参数包括嵌埋于靶序列中的登记序列。模板中加入登记序列在该模板序列内的给定位点提供了内部标记物。这种标记物可提供如下益处:即允许比对来自给定模板分子或多个拷贝的模板序列的重叠序列读数,其中该登记序列在所述模板序列的多个拷贝中所处的位置相同。此外,这些登记序列可在环形模板分子内提供给定的、相对位置的标记物,从而提供有关环形模板的复制和 / 或测序已结束的指示。

[0037] 本发明的登记序列尤其可用于已开发或正在开发的掺入测序方法的中,尤其是那些依赖于单个模板分子的序列测定的方法。掺入测序法通常基于模板导向的引物延伸反应中的掺入顺序而鉴定序列中的碱基。具体而言,鉴定掺入引物延伸产物中的碱基,通过互补性鉴别模板中与其配对的碱基。这种测序方法包括那些在模板依赖性引物延伸反应中逐步加入碱基的方法,其中,每件掺入事件中掺入的碱基均被鉴别。

[0038] 通常,这类方法使用染料标记的核苷酸,该核苷酸被封端或被阻断,例如在所加入的碱基的 3' 羟基上,以阻止超出该掺入事件以外的进一步的引物延伸。一旦掺入,检测该碱基上的染料,作为该特定碱基已被掺入的指示。然后处理该复合物,以除去阻断或封端的基团和染料或标记基团,并重复该方法以鉴别随后将掺入的碱基,从而鉴别基础模板的序列。此方法的不同变化为,在某个时间使用单类型的碱基来询问 (interrogate) 该复合物,鉴别该碱基是否已被掺入,然后尝试不同的碱基。或者,可同时使用多个不同标记的碱基,并根据所掺入的核苷酸上的染料的光谱特性鉴别哪种碱基被掺入。其它方法使用非荧光检测技术,这类技术通过使用发光酶报道子系统检测掺入反应的副产物(如焦磷酸盐 / 酯)的存在而评估掺入。

[0039] 在特别优选的方面,在实时单分子测序法中使用本发明。这些方法在例如美国专利 7,033,764、7,052,847、7,056,661 和 7,056,676 中有描述,出于所有目的,本文将这些专

利的全文以应用的方式纳入。简言之,这些方法通常提供固定在固相支持物上的模板 / 引物聚合酶复合物,这样单个复合物可与其它复合物光学分辨。将标记的核苷酸导入复合物中,可直接观察到它们的掺入。通过使用互相作用的标记物可以实现直接观察,所述标记物如能量转移供体 / 受体荧光团对,其中供体或受体荧光团之一连接于复合物活性位点近端,如在聚合酶分子的分开的位置上,而该对的另一成员则连接于被掺入的核苷酸。当掺入受体标记的核苷酸时,它被带至聚合酶上的供体附近,并产生可检测的荧光信号。其它策略使用荧光团,这些荧光团自身在多聚磷酸酯链从核苷酸类似物上解离时猝灭,在掺入过程中,它产生未猝灭的染料。

[0040] 其它优选的策略提供固定复合物的光学封闭 (optical confinement),这样掺入会将标记的核苷酸带入该封闭的照射 / 检测空间内,并持续足够的时间,从而可检测荧光,并使其与随机扩散的标记物或标记核苷酸 (具有更短暂的信号特征) 相区分。特别优选的封闭例子包括零模式波导的阵列。参见,例如美国专利 7,033,764、7,052,847、7,056,661 和 7,056,676 (先前已在本文中以引用的方式纳入本文)、和美国专利 6,991,726、7,013,054 和 7,181,122,出于所有目的,本文将其全文以应用的方式纳入。具体而言,这些 ZMW 的特征是置于透明基底之上的覆层,该覆层具有中空的芯,通过该芯可到达底层的基底,该芯被安排为跨过该覆层 / 基底。当光的频率低于芯的截留频率时,具有纳米级横截面大小 (例如横截面 (长、宽或直径) 约为 20 到约 200nm) 的这些芯通过减弱光通过该芯的传播而提供了光学封闭。结果,光仅能在芯内穿透短的距离,从而在该芯的末端 (光从该处导出) 提供非常小的照射体积。

