



(12) 发明专利申请

(10) 申请公布号 CN 117888207 A

(43) 申请公布日 2024. 04. 16

(21) 申请号 202410055914.1

(51) Int.Cl.

(22) 申请日 2018.03.14

C40B 50/06 (2006.01)

C12N 15/10 (2006.01)

(30) 优先权数据

62/471,723 2017.03.15 US

62/578,326 2017.10.27 US

(62) 分案原申请数据

201880032556.5 2018.03.14

(71) 申请人 特韦斯特生物科学公司

地址 美国加利福尼亚州

(72) 发明人 安东尼·考克斯 陈思远

查尔斯·莱多格

多米尼克·托普帕尼

(74) 专利代理机构 北京安信方达知识产权代理

有限公司 11262

专利代理师 韦昌金 郑霞

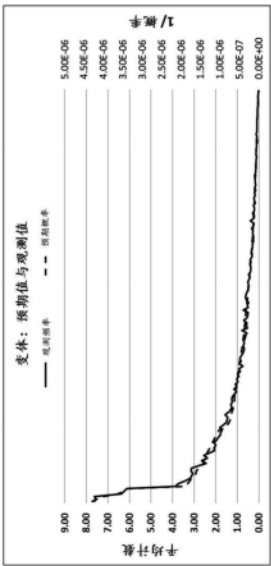
权利要求书1页 说明书61页
序列表(电子公布) 附图39页

(54) 发明名称

从头合成的组合核酸文库

(57) 摘要

本申请涉及从头合成的组合核酸文库。本文公开了用于生成编码核酸序列预定变体的高度准确的核酸文库的方法。变异程度可以是完全的,从而产生饱和的变体文库,或者变异程度可以是不完全的,从而产生非饱和的变体文库。本文所述的变异核酸文库可被设计用于通过转录或翻译进一步加工。本文所述的变异核酸文库可被设计用来生成变异RNA、DNA和/或蛋白质群体。本文进一步提供了用于鉴定具有增加或降低活性的变异种类的方法,其应用于调节用于治疗或减轻疾病的治疗剂的生物学功能及设计。



1. 一种合成变异核酸文库的方法,其包括:
 - a. 提供编码至少500个多核苷酸序列的预定序列,其中所述至少500个多核苷酸序列具有预选的密码子分布;
 - b. 合成编码所述至少500个多核苷酸序列的多个多核苷酸;
 - c. 测定由所述多个多核苷酸编码的核酸或基于所述多个多核苷酸翻译的蛋白质的活性;以及
 - d. 从步骤(c)的测定中收集结果,其中所述收集包括收集与阴性或无效结果相关的预定序列的结果。
2. 根据权利要求1所述的方法,其中步骤(d)包括收集至少80%的所述预定序列的结果。
3. 根据权利要求1所述的方法,其中步骤(d)包括收集至少90%的所述预定序列的结果。
4. 根据权利要求1所述的方法,其中步骤(d)包括收集至少100%的所述预定序列的结果。
5. 根据权利要求1所述的方法,其中呈现出预测多样性的至少约70%。
6. 根据权利要求1所述的方法,其中呈现出预测多样性的至少约90%。
7. 根据权利要求1所述的方法,其中呈现出预测多样性的至少约95%。
8. 根据权利要求1所述的方法,其中所述至少500个多核苷酸序列中的至少80%具有正确的大小。
9. 根据权利要求1所述的方法,其中所述至少500个多核苷酸序列中的至少约80%各自以所述文库中每个所述多核苷酸序列的平均频率的2倍以内的量存在于所述变异核酸文库中。
10. 根据权利要求1所述的方法,其进一步包括从步骤(c)的测定中收集与增强或降低的活性相关的预定序列的结果。

从头合成的组合核酸文库

[0001] 本申请是申请日为2018年03月14日,申请号为201880032556.5,发明名称为“从头合成的组合核酸文库”的申请的分案申请。

[0002] 交叉引用

[0003] 本申请要求2017年10月27日提交的第62/578,326号美国临时申请和2017年3月15日提交的第62/471,723号美国临时申请的权益,所述临时申请中的每一个均通过引用整体并入本文。

[0004] 序列表

[0005] 本申请含有以ASCII格式电子提交的序列表,并且其通过引用整体并入本文。创建于2018年3月13日的所述ASCII副本被命名为44854-729_601_SL.txt,大小为18,419个字节。

背景技术

[0006] 合成生物学的基石是设计、构建和测试过程——一个需要DNA,以使得便于快速且可行地生成并优化这些定制途径和生物体的迭代过程。在设计阶段,将构成DNA的A、C、T和G核苷酸规划成包含感兴趣的基因座或途径的多种基因序列,其中每种序列变体代表将进行测试的特定假设。这些变异基因序列代表序列空间(起源于进化生物学的一个概念)的子集,并且从属于构成基因、基因组、转录物组和蛋白质组的全部序列。

[0007] 通常针对每个设计-构建-测试循环设计许多不同的变体,以实现序列空间的充分采样并使优化设计的可能性最大化。尽管在概念上很简单,但与常规合成方法的速度、通量和质量相关的工艺瓶颈阻碍了这一循环进展的步伐,从而延长了开发时间。由于极其准确的DNA的高成本和当前合成技术的有限通量导致无法充分探索序列空间仍然是限速步骤。

[0008] 从构建阶段开始,有两个过程值得注意:核酸合成和基因合成。以往,通过分子克隆实现不同基因变体的合成。这种方法虽然稳定,但无法放大。早期的化学基因合成工作集中于产生大量具有重叠序列同源性的多核苷酸。随后将这些多核苷酸合并,并经历多轮聚合酶链反应(PCR),从而使重叠的多核苷酸连接成长双链基因。许多因素阻碍了这一方法,包括构建耗时耗力、需要大量的亚磷酰胺、原材料昂贵以及产生纳摩尔量的最终产物(显著低于下游步骤所需的量),并且大量单独的多核苷酸需要一个96孔板来建立一个基因的合成。

[0009] 在微阵列上合成多核苷酸使得基因合成的通量显著增加。可以在微阵列表面上合成大量的多核苷酸,然后切下并合并在一起。针对特定基因的每种多核苷酸含有独特的条形码序列,该条形码序列能够使特定的多核苷酸亚群区分开(depooled)并装配成感兴趣的基因。在该过程的这个阶段,将每个子池转移至96孔板中的一个孔中,从而使通量增加到96个基因。虽然其通量比经典方法高两个数量级,但由于缺乏成本效益且周转时间缓慢,它仍然不能充分支持一次需要数千个序列的设计、构建、测试循环。

发明内容

[0010] 本文提供了合成变异核酸文库的方法,其包括:(a)提供编码至少500个多核苷酸序列的预定序列,其中所述至少500个多核苷酸序列具有预选的密码子分布;(b)合成编码所述至少500个多核苷酸序列的多个多核苷酸;(c)测定由所述多个多核苷酸编码的核酸或基于所述多个多核苷酸翻译的蛋白质的活性;以及(d)从步骤(c)的测定中收集结果,其中所述收集包括收集与阴性或无效(null)结果相关的预定序列的结果。本文还提供了合成变异核酸文库的方法,其中步骤(d)包括收集至少90%的所述预定序列的结果。本文还提供了合成变异核酸文库的方法,其中步骤(d)包括收集至少100%的所述预定序列的结果。本文还提供了合成变异核酸文库的方法,其中呈现出(represent)预测多样性的至少约70%。本文还提供了合成变异核酸文库的方法,其中呈现出预测多样性的至少约90%。本文还提供了合成变异核酸文库的方法,其中呈现出预测多样性的至少约95%。本文还提供了合成变异核酸文库的方法,其中所述至少500个多核苷酸序列中的至少80%具有正确的大小。本文还提供了合成变异核酸文库的方法,其中所述至少500个多核苷酸序列中的至少约80%各自以所述文库中每个所述多核苷酸序列的平均频率的2倍以内的量存在于所述变异核酸文库中。本文还提供了合成变异核酸文库的方法,其进一步包括从步骤(c)的测定中收集与增强或降低的活性相关的预定序列的结果。本文还提供了合成变异核酸文库的方法,其中所述活性是细胞活性。本文还提供了合成变异核酸文库的方法,其中所述细胞活性包括增殖(reproduction)、生长、粘附、死亡、迁移、能量产生、氧利用、代谢活性、细胞信号传导、对自由基损伤的响应或其任意组合。本文还提供了合成变异核酸文库的方法,其中所述变异核酸文库编码变异基因或其片段的序列。本文还提供了合成变异核酸文库的方法,其中所述变异核酸文库编码抗体、酶或肽的至少一部分。本文还提供了合成变异核酸文库的方法,其中所述核酸文库编码指导RNA(gRNA)。本文还提供了合成变异核酸文库的方法,其中所述核酸文库编码siRNA、shRNA、RNAi或miRNA。

[0011] 本文提供了用于生成核酸组合文库的方法,该方法包括:(a)设计预定的序列,该序列编码:(i)第一多个多核苷酸,其中所述第一多个多核苷酸中的每个多核苷酸编码与单个参考序列相比的变异序列,和(ii)第二多个多核苷酸,其中所述第二多个多核苷酸中的每个多核苷酸编码与单个参考序列相比的变异序列;(b)合成所述第一多个多核苷酸和所述第二多个多核苷酸;以及(c)混合所述第一多个多核苷酸和所述第二多个多核苷酸以形成核酸的组合文库,其中呈现出预测多样性的至少约70%。本文还提供了用于生成核酸组合文库的方法,其中所述组合文库是非饱和组合文库。本文还提供了用于生成核酸组合文库的方法,其中所述组合文库是饱和组合文库。本文还提供了用于生成核酸组合文库的方法,其中合成了至少10,000个多核苷酸。本文还提供了用于生成核酸组合文库的方法,其中用于生成所述非饱和组合文库的多核苷酸的总数比用于生成饱和组合文库的多核苷酸的总数少至少25%。本文还提供了用于生成核酸组合文库的方法,其中至少80%的变体具有正确的大小。本文还提供了用于生成核酸组合文库的方法,其中呈现出预测多样性的至少约90%。本文还提供了用于生成核酸组合文库的方法,其中呈现出预测多样性的至少约95%。本文还提供了用于生成核酸组合文库的方法,其中所述组合文库编码第一参考序列或第二参考序列。本文还提供了用于生成核酸组合文库的方法,其中所述组合文库在翻译时编码蛋白质文库。本文还提供了用于生成核酸组合文库的方法,其中将所述组合文库的

核酸插入载体中。本文还提供了用于生成核酸组合文库的方法,其进一步包括使用所述组合文库作为PCR诱变反应的引物来进行核酸的PCR诱变。本文还提供了用于生成核酸组合文库的方法,其中所述组合文库编码变异基因或其片段的序列。本文还提供了用于生成核酸组合文库的方法,其中所述组合文库编码抗体、酶或肽的至少一部分。本文还提供了用于生成核酸组合文库的方法,其中所述组合文库编码所述抗体的可变区或恒定区的至少一部分。本文还提供了用于生成核酸组合文库的方法,其中所述组合文库编码所述抗体的至少一个CDR区。本文还提供了用于生成核酸组合文库的方法,其中所述组合文库编码在所述抗体的重链上的CDR1、CDR2和CDR3以及在其轻链上的CDR1、CDR2和CDR3。本文还提供了用于生成核酸组合文库的方法,其中所述组合文库编码指导RNA(gRNA)。

[0012] 本文提供了合成变异核酸文库的方法,其包括:(a)提供编码多个多核苷酸的预定序列,其中所述多核苷酸编码与单个参考序列相比具有变异序列的多个密码子;(b)为预定核酸参考序列中预选位置处的密码子选择分布值;(c)提供机器指令以随机生成一组具有与所选分布值相匹配(aligned)的分布值的核酸序列,其中该组核酸序列少于生成饱和密码子变体文库所需的核酸序列的量;以及(d)合成具有预选的分布的变异核酸文库,其中呈现出预测多样性的至少约70%。本文还提供了合成变异核酸文库的方法,其中至少80%的变体具有正确的大小。本文还提供了合成变异核酸文库的方法,其中呈现出预测多样性的至少约90%。本文还提供了合成变异核酸文库的方法,其中呈现出预测多样性的至少约95%。本文还提供了合成变异核酸文库的方法,其中所述变异核酸文库在翻译时编码蛋白质文库。本文还提供了合成变异核酸文库的方法,其中将所述变异核酸文库的核酸插入载体中。本文还提供了合成变异核酸文库的方法,其进一步包括使用所述变异核酸文库作为PCR诱变反应的引物来进行核酸的PCR诱变。本文还提供了合成变异核酸文库的方法,其中使用密码子分配来确定具有变异序列的所述多个密码子中的每个密码子。本文还提供了合成变异核酸文库的方法,其中所述密码子分配基于生物体中密码子序列的频率。本文还提供了合成变异核酸文库的方法,其中所述生物体是动物、植物、真菌、原生生物、古菌和细菌中的至少一种。本文还提供了合成变异核酸文库的方法,其中所述密码子分配基于所述密码子序列的多样性。