[0041] 如前所述,用于这些测序方法的模板中掺入登记序列为这些方法所获得的序列信息提供了参比或比对序列。这些比对序列可用于鉴别对环形模板测序获得的重复节段,或用于比对从相同的或同样的线性或环形模板序列组获得的序列。图 1 阐述了这些登记序列在获得这些比对中的应用。具体而言,如图 I 所示,在适当的引物序列 104 和核酸聚合酶 106 结合到模板分子 102 后进行引物延伸反应。模板序列 102 具有登记序列 108。该模板可以是与引物和聚合酶在引物延伸反应中反复反应的单模板分子,例如是可重复反应的线性分子,或者是适于进行连续反应的环形分子,或者该模板可含有多条与一种或多种不同的聚合酶和引物反应的相同模板序列。不论模板是单分子还是多个相同序列,在它的序列的某个位置上都包括登记序列 (以有阴影线的框表示)。

[0042] 如图 II 所示,从这些模板获得的序列信息本质上是有所不同的,例如,它们获自该靶 / 模板序列的不同部分,因此使用聚合酶 106 实施的重复模板导向的引物延伸产生了衍生自模板序列的序列信息或读数 110、112、114 和 116,其中包括与登记序列 108 互补的序列,分别显示为登记互补物 118、120、122 和 124。

[0043] 利用登记互补物 118-124 的存在,可确定这些测得的序列如何比对,如图 III 所示。如本文其它地方所述,这可提供关于序列测定的程度的指示,例如,是否已测到给定的点,或者可使用该指示确定已获得的序列覆盖水平。具体而言,在核酸测序中,测序方法的精确性通常依赖于给定序列区域的多倍测序或覆盖水平。通过鉴定每次通过模板序列的给定部分的开始和结束,可容易地确定单模板分子或其部分的覆盖水平。同样地,在测序登记序列在序列中处于相同位置的多个相同的模板分子 (或重叠模板分子) 时,提供了已围绕该登记序列进行的特定序列区域的测序的次数。例如,结合图 1 的 III,从比对的序列可鉴

定序列的 A 和 F 部分已被覆盖两次, B 和 E 部分已被覆盖了 3 次, C 部分(登记序列)和 D 部分具有四倍序列覆盖。

[0044] 在一些情况中,可能需要将一个以上登记序列掺入给定模板分子中,以将测序过程中获得的序列信息归在一起。这允许鉴别特定靶序列节段的起点和终点。给定模板中的这种多登记序列可含有相同或不同的碱基序列,以有助于它们的鉴定和区分。

[0045] 在其它方面,可在用于测序的遗传物质的分开片段中使用登记序列,各片段来自遗传物质较大部分(如基因组)的给定区域,用给定登记序列标记,而来自不同区域的模板则用不同的登记序列标记。然后,登记序列的测序提供测序方法所测的模板内的登记,和特定片段获自该较大基因组内容何处的鉴定。实践中,可将较大遗传靶材料如基因组分成独立的部分,如 100kb,或 100 万碱基,或染色体。然后将登记序列加到各部分的多个位置中,如通过部分限制性消化和标记的连接。通过这种方法,给定部分中的每一段(piece)都具有相同的登记序列,该登记序列将起到鉴定标签的作用,而所有的部分都提供有不同的登记序列。然后可合并所有部分的所有段,并测序。然后使用登记序列信息集合单个读数,如所有登记序列 1 的读数得自部分 1,所有登记序列 2 的读数得自部分 2。这种分成部分的方法对大靶标中的重复序列特别有帮助,因此这些重复可指定到其原始的位置。分成部分的方法还有助于集合通常难以集合的大靶标内的短读数。

[0046] 如本文其它地方已间接提到的,除了它们在比对序列信息的用途中外,某些模板(如登记序列位于模板的 5' 端的线性、或环形模板)内的登记序列的鉴定可提供整个模板序列已被测序或者已至少通过所述模板的 5' 端测序过的元件信息(element information)。具体而言,对于环形模板,模板测序中出现至少两次的登记序列的鉴定证实了该模板的整条序列在一倍覆盖中已被询问/测到。此外,如图 1 所示,这种登记序列可提供单个环形模板或多个相同模板的序列覆盖水平的直接读数。

[0047] 前述内容在图 2 中示意性阐述。如图 I 所示,提供环形模板分子 200,其与引物序列 202 和链取代聚合酶 204 形成复合物。模板序列 200 包括登记序列部分 206。在引物延伸反应期间(图 II),产生延伸反应产物 208 或新生链,其包括登记序列 206 的互补序列 210。使用例如优选的实时单分子测序方法从该反应获得的序列信息示意性地显示在图 IIIA、B 和 / 或 C 中。具体而言,如图 IIIA 所示,获得了从登记序列 212 的至少第一鉴定至该序列 214 的第二鉴定的序列信息,其中居间节段 216 的序列已被测定。结果,可容易地确定已对整个环形模板序列测序了至少一次。通过持续围绕着环形模板进行测序而获得多覆盖测序,如图 IIIB 所示。覆盖水平的测定仅仅需要登记序列的鉴定之间的居间序列的通过次数即可(所示的为 6 倍覆盖)。

[0048] 另一方面,可通过使用登记序列比对在相同环形模板序列上或多条相同模板序列上进行的多个不同的延伸反应获得的序列信息,并将其归类(coordinate),如图 IIIC 所示。例如,在某些情况中,可从环形模板不同部分的引物延伸获得序列信息的分立的、非连续的段。这可能是聚合酶解离、反应条件变化(有意或其它)或在相同模板序列上发生的多种不同的延伸反应(其中各不同反应提供可鉴别的序列输出)所得的给定引物延伸反应的中断导致的结果。然后,通过其登记序列可将序列信息的这些多个分立的片段相关联,如图 IIIC 所示。

[0049] 如前所述,可使用登记序列从给定模板分子获得共有序列信息。具体而言,通过基

于所鉴别的登记序列比对序列（包括例如至少部分比对），以及通过增加的比较，将其相同的核苷酸序列被重复测序的单模板分子的序列信息与它自身比较，获得该序列内各碱基位置的共有序列信息。这种共有序列测定的一个阐述性例子阐述于图 3 中。如图所示，基于它们的登记序列节段 314 比对其相同的核苷酸序列被重复测序的单模板获得的序列读数（显示为序列 302-312）。如图所示，从相同模板（如环形模板或重复的序列模板，如本文其它地方所述）内的相同序列的重复测序获得相同序列。将在一种或多种序列读数中有不同称谓的碱基（如碱基 316 和 318）与相同位置上的共有序列鉴定结果（如碱基 320 和 322）比较。通过获得给定序列位置的足够的共有序列信息，可获得所需的置信水平、该位置的碱基，并从而确定该模板的共有序列 324。

[0050] 应理解，可在给定位置获得的置信将依赖于通过该位置的重复序列读数的数量以及所使用的测序系统的固有的误差比例。但是，应注意的是，如果假定掺入期间非系统性的误差比例（如随机序列误差的比例）高达 25%，则对于给定序列，冗余测序法中 5 倍的序列覆盖将产生非常低的误差比例，理论上低于 0.1%。

[0051] 对于从冗长测序中读取 (call) 碱基，应理解的是必须建立给定读取的共有序列。因此，如果给定位置的反复测序产生不确定性，如在各测序轮次中鉴定的结果不同，那么共有序列的读取将需要在该位置读取至少 2 次，以建立多数一致性。此外，为了增加这种共有序列读取的置信，将需要在该位置进行三次、四次、五次、十次或更多次的读取。举例而言，如果鉴定给定位置，一次为 A，一次为 C，没有更多的鉴定，则将难以鉴定该位置的正确碱基读数，不确定性超过 50%。但是，在该位置实施一次额外的读取，如第 3 次读取为 A，则可以合理的确定性确定 C 读数是错误的。

[0052] 因此，结合本发明，比较给定模板的序列数据包括将该模板的相同核苷酸序列组的至少两个序列读数、优选至少 5 个序列读数、在某些情况中至少 10 个序列读数、在其它情况中至少 15 个甚至 20 个序列读数进行比较。

[0053] 一旦从给定靶序列获得多个读数，比对这些读数，产生共有序列。最简单的是，比对可通过人工地将共有序列元件置于相同序列位置中并针对遗漏或取代的碱基而调整来实现。但是，在优选的方面，通过计算机使用比对软件比对各序列读数。可容易地获得各种基因比对程序，包括例如 BLAST、FASTA、SSEARCH、SAM、DNABaser 等等。

[0054] 再次，一旦比对各序列读数，可建立共有序列。最简单的，可使用简单多数性读数建立给定位置的一致读数。但是，在一些情况下，可能需要建立绝对多数性的一致碱基读数，以消除可能产生自测序方法的误差比例的不确定性。例如，可能需要 51% 的多数性来建立一致读数，在优选的方面将需要大于 60%、大于 70%、大于 75%、大于 80% 甚至大于 90% 的一致性，以建立给定位置的一致读数。在优选的实施方式中，同样可通过适当编程以接受序列数据的计算机来实施序列比较和一致碱基测定，基于所有比对序列的一致性百分百认定一致读数。

[0055] III. 登记序列

[0056] 登记序列的特性将至少一定程度上依赖于使用该序列的应用的性质。具体而言，登记序列的长度、序列组成和在模板序列内所希望的位置（如果需要的话）可能会根据该应用而变化，下文将对此有更详细的描述。已知的登记序列优选是外源导入的登记序列，也称为“外源序列”或类似称呼，以与靶核酸序列中以其原始状态存在的已知部分相区别。通