[0013] 本文提供了合成变异核酸文库的方法,其包括:(a)提供编码多个多核苷酸的预定序列,其中所述多核苷酸编码与单个参考序列相比具有变异序列的密码子;(b)将所述多个多核苷酸分成多核苷酸的5'片段和多核苷酸的3'片段;(c)为预定核酸参考序列中预选位置处的密码子选择分布值;(d)提供机器指令以随机生成一组具有与所选分布值相匹配的分布值的核酸,其中该组核酸少于生成饱和核酸文库所需的核酸的量;(e)合成多核苷酸的5'片段和多核苷酸的3'片段;以及(f)混合多核苷酸的5'片段和多核苷酸的3'片段以形成变异核酸文库,其中呈现出预测多样性的至少约70%。本文还提供了合成变异核酸文库的方法,其中合成了至少10,000个多核苷酸。本文还提供了合成变异核酸文库的方法,其中至少80%的变体具有正确的大小。本文还提供了合成变异核酸文库的方法,其中呈现出预测多样性的至少约90%。本文还提供了合成变异核酸文库的方法,其中呈现出预测多样性的至少约95%。本文还提供了合成变异核酸文库的方法,其中将所述多个多核苷酸分成多于一个5'片段和多于一个3'片段中的至少一种。本文还提供了合成变异核酸文库的方法,其中所述变异核酸文库在翻译时编码蛋白质文库。本文还提供了合成变异核酸文库的方法,

其中将所述变异核酸文库的核酸插入载体中。本文还提供了合成变异核酸文库的方法,其进一步包括使用所述变异核酸文库作为PCR诱变反应的引物来进行核酸的PCR诱变。本文还提供了合成变异核酸文库的方法,其进一步包括鉴定具有增强或降低的活性的变异序列。本文还提供了合成变异核酸文库的方法,其中所述活性是细胞活性。本文还提供了合成变异核酸文库的方法,其中所述细胞活性包括增殖、生长、粘附、死亡、迁移、能量产生、氧利用、代谢活性、细胞信号传导、对自由基损伤的响应或其任意组合。本文还提供了合成变异核酸文库的方法,其中所述变异核酸文库编码变异基因或其片段的序列。本文还提供了合成变异核酸文库的方法,其中所述变异核酸文库编码抗体、酶或肽的至少一部分。本文还提供了合成变异核酸文库的方法,其中所述变异核酸文库编码所述抗体的可变区或恒定区的至少一部分。本文还提供了合成变异核酸文库的方法,其中所述变异核酸文库编码所述抗体的至少一个CDR区。本文还提供了合成变异核酸文库的方法,其中所述变异核酸文库编码在所述抗体的重链上的CDR1、CDR2和CDR3以及在其轻链上的CDR1、CDR2和CDR3。本文还提供了合成变异核酸文库的方法,其中在所述变异核酸文库中合成的不同序列的数目在50至1,000,000的范围内。本文还提供了合成变异核酸文库的方法,其中在所述变异核酸文库中合成的不同序列的数目在500至25000的范围内。本文还提供了合成变异核酸文库的方法,其中在所述变异核酸文库中合成的不同序列的数目在1000至15000的范围内。本文还提供了合成变异核酸文库的方法,其进一步包括使用所述变异核酸文库作为PCR诱变反应的引物来进行核酸的PCR诱变。本文还提供了合成变异核酸文库的方法,其中使用密码子分配来确定具有变异序列的密码子。本文还提供了合成变异核酸文库的方法,其中所述密码子分配基于生物体中密码子序列的频率。本文还提供了合成变异核酸文库的方法,其中所述生物体是动物、植物、真菌、原生生物、古菌和细菌中的至少一种。本文还提供了合成变异核酸文库的方法,其中所述密码子分配基于所述密码子序列的多样性。本文还提供了合成变异核酸文库的方法,其中所述核酸文库编码指导RNA(gRNA)。

[0014] 本文提供了用于生成核酸组合文库的方法,该方法包括:(a)提供预定的序列,该序列编码:(i)第一多个多核苷酸,其中所述第一多个多核苷酸中的每个多核苷酸编码与单个参考序列相比的变异序列,和(ii)第二多个多核苷酸,其中所述第二多个多核苷酸中的每个多核苷酸编码与单个参考序列相比的变异序列;(b)提供具有表面的结构;(c)合成所述第一多个多核苷酸,其中所述第一多个多核苷酸中的每个多核苷酸从所述表面延伸;(d)合成所述第二多个多核苷酸,其中所述第二多个多核苷酸中的每个多核苷酸从所述表面延伸;(e)从所述表面释放所述第一多个多核苷酸和所述第二多个多核苷酸;以及(f)混合所述第一多个多核苷酸和所述第二多个多核苷酸以形成核酸的组合文库,其中呈现出预测多样性的至少约70%。本文还提供了用于生成核酸组合文库的方法,其中呈现出预测多样性的至少约90%。本文还提供了用于生成核酸组合文库的方法,其中呈现出预测多样性的至少约95%。

[0015] 本文提供了合成变异核酸文库的方法,其包括:(a)设计编码多个多核苷酸的预定序列,其中所述多核苷酸编码与单个参考序列相比具有变异序列的多个密码子;(b)合成所述多个多核苷酸以生成变异核酸文库,其中呈现出预测多样性的至少约70%;(c)表达所述变异核酸文库;以及(d)评价与变异核酸文库相关的活性。本文还提供了合成变异核酸文库的方法,其中呈现出预测多样性的至少约90%。本文还提供了合成变异核酸文库的方法,其

中呈现出预测多样性的至少约95%。

[0016] 本文提供了用于生成核酸组合文库的方法,该方法包括:(a)提供预定的序列,该序列编码:(i)第一多个不同多核苷酸,其中所述第一多个不同多核苷酸中的每个不同多核苷酸编码与单个参考序列相比的变异序列,和(ii)第二多个不同多核苷酸,其中所述第二多个不同多核苷酸中的每个不同多核苷酸编码与单个参考序列相比的变异序列;(b)提供具有表面的结构;(c)合成所述第一多个不同多核苷酸,其中所述第一多个不同多核苷酸中的每个不同多核苷酸从所述表面延伸;(d)合成所述第二多个不同多核苷酸,其中所述第二多个不同多核苷酸中的每个不同多核苷酸从所述表面延伸;(e)从所述表面释放所述第一多个不同多核苷酸和所述第二多个不同多核苷酸;以及(f)混合所述第一多个多核苷酸和所述第二多个多核苷酸以形成核酸的组合文库,其中呈现出预测多样性的至少约70%。本文提供了用于生成核酸组合文库的方法,其中所述组合文库是非饱和组合文库。本文提供了用于生成核酸组合文库的方法,其中所述组合文库是饱和组合文库。本文提供了用于生成核酸组合文库的方法,其中合成了至少10,000个多核苷酸。本文提供了用于生成核酸组合文库的方法,其中用于生成所述非饱和组合文库的多核苷酸的总数比用于生成饱和组合文库的多核苷酸的总数少至少25%。本文提供了用于生成核酸组合文库的方法,其中至少80%的变体具有正确的大小。本文提供了用于生成核酸组合文库的方法,其中所述变体组合文库编码第一参考序列或第二参考序列。本文提供了用于生成核酸组合文库的方法,其中所述组合文库在翻译时编码蛋白质文库。本文提供了用于生成核酸组合文库的方法,其中将所述组合文库的核酸插入载体中。本文提供了用于生成核酸组合文库的方法,其进一步包括使用所述组合文库作为PCR诱变反应的引物来进行核酸的PCR诱变。本文提供了用于生成核酸组合文库的方法,其中所述组合文库编码变异基因或其片段的序列。本文提供了用于生成核酸组合文库的方法,其中所述组合文库编码抗体、酶或肽的至少一部分。本文提供了用于生成核酸组合文库的方法,其中所述组合文库编码所述抗体的可变区或恒定区的至少一部分。本文提供了用于生成核酸组合文库的方法,其中所述组合文库编码所述抗体的至少一个CDR区。本文提供了用于生成核酸组合文库的方法,其中所述组合文库编码在所述抗体的重链上的CDR1、CDR2和CDR3以及在其轻链上的CDR1、CDR2和CDR3。本文提供了用于生成核酸组合文库的方法,其中所述组合文库编码指导RNA(gRNA)。本文提供了用于生成核酸组合文库的方法,其中与预定序列相比,所述组合文库具有小于1/1000个碱基的总错误率。本文提供了用于生成核酸组合文库的方法,其中所述结构是固体支持物、凝胶或珠子,并且其中所述固体支持物是板或柱。

[0017] 本文提供了合成变异核酸文库的方法,其包括:(a)提供编码多个不同多核苷酸的预定序列,其中所述不同多核苷酸编码与单个参考序列相比具有变异序列的多个密码子;(b)为预定核酸参考序列中预选位置处的密码子选择分布值;(c)提供机器指令以随机生成一组核酸,其中该组核酸少于生成饱和密码子变体文库所需的核酸的量;以及(d)合成具有预选的分布的核酸文库,其中呈现出预测多样性的至少约70%。本文提供了合成变异核酸文库的方法,其中至少80%的变体具有正确的大小。本文提供了合成变异核酸文库的方法,其中所述组合文库在翻译时编码蛋白质文库。本文提供了合成变异核酸文库的方法,其中将所述组合文库的核酸插入载体中。本文提供了合成变异核酸文库的方法,其进一步包括使用所述组合文库作为PCR诱变反应的引物来进行核酸的PCR诱变。本文提供了合成变异核

酸文库的方法,其中使用密码子分配来确定具有变异序列的所述多个密码子中的每个密码子。本文提供了合成变异核酸文库的方法,其中所述密码子分配基于生物体中密码子序列的频率。本文提供了合成变异核酸文库的方法,其中所述生物体是动物、植物、真菌、原生生物、古菌和细菌中的至少一种。本文提供了合成变异核酸文库的方法,其中所述密码子分配基于所述密码子序列的多样性。

[0018] 本文提供了合成变异核酸文库的方法,其包括:(a) 提供编码多个不同多核苷酸的预定序列,其中所述不同多核苷酸编码与单个参考序列相比具有变异序列的密码子;(b) 将所述多个不同多核苷酸分成不同多核苷酸的5' 片段和不同多核苷酸的3' 片段;(c) 为预定核酸参考序列中预选位置处的密码子选择分布值;(d) 提供机器指令以随机生成一组核酸,其中该组核酸少于生成饱和核酸文库所需的核酸的量;(e) 合成不同多核苷酸的5' 片段和不同多核苷酸的3' 片段;以及(f) 混合不同多核苷酸的5' 片段和不同多核苷酸的3' 片段以形成变异核酸文库,其中呈现出预测多样性的至少约70%。本文提供了合成变异核酸文库的方法,其中合成了至少10,000个不同多核苷酸。本文提供了合成变异核酸文库的方法,其中至少80%的变体具有正确的大小。本文提供了合成变异核酸文库的方法,其中将所述多个不同多核苷酸分成多于一个5' 片段和多于一个3' 片段中的至少一种。本文提供了合成变异核酸文库的方法,其中所述组合文库在翻译时编码蛋白质文库。本文提供了合成变异核酸文库的方法,其中将所述组合文库的核酸插入载体中。本文提供了合成变异核酸文库的方法,其进一步包括使用所述组合文库作为PCR诱变反应的引物来进行核酸的PCR诱变。本文提供了合成变异核酸文库的方法,其进一步包括鉴定具有增强或降低的活性的变异序列。本文提供了合成变异核酸文库的方法,其中所述活性是细胞活性。本文提供了合成变异核酸文库的方法,其中所述细胞活性包括增殖、生长、粘附、死亡、迁移、能量产生、氧利用、代谢活性、细胞信号传导、对自由基损伤的响应或其任意组合。本文提供了合成变异核酸文库的方法,其中所述核酸文库编码变异基因或其片段的序列。本文提供了合成变异核酸文库的方法,其中所述核酸文库编码抗体、酶或肽的至少一部分。本文提供了合成变异核酸文库的方法,其中所述核酸文库编码指导RNA(gRNA)。本文提供了合成变异核酸文库的方法,其中所述核酸文库编码所述抗体的可变区或恒定区的至少一部分。本文提供了合成变异核酸文库的方法,其中所述核酸文库编码所述抗体的至少一个CDR区。本文提供了合成变异核酸文库的方法,其中所述核酸文库编码在所述抗体的重链上的CDR1、CDR2和CDR3以及在其轻链上的CDR1、CDR2和CDR3。本文提供了合成变异核酸文库的方法,其中与多个不同多核苷酸的预定序列相比,所述核酸文库具有小于1/1000个碱基的总错误率。本文提供了合成变异核酸文库的方法,其中在所述核酸文库中合成的不同序列的数目在约50至约1,000,000的范围内。本文提供了合成变异核酸文库的方法,其中在所述核酸文库中合成的不同序列的数目在约500至约25000的范围内。本文提供了合成变异核酸文库的方法,其中在所述核酸文库中合成的不同序列的数目在约1000至约15000的范围内。本文提供了合成变异核酸文库的方法,其进一步包括使用所述组合文库作为PCR诱变反应的引物来进行核酸的PCR诱变。本文提供了合成变异核酸文库的方法,其中使用密码子分配来确定具有变异序列的密码子。本文提供了合成变异核酸文库的方法,其中所述密码子分配基于生物体中密码子序列的频率。本文提供了合成变异核酸文库的方法,其中所述生物体是动物、植物、真菌、原生生物、古菌和细菌中的至少一种。本文提供了合成变异核酸文库的方法,其中所述密码子分

配基于所述密码子序列的多样性。

[0019] 本文提供了合成变异核酸文库的方法,其包括:(a)设计编码多个不同多核苷酸的预定序列,其中所述不同多核苷酸编码与单个参考序列相比具有变异序列的多个密码子;(b)合成所述多个不同多核苷酸以生成变异核酸文库,其中呈现出预测多样性的至少约70%;(c)表达所述变异核酸文库;以及(d)评价与变异核酸文库相关的活性。

[0020] 援引并入

[0021] 本说明书中所提及的所有出版物、专利和专利申请均通过引用而并入本文,其程度犹如具体地和个别地指出每一单独的出版物、专利或专利申请均通过引用而并入。

附图说明

[0022] 图1描绘了用于生成非饱和组合文库的示意图。

[0023] 图2描绘了用于生成饱和组合文库的示意图。

[0024] 图3A-图3D描绘了结合PCR诱变步骤的变异生物分子合成的处理工艺流程。

[0025] 图4A-图4D描绘了用于生成在单个预定密码子位点处包含与参考核酸序列不同的核酸序列的核酸的处理工艺流程。

[0026] 图5A-图5F描绘了从模板核酸生成一组核酸变体的备选工作流程,其中每个变体在单密码子位置处包含不同的核酸序列。每个变异核酸在其单密码子位置处编码不同的氨基酸,不同的密码子由X、Y和Z表示。

[0027] 图6A-图6E描绘了具有多个氨基酸(每个残基由单个圆圈表示)的参考氨基酸序列(图6A)和使用本文所述方法生成的变异氨基酸序列(图6B、图6C、图6D和图6E)。参考氨基酸序列和变异序列由通过本文所述的过程生成的核酸及其变体来编码。

[0028] 图7A-图7B描绘了参考氨基酸序列(图7A,SEQ ID NO:24)和变异氨基酸序列文库(图7B,按出现顺序分别为SEQ ID NO 25-31),每个变体包含单残基变体(由“X”表示)。参考氨基酸序列和变异序列由通过本文所述的过程生成的核酸及其变体来编码。

[0029] 图8A-图8B描绘了参考氨基酸序列(图8A)和变异氨基酸序列文库(图8B),每个变体包含两个位点的单位位置变体。每个变体由带不同图案的圆圈表示。参考氨基酸序列和变异序列由通过本文所述的过程生成的核酸及其变体来编码。

[0030] 图9A-图9B描绘了参考氨基酸序列(图9A)和变异氨基酸序列文库(图9B),每个变体包含一段氨基酸(由围绕圆圈的框表示),每一段具有在序列上与参考氨基酸序列不同的三个位点的位置变体(编码组氨酸)。参考氨基酸序列和变异序列由通过本文所述的过程生成的核酸及其变体来编码。

[0031] 图10A-图10B描绘了参考氨基酸序列(图10A)和变异氨基酸序列文库(图10B),每个变体包含两段氨基酸序列(由围绕圆圈的框表示),每一段具有在序列上与参考氨基酸序列不同的一个位点的单位位置变体(由带图案的圆圈表示)。参考氨基酸序列和变异序列由通过本文所述的过程生成的核酸及其变体来编码。

[0032] 图11A-图11B描绘了参考氨基酸序列(图11A)和氨基酸序列变体文库(图11B),每个变体包含一段氨基酸(由带图案的圆圈表示),每一段具有在序列上与参考氨基酸序列不同的单位点的多位置变体。在该图示中,5个位置发生改变,其中第一个位置具有50/50的K/R比;第二个位置具有50/25/25的V/L/S比,第三个位置具有50/25/25的Y/R/D比,第四个位

置对于所有氨基酸具有相等的比例,而第五个位置对于G/P具有75/25的比例。参考氨基酸序列和变异序列由通过本文所述的过程生成的核酸及其变体来编码。

[0033] 图12描绘了编码具有CDR1、CDR2和CDR3区的抗体的模板核酸,其中每个CDR区包含多个变异位点,每个单位点(由星号表示)包含单个位置和/或一段多个连续位置,该位置可与不同于模板核酸序列的任何密码子序列互换。

[0034] 图13描绘了预测的变体分布和所得到的变体多样性的图示。

[0035] 图14描绘了通过互换两个表达盒的区段(例如启动子、开放阅读框和终止子)以生成表达盒的变体文库而产生的示例性数目的变体。

[0036] 图15呈现了说明如本文所公开的基因合成的示例性处理工作流程的步骤图。

[0037] 图16示出了计算机系统的示例。

[0038] 图17是示出计算机系统的架构的框图。

[0039] 图18是说明网络的示图,该网络被配置用于并入多个计算机系统、多个蜂窝电话和个人数据助理,以及网络附加存储(NAS)。

[0040] 图19是使用共享虚拟地址存储空间的多处理器计算机系统的框图。

[0041] 图20描绘了通过凝胶电泳解析的PCR反应产物的BioAnalyzer迹线图。

[0042] 图21描绘了显示96组PCR产物的电泳图,每组PCR产物在序列上与单密码子位置处的野生型模板核酸不同,其中每组中的单密码子位置位于野生型模板核酸序列中的不同位点。每组PCR产物包含19个变异核酸,每个变体在其单密码子位置处编码不同的氨基酸。

[0043] 图22描绘了比较变体的观测频率和预期概率的图示。

[0044] 图23描绘了每个概率箱元(bin)的平均计数的图示。

[0045] 图24描绘了PCR产物的分析图。X轴是碱基对,Y轴是荧光单位。

[0046] 图25描绘了观察到的组合变体的分布图。

[0047] 图26A-图26D示出了非饱和组合文库的生成。

[0048] 图27A-图27C描绘了单个或多个CDR区中的变体的示意图。

[0049] 图28A描绘了单个或多个重链和轻链支架中的变体的示意图。

[0050] 图28B描绘了单个或多个框架中的变体的示意图。

具体实施方式

[0051] 除非另有说明,否则本公开采用在本领域技术范围内的常规分子生物学技术。除非另有定义,否则本文使用的所有技术和科学术语具有与本领域普通技术人员通常理解的相同的含义。

[0052] 定义

[0053] 贯穿本公开内容,数值特征以范围格式给出。应当理解,范围格式的描述只是为了方便和简明,而不应被解释为对任何实施方案的范围的硬性限制。因此,除非上下文另有明确规定,否则对范围的描述应被认为明确公开了所有可能的子范围以及该范围内精确到下限单位十分之一的各个数值。例如,对诸如从1至6的范围的描述应被认为已经明确公开了诸如从1至3、从1至4、从1至5、从2至4、从2至6、从3至6等子范围,以及该范围内的各个值,例如,1.1、2、2.3、5和5.9。无论范围的宽度如何,这都是适用的。这些中间范围的上限和下限可独立地包括在更小的范围内,并且也被涵盖于本发明之中,但受制于所声称范围中的任

何被明确排除的限值。除非上下文另有明确规定,否则当所声称的范围包括限值之一或全部两者时,排除了这些包括的限值之一或全部两者的范围也被包括在本发明中。

[0054] 本文使用的术语仅用于描述特定实施方案的目的,而非旨在限制任何实施方案。除非上下文另有明确规定,否则如本文所用的单数形式“一个”、“一种”和“该”也意欲包括复数形式。进一步应当理解,术语“包括”和/或“包含”在本说明书中使用指定所述特征、整体、步骤、操作、元件和/或组分存在,但不排除存在或添加一个或多个其它特征、整体、步骤、操作、元件、组分和/或其群体。如本文所用的,术语“和/或”包括一个或多个相关所列项目的任何及所有组合。

[0055] 除非特别说明或从上下文中可以明显看出,否则如本文所用的,关于数字或数字范围的术语“约”应被理解为表示所述数字及其 $\pm 10\%$ 的数字,或者对于范围列出的值,表示低于所列下限的10%至高于所列上限的10%。

[0056] 如本文所用的,术语“预选序列”、“预限定序列”或“预定序列”可互换使用。这些术语意指在聚合物的合成或装配之前,聚合物的序列是已知的和选定的。具体地,本发明的多个方面主要就核酸分子的制备在本文中进行了描述,寡核苷酸或多核苷酸的序列在核酸分子合成或装配之前是已知的和选定的。

[0057] 本文提供了用于产生合成的(即从头合成的或化学合成的)多核苷酸的方法和组合物。贯穿全文,术语寡核苷酸(oligonucleotide)、寡核苷酸(oligo)和多核苷酸被定义为同义词。本文所述的合成多核苷酸的文库可包含共同编码一种或多种基因或基因片段的多个多核苷酸。在一些情况下,多核苷酸文库包含编码序列或非编码序列。在一些情况下,多核苷酸文库编码多个cDNA序列。cDNA序列所基于的参考基因序列可含有内含子,而cDNA序列不含内含子。本文所述的多核苷酸可编码来自生物体的基因或基因片段。示例性生物体包括但不限于原核生物(例如,细菌)和真核生物(例如,小鼠、兔、人和非人灵长类动物)。在一些情况下,多核苷酸文库包含一个或多个多核苷酸,所述一个或多个多核苷酸中的每一个编码多个外显子的序列。本文所述的文库内的每个多核苷酸可以编码不同的序列,即,不相同的序列。在一些情况下,本文所述的文库内的每个多核苷酸包含至少一个与该文库内另一个多核苷酸的序列互补的部分。除非另有说明,否则本文所述的多核苷酸序列可包括DNA或RNA。

[0058] 本文提供了用于产生合成的(即从头合成的)基因的方法和组合物。包含合成基因的文库可以通过本文其它部分进一步详述的多种方法来构建,如PCA、非PCA基因装配方法或分层基因装配,从而将两个或更多个双链多核苷酸组合(“缝合”)以产生更大的DNA单元(即,底架)。大构建体的文库可包含长度为至少1、1.5、2、3、4、5、6、7、8、9、10、15、20、30、40、50、60、70、80、90、100、125、150、175、200、250、300、400、500kb或更长的多核苷酸。大构建体可被独立选择的约5000、10000、20000或50000个碱基对的上限所约束。任意数目的编码多肽区段的核苷酸序列的合成,该序列包括编码非核糖体肽(NRP)的序列,编码以下物质的序列:非核糖肽合成酶(NRPS)模块和合成变体、其它模块化蛋白质如抗体的多肽区段、来自其它蛋白质家族的多肽区段,包括非编码DNA或RNA,如调节序列,例如启动子、转录因子、增强子、siRNA、shRNA、RNAi、miRNA、衍生自微小RNA的核仁小RNA,或任何感兴趣的功能性或结构性DNA或RNA单元。以下是多核苷酸的非限制性实例:基因或基因片段的编码区或非编码区、基因间DNA、由连锁分析限定的基因座(多个基因座)、外显子、内含子、信使RNA(mRNA)、转移

RNA、核糖体RNA、短干扰RNA (siRNA)、短发夹RNA (shRNA)、微小RNA (miRNA)、核仁小RNA、核酶、互补DNA (cDNA) (其为mRNA的DNA呈现形式,通常通过信使RNA (mRNA) 的逆转录或通过扩增来获得);经合成或通过扩增产生的DNA分子、基因组DNA、重组多核苷酸、支链多核苷酸、质粒、载体、任何序列的分离的DNA、任何序列的分离的RNA、核酸探针和引物。编码本文提及的基因或基因片段的cDNA可包含至少一个编码外显子序列的区域,而没有在相应基因组序列中发现的居间内含子序列。或者,cDNA的相应基因组序列可能最初缺少内含子序列。