常,可预先选择这种外源序列,在样品制备步骤中将其导入靶序列中,在某些情况中根据待测序的模板的性质,例如长度、预期的序列前后关系、该模板中天然存在的该登记序列的复制可能性等等来设计该外源序列,以提供最佳的登记序列的识别。除了预先选择和/或设计外源登记序列外,通常还将登记序列置于靶序列内的某个可被选择的位置上。该选择的位置可以是确实已知的,或者是相对已知的,如已知在给定模板序列如环形模板中存在一定数量的次数(如一次),或者很靠近给定模板序列的3'或5'端,或者在该3'或5'端。

[0057] 可选择登记序列,使其在靶序列内被复制的可能性最小,确定地鉴别登记序列的机会最大。通常,靶序列内登记序列不被复制的可能性的变化将是该登记序列长度的函数。因此,靶序列内登记序列的长度通常将根据待测的靶序列的长度而变化。

[0058] 举例而言,5碱基登记序列存在于随机的1kb靶核酸内的可能性(虽然不高)将比1kb靶序列内出现的给定10碱基序列的可能性高。将理解,模板序列越长,登记序列的所需长度也越长,以解决提高的复制可能性。同样地,越希望避免模板序列内的天然复制,所需的登记序列将越长。为了大多数测序应用,如实时测序,其中通常需要对模板序列内的100到1000、10000或更多个连续碱基进行测序,通常登记序列的长度将为约5个到约100个碱基,甚至更长,优选的登记序列的长度为约10到约50个,在某些优选方法,为约10到约20个碱基。

[0059] 登记序列中的碱基的特定序列并不是关键的因素,因为较大序列中存在的n个碱基的任意序列的可能性都是相同的,而不论该序列中的所述特定碱基是什么。但是,在某些情况下,可能需要选择登记序列中的碱基的特定序列,以利用某些碱基优于其它碱基的测序精确性、易于可鉴别的染料组合、靶序列中的预期的序列偏差等等。不管靶序列中碱基的情况如何,根据本发明,登记序列由已知碱基序列组成,这样可易于从它所嵌埋的靶序列中鉴别该登记序列。

[0060] 如上所述,登记序列通读在模板序列的某个选定的相对位置提供。“选定的相对位置”指登记序列的位置未必确切知晓,但至少部分相对于一些其它成分而选择它的位置。例如,在某些情况中,如线性靶序列,登记序列可靠近或在该靶序列3'或5'端。

[0061] 但在环形序列的情况中,环形靶标内任意位置放置的登记序列都被视为是选定的相对位置,因为它的位置相对于它自身(一条全长的靶序列)是已知的。

[0062] 通常可通过多种方法实现靶序列或模板序列内登记序列的放置。例如,登记序列可作为复制靶序列(例如使用PCR或其它非线性扩增方法)而使用的引物序列上的粘附的标签提供,这样该靶序列基本上以登记序列位于待测序的链的3'或5'端的状态存在于样品内。也可使用连接方法将登记序列连接到双链序列的一条链的一端或另一端。随后的双链模板的解链和与聚合酶和引物的结合提供了包含有该登记序列的新合成的模板。

[0063] 在至少一个优选的方面,使用环形靶核酸作为测序的模板核酸,例如,如图2示意性所述。在某些例子中,由于能够被重复测序而不需要重新启动复合物的形成,因此这类环形模板特别优选。通过围绕着该环形模板重复测序,可获得序列冗余或多序列覆盖,结果,可证实所测定的序列。但是,在使用这种冗余测序法时,特别有用的是能够认识到模板已被完全测序或者被测了指定的次数。通过在模板中引入登记序列部分,可在每一次对环形模板进行测序后提供计数。除了证实环形模板已被测序至少一次外,登记序列还提供关于模板多次测序的比对的记录,以比对倍数覆盖(fold coverage)。可为单环形分子的测序提供

这种比对,或者,在某些情况中,为相同环形模板分子(其中登记序列在各模板中处于相同的位置)的多次测序提供比对。

[0064] 举例而言,在采用掺入的测序方法中,当采用环形模板时,直到有足够的序列冗余(达到所需的可能性水平)指示测序方法已围绕该环形模板实施了至少一次后才能够确定已获得整个模板序列。在重新对给定模板测序时,由于正确的序列是未知的,因此通常需要在获得结论之前将整个模板序列覆盖几次。在使用已知的登记序列的情况下,在已知的登记序列已被通读和鉴别两次时,可确定该序列已被通读过至少一次,从而保证已围绕该环完成了至少一次通读。此外,在寻求多倍覆盖的情况中,对单反应获得的序列输出内的登记序列的反复鉴定将提供获自该模板的序列覆盖水平的直接指示。