[0059] 变体文库合成

[0060] 本文所述的方法提供了合成各自编码至少一个预定参考核酸序列的预定变体的核酸文库。在一些情况下,预定参考序列是编码蛋白质的核酸序列,并且变体文库包含编码至少单个密码子的变异的序列,使得由合成核酸编码的后续蛋白质中单个残基的多个不同变体通过标准翻译过程生成。核酸序列中合成的特定变化可通过将核苷酸变化并入重叠或平端寡核苷酸引物中来引入。或者,多核苷酸群体可共同编码长核酸(例如,基因)及其变体。在这种布置中,多核苷酸群体可进行杂交并且经历标准分子生物技术以形成长核酸(例如,基因)及其变体。当长核酸(例如,基因)及其变体在细胞中表达时,可生成变异蛋白质文库。类似地,本文提供了合成编码RNA序列(例如,miRNA、shRNA和mRNA)或DNA序列(例如,增强子、启动子、UTR和终止子区)的变体文库的方法。在一些情况下,所述序列是外显子序列或编码序列。在一些情况下,所述序列不包含内含子序列。本文还提供了使用本文所述的方法合成的文库中所选择出的变体的下游应用。下游应用包括鉴定具有增强的生物学相关功能(例如,生物化学亲和力、酶活性、细胞活性变化)和用于治疗或预防疾病状态的变异核酸或蛋白质序列。

[0061] 组合核酸文库

[0062] 本文描述了用于有效合成高度准确的变异核酸文库的方法。本文还提供了用于合成基于组合的变体文库的方法。本文提供的方法的有利特征在于,可以准确地预测组合文库中装配的核酸的产物和频率,从而允许在准确理解与阴性或无效结果相关的那些组合产物以及与生化或细胞活性相关增强有关的那些组合产物的情况下筛查组合文库。这样的系统优于当前的方法,即噬菌体展示,后者没有有效的手段来收集关于阴性或无效结果的信息。本文提供的方法的另一个有利特征是,当设计并测试代表性的组合文库时,与完全饱和的文库相比,所需的材料和相关成本更少,同时还允许基于从第一代组合文库产物筛选中收集到的信息采用改进的进行多样化的标准(variegation criteria)快速生成第二代和第三代文库。

[0063] 如本文所述的用于有效且准确地合成变异核酸文库的方法可产生均匀且多样化的文库。使用本文所述方法生成的文库是非随机的。使用本文所述方法生成的文库能够以所需频率精确导入每个预期变体。使用本文所述方法生成的文库由于降低了呈现度(representation)的丢失率并提高了每个文库内多核苷酸或更长核酸的种类之间的均匀性而提供了高精度。另外,在多核苷酸合成水平上的这种精确性的益处允许在功能水平上有高精度,以用于下游应用,例如评估来自掺入在密码子水平上编码的预定变异的翻译产物的蛋白质活性。在一些情况下,本文所述的用于生成精确文库的方法允许后续文库的设计的改进。由于从第一文库收集的关于阴性或无效结果的信息,此类后续文库可能在设计中更加集中。例如,使用本文所述方法合成的第一变异核酸文库可以用来生成功能性RNA或

蛋白质的变体文库,可以针对某种活性对该变体文库进行筛查。基于对与精确定义的非随机文库相关的阳性和阴性结果的观察,对第二变体文库进行设计选择,然后将第二变体文库用于进一步的筛选步骤,以进一步筛查并选择与指定活性相关的种类。该过程可以重复1、2、3、4、5、6、7、8、9、10次或更多次。可以进行文库设计、构建、筛查和重复的方法,以鉴定与单一活性或多种活性(例如,结合亲和力、稳定性和表达)相关的增强的种类。

[0064] 通过使用计算机生成文库,序列可以是已知的并且是非随机的。在一些情况下,文库包含至少或大约 10^1 、 10^2 、 10^3 、 10^4 、 10^5 、 10^6 、 10^7 、 10^8 、 10^9 、 10^{10} 个或多于 10^{10} 个变体。在一些情况下,包含至少或大约 10^1 、 10^2 、 10^3 、 10^4 、 10^5 、 10^6 、 10^7 、 10^8 、 10^9 或 10^{10} 个变体的文库中每个变体的序列是已知的。在一些情况下,文库包含预测的变体多样性。在一些情况下,文库中所呈现的多样性是预测多样性的至少或大约60%、65%、70%、75%、80%、85%、90%、95%或大于95%。在一些情况下,文库中所呈现的多样性是预测多样性的至少或大约70%。在一些情况下,文库中所呈现的多样性是预测多样性的至少或大约80%。在一些情况下,文库中所呈现的多样性是预测多样性的至少或大约90%。在一些情况下,文库中所呈现的多样性至少是预测多样性的至少或大约99%。如本文所述,术语“预测多样性”是指包含所有可能的变体的群体中的总理论多样性。

[0065] 如本文所述生成高度均匀且多样化的文库,其中每个变体的序列是已知的,这导致对与增强或降低的活性相关的那些组合产物以及与阴性或无效结果相关的那些组合产物的准确理解。知道与增强或降低的活性相关的产物以及与阴性或无效结果相关的那些组合产物可以允许将文库有效地用于后续试验。例如,在进行大规模筛选时,会导致活性增强或降低的变异序列是已知的。在进行后续筛选时,可以排除导致阴性或无效结果的序列,从而仅筛选导致活性增强或降低的变异序列。

[0066] 在一些情况下,增强或降低的活性与细胞活性相关。该细胞活性包括但不限于增殖、生长、粘附、死亡、迁移、能量产生、氧利用、代谢活性、细胞信号传导、对自由基损伤的响应或其任意组合。

[0067] 在第一示例性过程中,生成非饱和组合文库。非饱和组合文库的生成可以减少合成步骤的数目。参见图1,第一核酸群体110在位置1、2、3和4处表现出多样性。第二核酸群体120在位置5、6、7和8处表现出多样性。将第一核酸群体110与第二核酸群体120组合以产生16种核酸片段组合。可以通过平端连接将第一核酸群体110与第二核酸群体120组合。在一些情况下,第一群体和第二群体被设计为使得它们具有包含限制酶识别区的互补重叠序列,使得在每个群体中的核酸切割之后,第一群体和第二群体能够互相退火。

[0068] 在一些情况下,用两个或更多个核酸片段合成核酸文库。可以用至少两个片段、至少3个片段、至少4个片段、至少5个片段或更多片段合成核酸文库。每个核酸片段的长度或合成的核酸的平均长度可以是至少或大约至少10、15、20、25、30、35、40、45、50、100、150、200、300、400、500、2000个或更多个核苷酸。每个核酸片段的长度或合成的核酸的平均长度可以是至多或大约至多2000、500、400、300、200、150、100、50、45、35、30、25、20、19、18、17、16、15、14、13、12、11、10个或更少的核苷酸。每个核酸片段的长度或合成的核酸的平均长度可以是10-2000、10-500、9-400、11-300、12-200、13-150、14-100、15-50、16-45、17-40、18-35、19-25。

[0069] 各种混合方法,例如通过连接进行的混合,以及试剂,是本领域中已知的,并且可

用于实施本文提供的方法。可利用平端连接将来自一个核酸群体的片段与来自第二核酸群体的片段连接起来。连接酶可包括但不限于大肠杆菌连接酶、T4连接酶、哺乳动物连接酶(例如,DNA连接酶I、DNA连接酶II、DNA连接酶III、DNA连接酶IV)、热稳定连接酶和快速连接酶。在一些情况下,利用PCR延伸重叠法使两个片段退火并连接,以形成更长的核酸。在这样的布置中,第一片段具有与第二片段互补的区域,使得在DNA聚合酶和扩增试剂如dNTP、缓冲溶液和ATP的存在下,每个片段充当另一个片段的引物,以进行从退火位置延伸的扩增反应。在一些情况下,通过在切割限制酶识别区之后进行连接,将来自一个核酸群体的片段与来自第二核酸群体的片段连接起来。在一些情况下,限制酶产生突出端,然后通过连接酶连接这些突出端。可以采用一个核酸片段与另一个核酸片段的1:1摩尔比。在一些情况下,该摩尔比为至少1:1、至少1:2、至少1:3、至少1:4或更大。或者,该摩尔比可以是至少2:1、至少3:1、至少4:1或更大。所连接的核酸片段的总摩尔质量或每个核酸片段的摩尔质量可以是至少或至少约1、10、20、30、40、50、100、250、500、750、1000、2000、3000、4000、5000、6000、7000、8000、9000、10000、25000、50000、75000、100000皮摩尔或更高。

[0070] 在一些情况下,通过本文所述方法生成的核酸片段在连接之前进行平端化。可以使用T4 DNA聚合酶或Klenow片段对核酸进行平端化。或者,使用直接产生平端的酶(例如,Sma I、Dpn I、Pvu II、Eco RV)。在一些情况下,使用DNA核酸内切酶或DNA核酸外切酶产生平端。

[0071] 在第二示例性工作流程中,生成饱和组合文库。参见图2,第一核酸群体210在位置1、2、3和4处表现出多样性。第二核酸群体220在位置5、6、7和8处表现出多样性。如图2所示,基因片段“左侧”的核酸群体210具有 4^4 的多样性。基因片段“右侧”的核酸群体220具有 4^4 的多样性。然后可以合成长的基因片段,其在所需基因的“左”半部分具有多样性,与在所需基因的“右”半部分具有多样性的另一个片段组合,产生 4^8 的总多样性。每个核酸片段的长度或合成的核酸的平均长度可以是至少或大约至少10、15、20、25、30、35、40、45、50、100、150、200、300、400、500、2000个或更多个核苷酸。每个核酸片段的长度或合成的核酸的平均长度可以是至多或大约至多2000、500、400、300、200、150、100、50、45、35、30、25、20、19、18、17、16、15、14、13、12、11、10个或更少的核苷酸。每个核酸片段的长度或合成的核酸的平均长度可以是10-2000、10-500、9-400、11-300、12-200、13-150、14-100、15-50、16-45、17-40、18-35、19-25。

[0072] 可以验证所得到的核酸。在一些情况下,通过测序验证核酸。在一些情况下,通过高通量测序,例如通过下一代测序来验证核酸。测序文库的测序可以使用任何合适的测序技术进行,包括但不限于单分子实时(SMRT)测序、聚合酶克隆(Polony)测序、连接测序、可逆终止子测序、质子检测测序、离子半导体测序、纳米孔测序、电子测序、焦磷酸测序、Maxam-Gilbert测序、链终止(例如Sanger)测序、+S测序或合成测序。

[0073] 本文提供了合成高度准确的、在其变异程度上是非饱和的或饱和的核酸文库的方法。在一些情况下,约70%的核酸是无插入和缺失的。在一些情况下,至少60%、65%、70%、75%、80%、85%、90%、95%、99%或超过99%的核酸是无插入和缺失的。在一些情况下,约60%、65%、70%、75%、80%、85%、90%、95%、99%或超过99%的核酸是无插入和缺失的。在一些情况下,超过90%的核酸是无插入和缺失的。在一些情况下,至少80%的核酸没有错误。在一些情况下,至少约70%、75%、80%、85%、90%、95%、99%或更多的核酸没有

错误。

[0074] 本文提供了合成高度准确的、在其变异程度上是非饱和的或饱和的核酸文库的方法。在一些情况下,本文所述的从头合成的核酸文库中超过80%的核酸在扩增后整个文库的平均呈现度的至少约1.5倍以内被呈现。在一些情况下,本文所述的从头合成的核酸文库中超过80%的核酸在扩增后整个文库的平均呈现度的至少约1.5倍、2倍、2.5倍、3倍、3.5倍或4倍以内被呈现。在一些情况下,本文所述的从头合成的核酸文库中超过90%的核酸在扩增后整个文库的平均呈现度的至少约1.5倍以内被呈现。在一些情况下,本文所述的从头合成的核酸文库中超过90%的核酸在扩增后整个文库的平均呈现度的至少约1.5倍、2倍、2.5倍、3倍、3.5倍或4倍以内被呈现。在一些情况下,本文所述的从头合成的核酸文库中超过80%的核酸在扩增后整个文库的平均呈现度的至少约2倍以内被呈现。在一些情况下,本文所述的从头合成的核酸文库中超过80%的核酸在扩增后整个文库的平均呈现度的至少约2倍以内被呈现。

[0075] 代表性核酸文库的生成

[0076] 本文描述了用于合成具有变异密码子编码区的预选分布的核酸文库的方法。而且,这样的文库对于预选的分布可以是非饱和的,同时提供对代表性分布的了解。本文还提供了与核酸生成有关的方法,所述核酸一旦被翻译,即可在特定位置提供预选的氨基酸分布。通过从预选的分布中生成随机样本,设计了低于饱和的核酸文库,使其代表性分布接近于预选的群体分布。具有接近预选群体分布的代表性分布的本文所述核酸文库可以进一步包括以期望的预选分布精确引入每个预期的变体。

[0077] 本文所述的计算技术包括但不限于随机采样。在第一过程中,对于每个位置处的密码子变异的预选分布,计算每个位置的累积分布值。在一些情况下,累积分布值映射到约0.0至1.0之间的概率。对于核酸群体,累积分布值用于确定在特定位置处的密码子变体的可能性。例如,将密码子变体在整个核酸群体中在每个位置出现的次数相加,然后可以确定每个氨基酸在每个位置出现的百分比。然后将核酸样本群体中的百分比与预选的分布进行比较。在群体中具有足够数量的核酸时,会生成与预选分布匹配的样本分布。在一些情况下,进行的采样是采用均匀随机采样的蒙特卡洛 (Monte Carlo) 采样的形式。

[0078] 在一些情况下,与饱和核酸文库相比,经设计并合成以具有预选分布的核酸文库编码约1%、5%、10%、15%、20%、25%、30%、35%、40%、45%、50%、55%、60%或超过60%的不同核酸。在一些情况下,与饱和核酸文库相比,经设计并合成以具有预选分布的核酸文库编码至少1%、5%、10%、15%、20%、25%、30%、35%、40%、45%、50%、55%、60%或超过60%的不同核酸。

[0079] 在一些情况下,与较大的核酸文库相比,经设计并合成以具有预选分布的核酸文库编码约1%、5%、10%、15%、20%、25%、30%、35%、40%、45%、50%、55%、60%或超过60%的不同核酸。在一些情况下,与较大的核酸文库相比,经设计并合成以具有预选分布的核酸文库编码至少1%、5%、10%、15%、20%、25%、30%、35%、40%、45%、50%、55%、60%或超过60%的不同核酸。

[0080] 在一些情况下,来自较大变异核酸文库的代表性亚群中经设计并合成的核酸的数目在约50-100000、100-75000、250-50000、500-25000和1000-15000、2000-10000和4000-8000个序列的范围内。在一些情况下,核酸群体是500个序列。在一些情况下,核酸群体是

5000、10000或15000个序列。在一些情况下,核酸群体具有至少50、100、150、500、1000、2000、5000、10000、20000、50000、100000、200000、400000、800000、1000000个或更多个不同的序列。在一些情况下,每个核酸群体是至多50、100、500、1000、2000、5000、10000、20000、50000、100000、200000、400000、800000或1000000个。

[0081] 在一些情况下,通过组合方法合成核酸文库以达到变异密码子编码区的预选分布呈现出预测多样性的70%至99%。在一些情况下,通过组合方法合成核酸文库以达到变异密码子编码区的预选分布呈现出预测多样性的至少70%。在一些情况下,通过组合方法合成核酸文库以达到变异密码子编码区的预选分布呈现出预测多样性的70%至75%、70%至80%、70%至85%、70%至90%、70%至95%、70%至97%、70%至99%、75%至80%、75%至85%、75%至90%、75%至95%、75%至97%、75%至99%、80%至85%、80%至90%、80%至95%、80%至97%、80%至99%、85%至90%、85%至95%、85%至97%、85%至99%、90%至95%、90%至97%、90%至99%、95%至97%、95%至99%或97%至99%。在一些情况下,合成的代表性核酸群体所呈现出的多样性是预测多样性的至少或大约60%、65%、70%、75%、80%、85%、90%、95%或超过95%。在一些情况下,合成的代表性核酸群体所呈现出的多样性是预测多样性的99%。

[0082] 使用组合方法生成代表性核酸文库

[0083] 本文提供了通过组合方法合成核酸文库以达到变异密码子编码区的预选分布的方法。在一些情况下,将用作合成核酸群体的变体的模板的参考序列分开,使得第一部分是核酸的第一变体群体的参考序列,而第二部分是核酸的第二变体群体的参考序列。

[0084] 在一些情况下,使用本文所述的随机采样方法为来自较大变体文库的部分生成代表性变体分布。合成代表完整参考序列第一部分的变体的第一代表性核酸群体和代表完整参考序列第二部分的变体的第二代表性核酸群体,然后通过连接,例如通过平端连接或通过本领域已知的一些其它生物化学技术进行组合。在一些情况下,所得到的核酸文库为饱和的。在一些情况下,所得到的核酸文库为非饱和的。

[0085] 在一些情况下,用两个或更多个变异核酸群体合成核酸文库,当这些群体连接时,产生所需的更长的核酸变体文库。可以用至少2、3、4、5、6、7、8、9、10个或多于10个群体合成核酸文库,每个群体编码参考核酸的不同区域。在一些情况下,每个核酸群体在约50-100000、100-75000、250-50000、500-25000和1000-15000、2000-10000和4000-8000个序列的范围内。在一些情况下,每个核酸群体是约500、1000、5000、10000、15000个或更多个序列。在一些情况下,每个核酸群体是至少50、100、150、500、1000、2000、5000、10000、20000、50000、100000、200000、400000、800000、1000000个或更多。在一些情况下,每个核酸群体是至多50、100、500、1000、2000、5000、10000、20000、50000、100000、200000、400000、800000和1000000个。

[0086] 在一些情况下,通过组合方法合成核酸文库以达到变异密码子编码区的预选分布呈现出预测多样性的70%至99%。在一些情况下,通过组合方法合成核酸文库以达到变异密码子编码区的预选分布呈现出预测多样性的至少70%。在一些情况下,通过组合方法合成核酸文库以达到变异密码子编码区的预选分布呈现出预测多样性的70%至75%、70%至80%、70%至85%、70%至90%、70%至95%、70%至97%、70%至99%、75%至80%、75%至85%、75%至90%、75%至95%、75%至97%、75%至99%、80%至85%、80%至90%、80%至

95%、80%至97%、80%至99%、85%至90%、85%至95%、85%至97%、85%至99%、90%至95%、90%至97%、90%至99%、95%至97%、95%至99%或97%至99%。在一些情况下,通过组合方法合成核酸文库以达到变异密码子编码区的预选分布呈现出预测多样性的至少或大约60%、65%、70%、75%、80%、85%、90%、95%或超过95%。在一些情况下,合成的代表性核酸群体所呈现出的多样性是预测多样性的99%。

[0087] 合成后进行PCR诱变

[0088] 通过本文所述的组合方法生成的核酸文库(例如饱和或非饱和的)可以用于PCR诱变方法。在一些情况下,具有预选分布的代表性核酸文库用于PCR诱变方法。在该工作流程中,合成多个多核苷酸,其中每个多核苷酸编码参考核酸序列的预定变体的预定序列。参见附图,图3A-图3D中描绘了示例性工作流程,其中多核苷酸在表面上生成。图3A描绘了具有121个座位的表面的单簇的放大视图。图3B中描绘的每个核酸均为可用于从参考核酸序列扩增以产生变异长核酸文库(图3C)的引物。然后,变异长核酸文库任选地经历转录和/或翻译以生成变异RNA或蛋白质文库,图3D。在该示例性说明中,描绘了具有基本上为平面的表面的装置,其用于从头合成多核苷酸,图3A。在一些情况下,该装置包含一簇座位,其中每个座位为多核苷酸延伸的位点。在一些情况下,单簇包含生成所期望的变异序列文库所需的所有多核苷酸变体。在备选的布置中,板包含未分隔成簇的一片座位。

[0089] 本文提供了在簇内合成多核苷酸(例如,如图3所示),然后在单个簇内扩增多核苷酸的方法。与在没有成簇布置的情况下在整个板上扩增不相同的多核苷酸相比,这样的布置提供改进的核酸呈现。在一些情况下,由于反复合成具有高GC含量的多核苷酸的大多核苷酸群体,在簇内座位表面上合成的多核苷酸的扩增克服了对呈现的负面影响。在一些情况下,本文描述的簇包含约50-1000、75-900、100-800、125-700、150-600、200-500或300-400个离散的座位。在一些情况下,座位是斑点、孔、微孔、通道或柱杆(post)。在一些情况下,每个簇具有至少1X、2X、3X、4X、5X、6X、7X、8X、9X、10X或更高丰余度的支持延伸具有相同序列的多核苷酸的单独特征。在一些情况下,1X丰余度意味着没有具有相同序列的多核苷酸。

[0090] 本文所述的从头合成的多核苷酸文库可包含多个多核苷酸,每个多核苷酸在第一位置(位置“x”)处有至少一个变异序列,并且每个变异多核苷酸在第一轮PCR中用作引物以生成第一延伸产物。在该实例中,第一多核苷酸420中的位置“x”编码变异密码子序列,即来自参考序列的19个可能的变体之一。参见图4A。包含与第一多核苷酸的序列重叠的序列的第二多核苷酸425也在另一轮的PCR中用作引物以生成第二延伸产物。另外,外部引物415、430可用于扩增来自长核酸序列的片段。所得到的扩增产物是长核酸序列的片段435、440。参见图4B。然后使长核酸序列的片段435、440杂交,并经历延伸反应以形成长核酸的变体445。参见图4C。第一和第二延伸产物的重叠末端可充当第二轮PCR的引物,从而生成含有该变体的第三延伸产物(图4D)。为了提高产率,长核酸的变体在包括DNA聚合酶、扩增试剂和外部引物415、430的反应中进行扩增。在一些情况下,第二多核苷酸包含邻近但不包括变异位点的序列。在备选的布置中,生成具有与第二多核苷酸相重叠的区域的第一多核苷酸。在这种情境下,针对至多19个变体合成在单个密码子处具有变异的第一核酸。第二核酸不包含变异序列。任选地,第一群体包含第一多核苷酸变体和编码不同密码子位点处的变体的其它多核苷酸。或者,第一多核苷酸和第二多核苷酸可被设计用于平端连接。

[0091] 图5A-图5F描绘了备选的诱变PCR方法。在这样的过程中,包含第一和第二链505、510的模板核酸分子500在含有第一引物515和第二引物520的PCR反应中扩增(图5A)。扩增反应包括作为核苷酸试剂的尿嘧啶。生成尿嘧啶标记的延伸产物525(图5B),任选地进行纯化,并且充当使用第一多核苷酸535和多个第二多核苷酸530生成第一延伸产物540和545的后续PCR反应的模板(图5C-图5D)。在该过程中,多个多核苷酸530包含编码变异序列的多核苷酸(在图5C中表示为X、Y和Z)。尿嘧啶标记的模板核酸用尿嘧啶特异性切除试剂,例如从New England Biolabs商购获得的USER digest进行消化。添加变体535和具有变体X、Y和Z的不同密码子530,并且进行有限的PCR步骤以生成图5D。在将含尿嘧啶的模板消化后,延伸产物的重叠末端用来引发PCR反应,其中第一延伸产物540和545与第一外部引物550和第二外部引物555组合起到引物的作用,从而生成在变异位点处含有多个变体X、Y和Z的核酸分子560的文库,图5F。

[0092] 具有长核酸的变体和非变体部分的群体的从头合成

[0093] 通过本文所述的组合方法生成的核酸文库(例如饱和或非饱和的)可以用于从成长核酸的多个片段,其中至少一个片段以多种形式合成,每种形式具有不同的变异序列。在一些情况下,具有预选分布的代表性核酸文库用于从头合成,其中至少一个片段以多种形式合成,每种形式具有不同的变异序列。在这种布置中,从头合成装配变异长程核酸文库所需的全部片段。合成的片段可具有重叠的序列,使得在合成之后,片段文库经历杂交。杂交后,可进行延伸反应以补平任何互补缺口。

[0094] 或者,合成的片段可以用引物来扩增,随后经历平端连接或重叠杂交。在一些情况下,该装置包含一簇座位,其中每个座位是多核苷酸延伸的位点。在一些情况下,单簇包含预定长核酸的所有多核苷酸变体和其它片段序列,以生成所期望的变异核酸序列文库。该簇可包含约50至500个座位。在一些布置中,簇包含超过500个座位。

[0095] 第一多核苷酸群体中的每个单独的多核苷酸可在簇的单独的、可单独寻址的座位上生成。一个多核苷酸变体可以由多个可单独寻址的座位呈现。第一多核苷酸群体中的每个变体可以呈现1、2、3、4、5、6、7、8、9、10次或更多次。在一些情况下,第一多核苷酸群体中的每个变体在3个或更少的座位处呈现。在一些情况下,第一多核苷酸群体中的每个变体在两个座位处呈现。在一些情况下,第一多核苷酸群体中的每个变体仅在单个座位处呈现。

[0096] 本文提供了生成丰度度降低的核酸文库的方法。在一些情况下,可以在不需要超过1次合成变异核酸的情况下生成变异核酸,以获得所需变异核酸。在一些情况下,本公开提供了在不需要超过1、2、3、4、5次、6、7、8、9、10次或更多次合成变异核酸的情况下生成变异核酸以生成所需变异核酸的方法。

[0097] 可以在不需要在超过1个离散位点处合成变异核酸的情况下生成变异核酸,以获得所需变异核酸。本公开提供了在不需要在超过1个位点、2个位点、3个位点、4个位点、5个位点、6个位点、7个位点、8个位点、9个位点或10个位点处合成变异核酸的情况下生成变异核酸以生成所需变异核酸的方法。在一些情况下,在至多6、5、4、3、2或1个离散位点处合成核酸。相同的核酸可以在表面上的1、2或3个离散座位中合成。