[0065] 可采用已知的样品制备方法容易地获得环形模板,其中,将登记序列附加或连接到线性模板的末端部分(或作为预先合成的引物上的标签而加入,该引物用于从基础靶序列上制备模板),接着使用例如环化连接酶或其它已知的环化技术将模板环化。

[0066] 应理解,登记序列可以以其它序列部分的整合的成分存在。例如,登记序列可作为测序引物识别序列,或连接到测序引物识别序列上。这样,登记序列的引入可与引导序列的引入同时进行。其它功能性序列组分可包括限制性位点、二级结构诱导序列,如发夹等。

[0067] 虽然本发明以实时读取序列信息的方式进行一般性的描述,但是应理解的是,优选的是提供测序部分或整个模板序列的序列数据,接着根据登记序列的存在评估该序列数据。序列数据中鉴定到转录的登记序列指示两次表明该整个模板序列已提供在该读数中。

[0068] 虽然以固定的模板长度进行阐述,也应理解的是本文所述的登记序列可包括在不同长度模板的一端,该模板然后被环化。如此处理将为该模板序列的节段提供差异化的登记序列读数,所述节段在线性模板中与登记序列的距离不同,但环化时将位于靠近登记序列的位置上。可采用各种方法提供不同长度的模板序列,例如沿着模板的长度在不同的位点限制或切割模板序列,以产生原始模板序列嵌套片段组,各片段在一个末端有登记序列,而整条序列的另一末端有不同的位点。当模板片段被环化时,从登记序列开始向新连接的片段末端的方向进行测序,结合其它不同长度的片段的其它测定序列,可从该新序列获得序列前后关系信息。结果将通过对该多嵌套模板片段进行测序而提供虚拟的模板序列的倍数覆盖。这在附图 4 由示意性阐述。

[0069] 如图 4A 所示,提供共用序列的一组不同长度的片段(402,404,406)。将登记序列 408 连接到这些片段的相同末端,如在相同的序列位置上,然后将它们环化。然后使用该登记序列作为环形序列的共同的起始位点的指示,这样将知道靠近该登记序列的非共有区域被连接,但与该共同的起始位点的距离不同。具体而言,单模板序列的各不同长度的嵌套片段(402,404,406)的共同末端连接有登记序列 408。然后环化各片段,提供环形模板(如环形模板 412,414,416)。在从该登记序列向环化片段的未共有末端进行测序时(如箭头所示),可从各环形模板 412、414 和 416 产生不同的序列信息(分别如图 4B 示意性所示的序列读数 422、424 和 426)。

[0070] 由于各序列读数得自相同的基础模板,可以知晓这些片段得自相同的序列前后关系。此外,通过确认不同片段的相对长度(例如通过对这些片段实施基于大小的分离),可提供整个模板序列前后关系内各非共有序列的相对位置。这在图 4C 有示意性阐述。如图所示,基于特定大小的片段的来源,能够确定各测定序列读数在基础序列 430 中的前后关