[0098] 在一些情况下,呈现单变异核酸的座位的量是下游加工(例如,扩增反应或细胞试验)所需的核酸材料的量的函数。在一些情况下,呈现单变异核酸的座位的量是单簇中可用座位的函数。

[0099] 本文提供了用于生成核酸文库的方法,该核酸文库包含在参考核酸的多个位点处不同的变异核酸。在这类情况下,每个变体文库均在一簇座位内的可单独寻址的座位上生成。应当理解,由核酸文库呈现的变异位点的数目将取决于该簇中可单独寻址的座位的数目和每个位点处所需变体的数目。在一些情况下,每个簇包含约50至500个座位。在一些情况下,每个簇包含100至150个座位。

[0100] 在示例性布置中,19个变体在变异位点处呈现,其对应于编码19个可能的变异氨基酸中的每一个的密码子。在另一个示例性情况下,61个变体在变异位点处呈现,其对应于编码19个可能的变异氨基酸中的每一个的三联体。在非限制性实例中,簇包含121个可单独寻址的座位。在该实例中,核酸群体包含每个单位点变体的6次重复(6次重复 $\times$ 1个变异位点 $\times$ 19个变体=114个座位)、每个双位点变体的3次重复(3次重复 $\times$ 2个变异位点 $\times$ 19个变体=114个座位)或每个三位点变体的2次重复(2次重复 $\times$ 3个变异位点 $\times$ 19个变体=114个座位)。在一些情况下,核酸群体在四个、五个、六个或超过六个变异位点处包含变体。

[0101] 本文提供了用于产生合成的(即从头合成或化学合成的)核酸的方法和组合物。本文所述的合成核酸的文库可包含多个共同编码一个或多个基因或基因片段的核酸。在一些情况下,核酸文库包含编码序列或非编码序列。在一些情况下,核酸文库编码多个cDNA序列。在一些情况下,核酸文库包含一个或多个核酸,所述一个或多个核酸中的每一个核酸编码多个外显子的序列。本文所述文库内的每个核酸可编码不同的序列,即,不相同的序列。在一些情况下,本文所述文库内的每个核酸包含与该文库内的另一个核酸的序列互补的至少一部分。除非另有说明,否则本文所述的核酸序列可包含DNA或RNA。

[0102] 本文提供了用于产生合成的(即从头合成的)基因的方法和组合物。包含合成基因的文库可以通过本文其它部分进一步详述的多种方法来构建,如PCA、非PCA基因装配方法或分层基因装配,从而将两个或更多个双链核酸组合(“缝合”)以产生更大的DNA单元(即,底架)。大构建体的文库可包含长度为至少1、1.5、2、3、4、5、6、7、8、9、10、15、20、30、40、50、60、70、80、90、100、125、150、175、200、250、300、400、500kb或更长的核酸。大构建体可被独立选择的约5000、10000、20000或50000个碱基对的上限所约束。任意数目的编码多肽区段的核苷酸序列的合成可包括编码非核糖体肽(NRP)的序列,编码以下物质的序列:非核糖肽合成酶(NRPS)模块和合成变体、其它模块化蛋白质如抗体的多肽区段、来自其它蛋白质家族的多肽区段,包括非编码DNA或RNA,如调节序列,例如启动子、转录因子、增强子、siRNA、shRNA、RNAi、miRNA、衍生自微小RNA的核仁小RNA,或任何感兴趣的功能性或结构性DNA或RNA单元。以下是核酸的非限制性实例:基因或基因片段的编码区或非编码区、基因间DNA、由连锁分析限定的基因座(多个基因座)、外显子、内含子、信使RNA(mRNA)、转移RNA、核糖体RNA、短干扰RNA(siRNA)、短发夹RNA(shRNA)、微小RNA(miRNA)、核仁小RNA、核酶、cDNA(其为mRNA的DNA呈现形式,通常通过信使RNA(mRNA)的逆转录或通过扩增来获得);经合成或通过扩增产生的DNA分子、基因组DNA、重组多核苷酸、支链多核苷酸、质粒、载体、任何序列的分离的DNA、任何序列的分离的RNA、核酸探针和引物。在cDNA的语境中,术语基因或基因片段是指包含至少一个编码外显子序列的区域而没有间插内含子序列的DNA核酸序列。

[0103] 在各个实施方案中,本文所述的方法和组合物涉及基因文库。基因文库可包含多个亚区段。在一个或多个亚区段中,文库的基因可以共价连接在一起。在一个或多个亚区段中,文库的基因可编码具有一个或多个代谢终产物的第一代代谢途径的组成部分。在一个或

多个亚区段中,可以基于一种或多种靶向代谢终产物的制备过程来选择文库的基因。所述一种或多种代谢终产物可以包含生物燃料。在一个或多个亚区段中,文库的基因可以编码具有一种或多种代谢终产物的第二代谢途径的组成部分。第一和第二代谢途径的一种或多种终产物可以包含一种或多种共同的终产物。在一些情况下,第一代谢途径包含在第二代谢途径中操纵的终产物。

[0104] 用于生物体的变异核酸文库

[0105] 通过本文所述的方法生成的变异核酸文库可以编码生物体的至少一个基因。在一些情况下,该核酸文库编码生物体的单个基因、途径或整个基因组。在一些情况下,该变异核酸文库编码基因(例如1000个碱基对)、部分(例如3-10个基因)、途径(例如10-100个基因)或底架(例如,100-1000个基因)中的至少一种。表1提供了模型生物体的非限制性示例性列表。

[0106] 表1.模型生物体和基因编号

	模型生物体	蛋白质编码基因*
[0107]	拟南芥 (<i>Arabidopsis thaliana</i>)	27000
	秀丽隐杆线虫 (<i>Caenorhabditis elegans</i>)	20000
	家犬 (<i>Canis lupus familiaris</i>)	19000
	莱茵衣藻(<i>Chlamydomonas reinhardtii</i>)	14000
	斑马鱼 (<i>Danio rerio</i>)	26000
[0108]	盘基网柄菌 (<i>Dictyostelium discoideum</i>)	13000
	黑腹果蝇 (<i>Drosophila melanogaster</i>)	14000
	大肠杆菌 (<i>Escherichia coli</i>)	4300
	猕猴 (<i>Macaca mulatta</i>)	22000
	小家鼠 (<i>Mus musculus</i>)	20000
	家兔 (<i>Oryctolagus cuniculus</i>)	27000
	大鼠 (<i>Rattus norvegicus</i>)	22000
	酿酒酵母 (<i>Saccharomyces cerevisiae</i>)	6600
	野猪 (<i>Sus scrofa</i>)	21000

[0109] *此处的数字反映蛋白质编码基因的数目,不包括tRNA和非编码RNA。Ron Milo& Rob Phillips,Cell Biology by the Numbers 286(2015)。

[0110] 密码子变异

[0111] 本文所述的变异核酸文库可包含多个核酸,其中每个核酸编码与参考核酸序列相比的变异密码子序列。在一些情况下,第一核酸群体中的每个核酸在单变异位点处含有变

体。在一些情况下,第一核酸群体在单变异位点处含有多个变体,使得第一核酸群体在相同变异位点处含有超过一个变体。第一核酸群体可包含在相同变异位点处共同编码多个密码子变体的核酸。第一核酸群体可包含在相同位置处共同编码多达19个或更多个密码子的核酸。第一核酸群体可包含在相同位置处共同编码多达60个变异三联体的核酸,或者第一核酸群体可包含在相同位置处共同编码多达61个不同密码子三联体的核酸。每个变体可编码在翻译过程中产生不同氨基酸的密码子。表2提供了对于变异位点可能的每个密码子(和代表性氨基酸)的列表。

[0112] 表2.密码子和氨基酸列表

氨基酸	单字母 代码	三字母 代码	密码子					
丙氨酸	A	Ala	GCA	GCC	GCG	GCT		
半胱氨酸	C	Cys	TGC	TGT				
天冬氨酸	D	Asp	GAC	GAT				
谷氨酸	E	Glu	GAA	GAG				
苯丙氨酸	F	Phe	TTC	TTT				
甘氨酸	G	Gly	GGA	GGC	GGG	GGT		
组氨酸	H	His	CAC	CAT				
异亮氨酸	I	Iso	ATA	ATC	ATT			
赖氨酸	K	Lys	AAA	AAG				
亮氨酸	L	Leu	TTA	TTG	CTA	CTC	CTG	CTT
甲硫氨酸	M	Met	ATG					
天冬酰胺	N	Asn	AAC	AAT				
脯氨酸	P	Pro	CCA	CCC	CCG	CCT		
谷氨酰胺	Q	Gln	CAA	CAG				
精氨酸	R	Arg	AGA	AGG	CGA	CGC	CGG	CGT
丝氨酸	S	Ser	AGC	AGT	TCA	TCC	TCG	TCT
苏氨酸	T	Thr	ACA	ACC	ACG	ACT		
缬氨酸	V	Val	GTA	GTC	GTG	GTT		
色氨酸	W	Trp	TGG					
酪氨酸	Y	Tyr	TAC	TAT				

[0114] 本文提供了变异核酸文库,其包含编码与参考核酸序列相比的变异密码子序列的核酸,其中该变异密码子序列基于密码子分配来选择。示例性的密码子分配在表3中示出,其中按从左至右的优先顺序选择变异密码子序列。在一些情况下,密码子分配基于生物体中密码子的频率。示例性生物体包括但不限于动物、植物、真菌、原生生物、古菌或细菌。例如,密码子分配基于大肠杆菌或智人(Homo sapiens)。

[0115] 表3.密码子分配

[0116]

氨基酸	单字母 代码	三字母 代码	密码子				
丙氨酸	A	Ala	GCT	GCA	GCC	GCG	
半胱氨酸	C	Cys	TGC	TGT			
天冬氨酸	D	Asp	GAT	GAC			
谷氨酸	E	Glu	GAG	GAA			
苯丙氨酸	F	Phe	TTC	TTT			
甘氨酸	G	Gly	GGT	GGA	GGC	GGG	
组氨酸	H	His	CAC	CAT			
异亮氨酸	I	Iso	ATC	ATT	ATA		
赖氨酸	K	Lys	AAG	AAA			
亮氨酸	L	Leu	CTG	CTC	CTT	TTG	TTA CTA
甲硫氨酸	M	Met	ATG				
天冬酰胺	N	Asn	AAC	AAT			
脯氨酸	P	Pro	CCT	CCA	CCG	CCC	
谷氨酰胺	Q	Gln	CAG	CAA			
精氨酸	R	Arg	AGA	CGT	AGG	CGA	CGC CGG
丝氨酸	S	Ser	AGC	TCT	TCC	AGT	TCA TCG
苏氨酸	T	Thr	ACC	ACA	ACT	ACG	
缬氨酸	V	Val	GTG	GTT	GTC	GTA	
色氨酸	W	Trp	TGG				
酪氨酸	Y	Tyr	TAC	TAT			
终止密码子			TGA	TAA	TAG		

[0117] 本文提供了变异核酸文库,其包含编码与参考核酸序列相比的变异密码子序列的核酸,其中基于密码子分配的变异密码子序列取决于多种因素。在一些情况下,该变异密码子序列基于密码子序列的复杂性或多样性来选择。例如,选择包含三个不同核碱基的密码子序列,而不是包含两个不同核碱基的密码子序列或包含相同核碱基的密码子序列。在一些情况下,该密码子序列基于下游应用来选择。下游应用包括但不限于使对蛋白质翻译后的表达水平的影响最小化或改善通过下一代测序对变异密码子序列的检测。改善通过下一代测序对变异密码子序列的检测可以包括避免具有高错误率的均聚物。在一些情况下,选择密码子序列,除非该密码子序列导致引起序列破坏的位点,如限制酶位点。

[0118] 基于本文所述密码子分配的变异位点的密码子序列可以是随机化的。在一些情况下,该密码子序列不是随机化的。例如,对于每个肽选择一个突变的单变体文库,密码子序列不是随机化的。在一些情况下,多变体文库包含随机化的密码子序列。

[0119] 核酸群体可包含在多个位置处共同编码至多20个密码子变异的改变的核酸。在这类情况下,该群体中的每个核酸包含在相同核酸中超过一个位置处的密码子变异。在一些

情况下,该群体中的每个核酸包含在单个核酸中的1、2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20个或更多个密码子处的密码子变异。在一些情况下,每个变异长核酸包含在单个长核酸中的1、2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20、21、22、23、24、25、26、27、28、29、30个或更多个密码子处的密码子变异。在一些情况下,该变异核酸群体包含在单个核酸中的1、2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20、21、22、23、24、25、26、27、28、29、30个或更多个密码子处的密码子变异。在一些情况下,该变异核酸群体包含在单个长核酸中的至少约10、20、30、40、50、60、70、80、90、100、125、150、175、200、225、250、275、300个或更多个密码子处的密码子变异。