系。此外,如果任意重叠的序列部分存在于任意两个或多个序列读数的话,可确定这些读数的连续的前后关系。

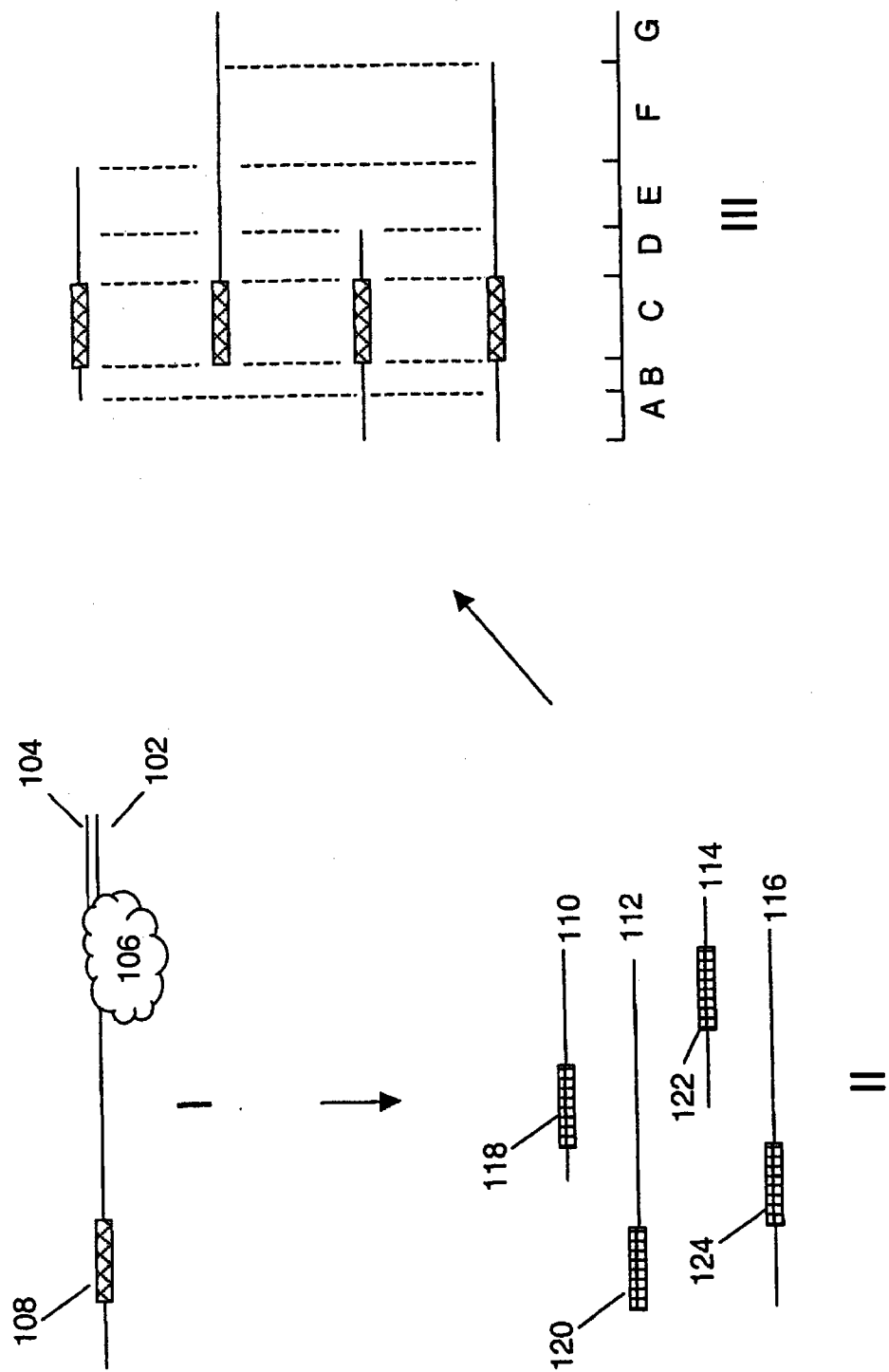