[0120] 本文提供了其中在含有多个可单独寻址的座位的第二簇上生成第二核酸群体的过程。第二核酸群体可包含对于每个密码子位置而言恒定(即,在每个位置处编码相同的氨基酸)的多个第二核酸。第二核酸可与第一核酸的至少一部分重叠。在一些情况下,第二核酸不包含在第一核酸上所呈现的变异位点。或者,第二核酸群体可包含多个第二核酸,该第二核酸含有至少一个针对一个或多个密码子位置的变异。

[0121] 本文提供了用于合成核酸文库的方法,其中生成在多个密码子位置处包含变体的单个核酸群体。第一核酸群体可在含有多个可单独寻址的座位的第一簇上生成。在这类情况下,第一核酸群体在不同密码子位置处包含变体。在一些情况下,所述不同位点是连续的(即,编码连续的氨基酸)。例如,第一核酸群体在两个连续密码子位置处包含变体,在一个位置处编码多达19个变体。在一些情况下,第一核酸群体在两个连续的密码子位置处包含变体,在一个位置处编码约1至约19个变体。在一些情况下,合成约38个核酸。第一核酸群体可包含在相同或另外的变异位点处共同编码至多19个密码子变体的改变的核酸。第一核酸群体可包括多个第一核酸,其在位置x处含有至多19个变体、在位置y处含有至多19个变体且在位置z处含有至多19个变体。在这样的布置中,每个变体编码不同的氨基酸,使得在每个不同的变异位点处编码至多19个氨基酸变体。在另外的情况下,第二核酸群体在含有多个可单独寻址的座位的第二簇上生成。第二核酸群体可包含对于每个密码子位置而言恒定(即,在每个位置处编码相同的氨基酸)的多个第二核酸。第二核酸可与第一核酸的至少一部分重叠。第二核酸可不包含在第一核酸上所呈现的变异位点。

[0122] 通过本文所述的过程生成的变异核酸文库提供了变异蛋白质文库的生成。在第一个示例性布置中,模板核酸编码序列,该序列在转录并翻译时产生具有多个密码子位置的参考氨基酸序列(图6A),这些位置由单个圆圈表示。模板的核酸变体可使用本文所述的方法生成。在一些情况下,核酸中存在单个变体,导致单变异氨基酸序列(图6B)。在一些情况下,核酸中存在多于一个变体,其中这些变体被一个或多个密码子隔开,导致在变异残基之间具有间隔的蛋白质(图6C)。在一些情况下,核酸中存在多于一个变体,其中这些变体是顺序的并且彼此相邻或连续,导致间隔的变异残基段(图6D)。在一些情况下,核酸中存在两段变体,其中每段变体包含顺序的且相邻或连续的变体(图6E)。

[0123] 本文提供了生成核酸变体文库的方法,其中每个变体包含单位置密码子变体。在一个实例中,模板核酸具有多个密码子位置,其中示例性氨基酸残基由带有它们各自的单字母代码蛋白质密码子的圆圈表示,图7A。图7B描绘了由变异核酸文库编码的氨基酸变体文库,其中每个变体包含位于不同单个位点处的单位置变体(由“X”表示)。第一位置变体用任意密码子来代替丙氨酸,第二个变体用由变异核酸文库编码的任意密码子来代替色氨

酸,第三个变体用任意密码子来代替异亮氨酸,第四个变体用任意密码子来代替赖氨酸,第五个变体用任意密码子来代替精氨酸,第六个变体用任意密码子来代替谷氨酸,而第七个变体用任意密码子来代替谷氨酰胺。当全部或少于全部密码子变体由变异核酸文库编码时,在蛋白质表达(即,DNA转录的标准细胞事件之后进行翻译和加工事件)之后生成相应的氨基酸序列变体群体。

[0124] 在一些布置中,生成具有多位点的单位位置变体的文库。如图8A所示,提供了野生型模板。图8B描绘了具有两个位点的单位位置密码子变体的所得氨基酸序列,其中编码不同氨基酸的每个密码子变体由带不同图案的圆圈表示。

[0125] 本文提供了生成具有一段多位点、单位位置变体的文库的方法。每段核酸可具有1、2、3、4、5个或更多个变体。每段核酸可具有至少1个变体。每段核酸可具有至少2个变体。每段核酸可具有至少3个变体。例如,一段5个核酸可具有1个变体。一段5个核酸可具有2个变体。一段5个核酸可具有3个变体。一段5个核酸可具有4个变体。例如,一段4个核酸可具有1个变体。一段4个核酸可具有2个变体。一段4个核酸可具有3个变体。一段4个核酸可具有4个变体。

[0126] 在一些情况下,单位位置变体可全部编码相同的氨基酸,例如组氨酸。如图9A所示,提供了参考氨基酸序列。在这种布置中,一段核酸编码多位点的单位位置变体,并且在表达时产生具有编码组氨酸的所有单位位置变体的氨基酸序列,图9B。在一些实施方案中,通过本文所述的方法合成的变体文库在所得到的氨基酸序列中未编码多于4个组氨酸残基。

[0127] 在一些情况下,通过本文所述的方法生成的核酸变体文库提供具有单独的变异段的氨基酸序列的表达。图10A中描绘了模板氨基酸序列。一段核酸可以在两个区段中仅具有1个变异密码子,并且当表达时产生图10B中所描绘的氨基酸序列。在图10B中由带不同图案的圆圈描绘变体,以表明氨基酸的变异处于单一区段中不同的位置。

[0128] 本文提供了合成具有1、2、3个或更多个密码子变体的核酸文库的方法和装置,其中选择性地控制每个位点的变体。单位点变体的两种氨基酸之比可以是约1:100、1:50、1:10、1:5、1:3、1:2、1:1。单位点变体的三种氨基酸之比可以是约1:1:100、1:1:50、1:1:20、1:1:10、1:1:5、1:1:3、1:1:2、1:1:1、1:10:10、1:5:5、1:3:3或1:2:2。图11A描绘了由野生型核酸序列编码的野生型参考氨基酸序列。图11B描绘了氨基酸变体文库,其中每个变体包含一段序列(由带图案的圆圈表示),其中每个位置可以在所得到的变异蛋白质文库中具有一定比例的氨基酸。所得到的变异蛋白质文库由通过本文所述方法生成的变异核酸文库编码。在该图示中,5个位置发生改变:第一个位置1100具有50/50的K/R比;第二个位置1110具有50/25/25的V/L/S比,第三个位置1120具有50/25/25的Y/R/D比,第四个位置1130对于所有20种氨基酸具有相等的比例,而第五个位置1140对于G/P具有75/25的比例。本文所述的比例仅是示例性的。

[0129] 在一些情况下,生成合成的变体文库,其编码最终翻译成蛋白质的氨基酸序列的核酸序列。示例性氨基酸序列包括编码小肽以及大肽(例如抗体序列)至少一部分的氨基酸序列。在一些情况下,合成的核酸各自编码抗体序列一部分中的变异密码子。合成的变异核酸之部分所编码的示例性抗体序列包括其抗原结合区或可变区,或其片段。本文所述的核酸编码其一部分的抗体片段实例包括但不限于Fab、Fab'、F(ab')₂和Fv片段,双抗体,线性抗体,单链抗体分子,和由抗体片段形成的多特异性抗体。本文所述的核酸编码其一部分的

示例抗体区域包括但不限于Fc区,Fab区,Fab区的可变区,Fab区的恒定区,重链或轻链的可变区(V_H 或 V_L),或 V_H 或 V_L 的特异性互补决定区(CDR)。通过本文公开的方法生成的变体文库可导致本文所述的一个或多个抗体区域的变异。在一个示例性过程中,生成编码几个CDR的核酸的变体文库。参见图12。编码具有CDR1 1210、CDR21220和CDR3 1230区的抗体的模板核酸通过本文所述的方法进行修饰,其中每个CDR区包含多个变异位点。生成重链或轻链的单个可变域中3个CDR中的每一个的变异1215、1225和1235。每个位点(由星号表示)可包含单个位置、一段多个连续位置或两者,该位置可与不同于模板核酸序列的任何密码子序列互换。变体文库的多样性可通过使用本文提供的方法而显著增加,具有高达约 10^{10} 或更高的多样性。

[0130] 在一些情况下,变体文库包含重链或轻链可变域(V_H 或 V_L)的单个或多个变体。在一些情况下,变体文库在 V_H 区中包含单个或多个变体。示例性 V_H 区包括但不限于IGHV1、IGHV2、IGHV3、IGHV4、IGHV5、IGHV6和IGHV7。在一些情况下,变体文库在 V_L 区中包含单个或多个变体。示例性 V_L 区包括但不限于IGKV1、IGKV2、IGKV3、IGKV4、IGKV5、IGLV1、IGLV2和IGLV3。

[0131] 表达盒中的变异

[0132] 在一些情况下,生成合成的变体文库,其编码表达构建体的一部分。表达构建体的示例性部分包括启动子、开放阅读框和终止区。在一些情况下,表达构建体编码一个、两个、三个或更多个表达盒。如图14所示,可生成核酸文库,其编码在构成表达构建体盒之部分的单独区域的单个位点或多个位点处的密码子变异。为了生成表达两个构建体的盒,合成编码第一启动子1410、第一开放阅读框1420、第一终止子1430、第二启动子1440、第二开放阅读框1450或第二终止子序列1460的变异序列的至少一部分的变异核酸。如前述实例中所述,在数轮扩增后,生成具有1,024个表达构建体的文库。图14提供了一个示例性布置。在一些情况下,另外的调节序列如非翻译调节区(UTR)或增强子区也包括在本文提到的表达盒中。表达盒可包含1、2、3、4、5、6、7、8、9、10个或更多个组分,其变异序列通过本文所述的方法生成。在一些情况下,该表达构建体在多顺反子载体中包含多于一个基因。在一个实例中,将合成的DNA核酸插入到病毒载体(例如,慢病毒)中,随后包装以供转导至细胞中,或者插入到非病毒载体中以便转移至细胞中,随后进行筛选和分析。

[0133] 本文公开的用于插入核酸的表达载体包含真核(例如,细菌和真菌)和原核(例如,哺乳动物、植物和昆虫)表达载体。示例性表达载体包括但不限于哺乳动物表达载体:pSF-CMV-NEO-NH2-PPT-3XFLAG、pSF-CMV-NEO-COOH-3XFLAG、pSF-CMV-PURO-NH2-GST-TEV、pSF-OXB20-COOH-TEV-FLAG(R)-6His(“6His”被披露为SEQ ID NO:32)、pCEP4 pDEST27、pSF-CMV-Ub-KrYFP、pSF-CMV-FMDV-daGFP、pEF1a-mCherry-N1载体、pEF1a-tdTomato载体、pSF-CMV-FMDV-Hygro、pSF-CMV-PGK-Puro、pMCP-tag(m)和pSF-CMV-PURO-NH2-CMYC;细菌表达载体:pSF-OXB20-BetaGal、pSF-OXB20-Fluc、pSF-OXB20和pSF-Tac;植物表达载体:pRI 101-AN DNA和pCambia2301;和酵母表达载体:pTYB21和pKLAC2,以及昆虫载体:pAc5.1/V5-His A和pDEST8。示例性细胞包括但不限于原核细胞和真核细胞。示例性真核细胞包括但不限于动物、植物和真菌细胞。示例性动物细胞包括但不限于昆虫、鱼和哺乳动物细胞。示例性哺乳动物细胞包括小鼠、人和灵长类动物细胞。通过本文所述的方法合成的核酸可以通过本领域已知的各种方法(包括但不限于转染、转导和电穿孔)转移至细胞中。所测试的示例性细胞功能包括但不限于细胞增殖、迁移/粘附、代谢和细胞信号传导活性的改变。

[0134] 高度平行的核酸合成

[0135] 本文提供了一种平台方法,其利用从多核苷酸合成到硅上纳米孔内基因装配的端到端过程的小型化、平行化及垂直整合来创建革命性的合成平台。本文所述的装置采用与96孔板相同的占地面积(footprint)提供了这样一种硅合成平台,与传统合成方法相比,该硅合成平台能够将通量提高高达1,000倍或更多,其中在单次高度平行化运行中产生高达约1,000,000个或更多个多核苷酸或10,000个或更多个基因。

[0136] 随着新一代测序的出现,高分辨率基因组数据已成为深入研究各种基因在正常生物学和疾病发病机理中的生物学作用的研究的重要因素。本研究的核心是分子生物学的中心法则和“连续信息的逐残基转移”的概念。将DNA中编码的基因组信息转录成信息,随后将其翻译成蛋白质,该蛋白质是给定生物学途径内的活性产物。