图 1

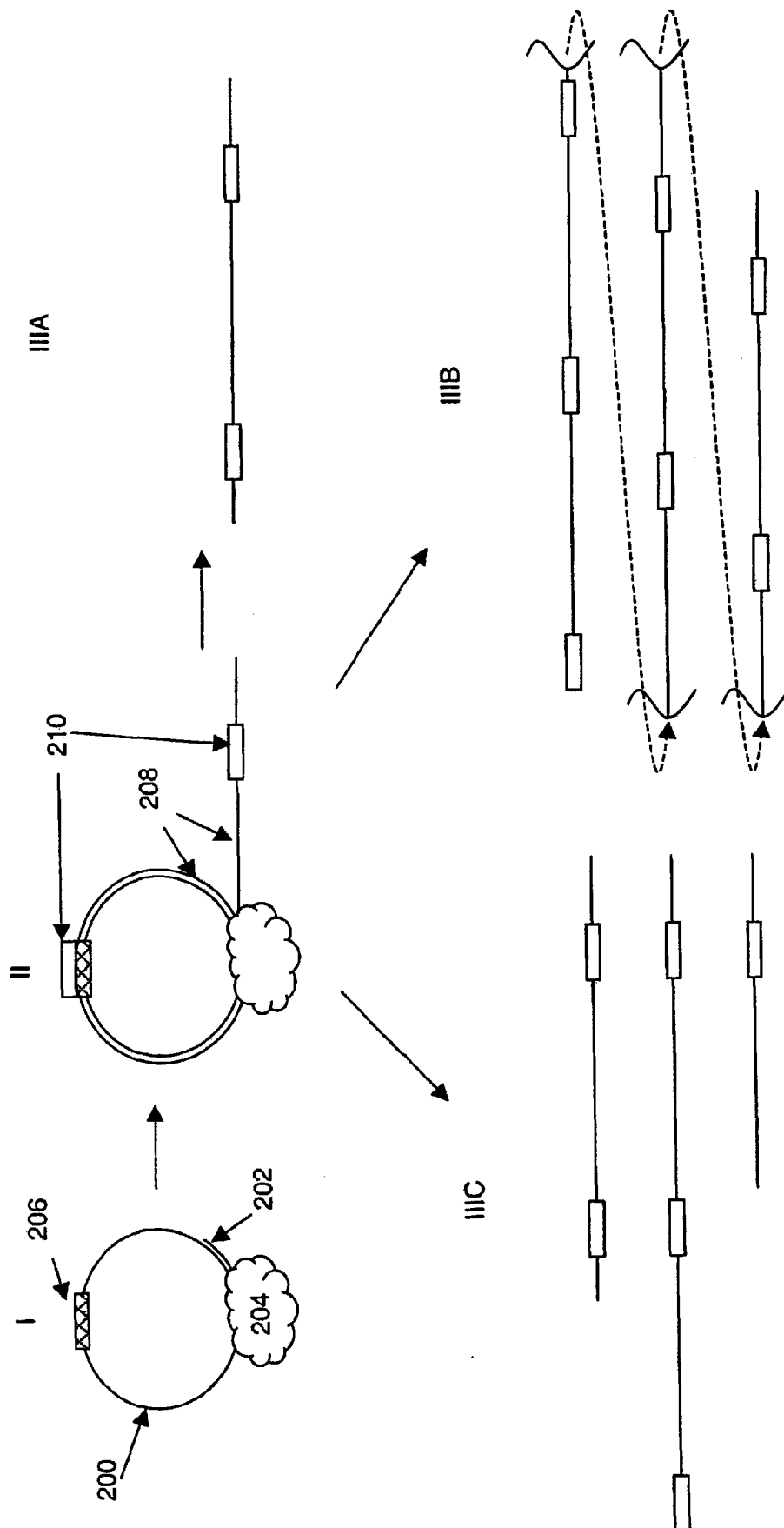


图 2

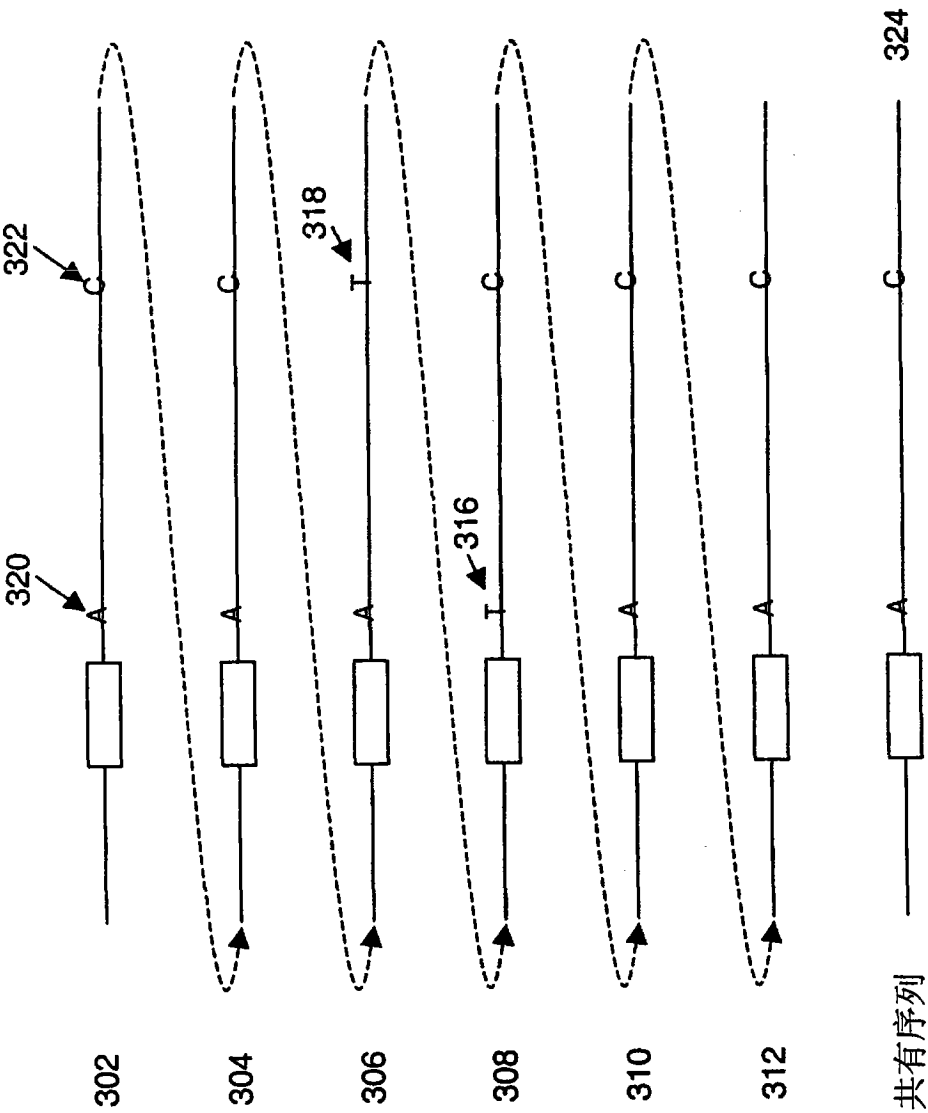


图 3

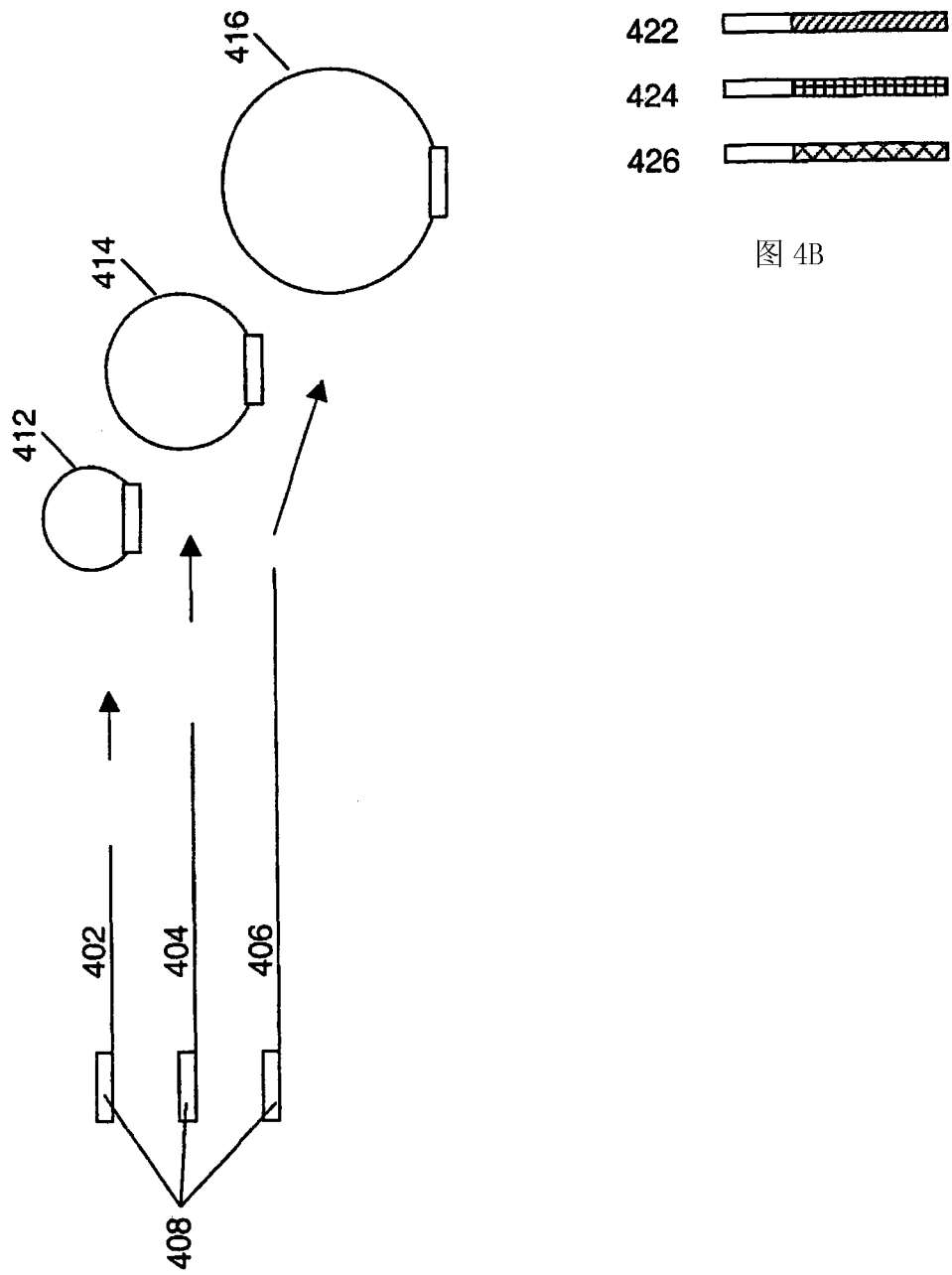


图 4A



图 4C