[0137] 另一个令人兴奋的研究领域是关于着眼于高度特异性细胞靶标的治疗性分子的发现、研发和制备。高度多样性的DNA序列文库是靶向治疗剂的开发流程的核心。在设计、构建和测试蛋白质工程循环中使用基因突变体表达蛋白质,在理想情况下该循环得到针对对其治疗靶标具有高亲和力的蛋白质的高度表达而优化的基因。作为实例,考虑受体的结合口袋。同时测试结合口袋内所有残基的所有序列排列的能力将允许进行彻底的探索,从而增加成功的可能性。饱和诱变(其中研究人员试图在受体内的特定位点处生成所有可能的突变)代表了针对这种开发挑战的一种方法。虽然其成本高、耗时且耗力,但它能够将每个变体引入到每个位置。相反,组合诱变(其中几个选定的位置或短DNA段可得到广泛修饰)生成具有偏向呈现的变体的不完全组库。

[0138] 为了加速药物开发流程,具有在可用于测试的正确位置处以预期频率可获得的所需变体的文库(换言之,精确文库)使得能够降低成本以及筛选的周转时间。本文提供了用于合成核酸合成变体文库的方法,其能够以所需的频率精确引入每种期望的变体。对于最终用户来说,这意味着不仅能够彻底对序列空间进行采样,而且能够以有效的方式查询这些假设,从而降低成本和筛选时间。全基因组编辑可以阐明重要的途径,可以检测每个变体和序列排列以获得最佳功能性的文库,并且可以使用数以千计的基因重建整个途径和基因组,以重新改造生物系统以供药物发现。

[0139] 在第一个实例中,药物本身可以使用本文所述的方法进行优化。例如,为了改善抗体的指定功能,设计并合成编码抗体一部分的变异核酸文库。然后通过本文所述的过程(例如,PCR诱变之后插入载体中)生成抗体的变异核酸文库。然后在生产细胞系中表达该抗体,并针对增强的活性进行筛选。示例筛选包括检查对抗原的结合亲和力、稳定性或效应物功能(例如,ADCC、补体或凋亡)的调节。用来优化抗体的示例性区域包括但不限于Fc区、Fab区、Fab区的可变区、Fab区的恒定区、重链或轻链的可变域(V_H 或 V_L)以及 V_H 或 V_L 的特定互补决定区(CDR)。

[0140] 或者,待优化的分子是用作活化剂或竞争性抑制剂的受体结合表位。在合成核酸的变体文库之后,可以将核酸的变体文库插入到载体序列中,随后在细胞中表达。受体抗原可以在细胞(例如,昆虫、哺乳动物或细菌细胞)中表达,随后进行纯化,或者其可以在细胞(例如,哺乳动物细胞)中表达以检测来自序列变异的功能性后果。功能性后果包括但不限于蛋白质表达、结合亲和力和稳定性的变化。细胞功能性后果包括但不限于增殖、生长、粘附、死亡、迁移、能量产生、氧利用、代谢活性、细胞信号传导、老化、对自由基损伤的响应或

其任意组合的变化。在一些实施方案中,为优化而选择的蛋白质的类型是酶、转运蛋白、G蛋白偶联受体、电压门控离子通道、转录因子、聚合酶、衔接蛋白(没有酶活性的蛋白质,用于将两种其它蛋白质结合在一起)和细胞骨架蛋白。酶的示例性类型包括但不限于信号传导酶(如蛋白激酶、蛋白磷酸酶、磷酸二酯酶、组蛋白脱乙酰酶和GTP酶)。

[0141] 本文提供了包含参与整个途径或整个基因组的分子的变体的变异核酸文库。示例性的途径包括但不限于代谢、细胞死亡、细胞周期进展、免疫细胞活化、炎症应答、血管生成、淋巴生成、低氧和氧化应激应答或细胞粘附/迁移途径。细胞死亡途径中的示例性蛋白质包括但不限于Fas、Cadd、胱天蛋白酶3、胱天蛋白酶6、胱天蛋白酶8、胱天蛋白酶9、胱天蛋白酶10、IAP、TNFR1、TNF、TNFR2、NF- κ B、TRAFs、ASK、BAD和Akt。细胞周期途径中的示例性蛋白质包括但不限于NF κ B、E2F、Rb、p53、p21、细胞周期蛋白A、细胞周期蛋白B、细胞周期蛋白D、细胞周期蛋白E和cdc 25。细胞迁移途径中的示例性蛋白质包括但不限于Ras、Raf、PLC、丝切蛋白、MEK、ERK、MLP、LIMK、ROCK、RhoA、Src、Rac、肌球蛋白II、ARP2/3、MAPK、PIP2、整联蛋白、踝蛋白、kindlin、migfilin和细丝蛋白。

[0142] 通过本文所述的方法合成的核酸文库可以在各种细胞类型中表达。示例性细胞类型包括原核细胞(例如,细菌和真菌)和真核细胞(例如,植物和动物)。示例性的动物包括但不限于小鼠、兔子、灵长类动物、鱼和昆虫。示例性的植物包括但不限于单子叶植物和双子叶植物。示例性的植物还包括但不限于微藻类,海带,蓝藻细菌和绿色、棕色和红色藻类,小麦,烟草和玉米,水稻,棉花,蔬菜,和水果。

[0143] 通过本文所述的方法合成的核酸文库可以在与疾病状态相关的各种细胞中表达。与疾病状态相关的细胞包括细胞系、组织样品、来自受试者的原代细胞、从受试者扩充的培养细胞或模型系统中的细胞。示例性的模型系统包括但不限于疾病状态的植物和动物模型。

[0144] 通过本文所述的方法合成的核酸文库可以在各种细胞类型中表达以评估细胞活性的变化。示例性的细胞活性包括但不限于增殖、周期进展、细胞死亡、粘附、迁移、增殖、细胞信号传导、能量产生、氧利用、代谢活性和老化、对自由基损伤的响应或其任意组合。

[0145] 为了鉴定与疾病状态的预防、减轻或治疗相关的变异分子,本文所述的变异核酸文库在与疾病状态相关的细胞中表达,或者在可以诱发疾病状态的细胞中表达。在一些情况下,使用药剂在细胞中诱发疾病状态。用于疾病状态诱发的示例性工具包括但不限于Cre/Lox重组系统、LPS炎症诱发和用来诱发低血糖的链脲佐菌素。与疾病状态相关的细胞可以是来自模型系统的细胞或培养的细胞,以及来自具有特定疾病状况的受试者的细胞。示例性疾病状况包括细菌、真菌、病毒、自身免疫性或增生性病症(例如,癌症)。在一些情况下,所述变异核酸文库在模型系统、细胞系或来源于受试者的原代细胞中表达,并针对至少一种细胞活性的改变进行筛选。示例性的细胞活性包括但不限于增殖、周期进展、细胞死亡、粘附、迁移、增殖、细胞信号传导、能量产生、氧利用、代谢活性和老化、对自由基损伤的响应或其任意组合。

[0146] 基底

[0147] 本文提供了包含多个簇的基底,其中每个簇包含多个支持多核苷酸附着和合成的座位。如本文所用的术语“座位”是指结构上的离散区域,其提供了对编码单个预定序列的多核苷酸从该表面延伸的支持。在一些情况下,座位在二维表面(例如,基本上为平面的表

面)上。在一些情况下,座位是指表面上离散的凸起或凹陷的位点,例如孔、微孔、通道或柱杆。在一些情况下,座位的表面包含这样的材料,该材料被活化官能化,以附着至少一个核苷酸以供多核苷酸合成,或者优选地,附着相同核苷酸的群体以供多核苷酸群体合成。在一些情况下,多核苷酸是指编码相同核酸序列的多核苷酸群体。在一些情况下,装置的表面包括基底的一个或多个表面。

[0148] 使用所提供的系统和方法在文库内合成的多核苷酸的平均错误率常常可以小于1/1000、小于1/1250、小于1/1500、小于1/2000、小于1/3000或更低。在一些情况下,使用所提供的系统和方法在文库内合成的多核苷酸的平均错误率小于1/500、1/600、1/700、1/800、1/900、1/1000、1/1100、1/1200、1/1250、1/1300、1/1400、1/1500、1/1600、1/1700、1/1800、1/1900、1/2000、1/3000或更低。在一些情况下,使用所提供的系统和方法在文库内合成的多核苷酸的平均错误率小于1/1000。

[0149] 在一些情况下,与预定序列相比,使用所提供的系统和方法在文库内合成的多核苷酸的总错误率小于1/500、1/600、1/700、1/800、1/900、1/1000、1/1100、1/1200、1/1250、1/1300、1/1400、1/1500、1/1600、1/1700、1/1800、1/1900、1/2000、1/3000或更低。在一些情况下,使用所提供的系统和方法在文库内合成的多核苷酸的总错误率小于1/500、1/600、1/700、1/800、1/900或1/1000。在一些情况下,与预定序列相比,使用本文提供的系统和方法在文库内合成的多核苷酸的总错误率小于1/500或更低。

[0150] 在一些情况下,错误校正酶可用于使用所提供的系统和方法在文库内合成的多核苷酸。在一些情况下,与预定序列相比,经错误校正的多核苷酸的总错误率可小于1/500、1/600、1/700、1/800、1/900、1/1000、1/1100、1/1200、1/1300、1/1400、1/1500、1/1600、1/1700、1/1800、1/1900、1/2000、1/3000或更低。在一些情况下,使用所提供的系统和方法在文库内合成的多核苷酸经错误校正后的总错误率可小于1/500、1/600、1/700、1/800、1/900或1/1000。在一些情况下,使用所提供的系统和方法在文库内合成的多核苷酸经错误校正后的总错误率可小于1/1000。

[0151] 错误率可限制基因合成在产生基因变体文库方面的价值。错误率为1/300时,在1500个碱基对的基因中约0.7%的克隆将是正确的。由于大多数来自多核苷酸合成的错误导致移码突变,所以在这样的文库中超过99%的克隆将不会产生全长蛋白质。将错误率降低75%将使正确克隆的比例提高40倍。本公开的方法和组合物允许快速从头合成大核酸和基因文库,其错误率低于基因合成方法通常观察到的错误率,这是由于合成质量的改善以及能够以大规模平行且具时效性的方式进行的错误校正方法的适用性。因此,可以合成文库,其中在整个文库中或超过80%、85%、90%、93%、95%、96%、97%、98%、99%、99.5%、99.8%、99.9%、99.95%、99.98%、99.99%或更多的文库中具有低于1/300、1/400、1/500、1/600、1/700、1/800、1/900、1/1000、1/1250、1/1500、1/2000、1/2500、1/3000、1/4000、1/5000、1/6000、1/7000、1/8000、1/9000、1/10000、1/12000、1/15000、1/20000、1/25000、1/30000、1/40000、1/50000、1/60000、1/70000、1/80000、1/90000、1/100000、1/125000、1/150000、1/200000、1/300000、1/400000、1/500000、1/600000、1/700000、1/800000、1/900000、1/1000000或更低的碱基插入、缺失、置换或总错误率。本公开的方法和组合物还涉及具有低错误率的大合成核酸和基因文库,该错误率与该文库的至少一个子集中至少30%、40%、50%、60%、70%、75%、80%、85%、90%、93%、95%、96%、97%、98%、99%、

99.5%、99.8%、99.9%、99.95%、99.98%、99.99%或更多的多核苷酸或基因相关,从而涉及与预定/预选序列相比的无错误序列。在一些情况下,文库内的隔离体积中至少30%、40%、50%、60%、70%、75%、80%、85%、90%、93%、95%、96%、97%、98%、99%、99.5%、99.8%、99.9%、99.95%、99.98%、99.99%或更多的多核苷酸或基因具有相同的序列。在一些情况下,与超过95%、96%、97%、98%、99%、99.5%、99.6%、99.7%、99.8%、99.9%或更高的相似性或同一性有关的任意多核苷酸或基因中的至少30%、40%、50%、60%、70%、75%、80%、85%、90%、93%、95%、96%、97%、98%、99%、99.5%、99.8%、99.9%、99.95%、99.98%、99.99%或更多具有相同的序列。在一些情况下,优化与多核苷酸或基因上的指定基因座有关的错误率。因此,作为大文库的部分的一个或多个多核苷酸或基因的给定基因座或多个选定基因座可各自具有低于1/300、1/400、1/500、1/600、1/700、1/800、1/900、1/1000、1/1250、1/1500、1/2000、1/2500、1/3000、1/4000、1/5000、1/6000、1/7000、1/8000、1/9000、1/10000、1/12000、1/15000、1/20000、1/25000、1/30000、1/40000、1/50000、1/60000、1/70000、1/80000、1/90000、1/100000、1/125000、1/150000、1/200000、1/300000、1/400000、1/500000、1/600000、1/700000、1/800000、1/900000、1/1000000或更低的错误率。在各种情况下,这类错误优化的基因座可包含至少1、2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20、25、30、35、40、45、50、60、70、80、90、100、200、300、400、500、600、700、800、900、1000、1500、2000、2500、3000、4000、5000、6000、7000、8000、9000、10000、30000、50000、75000、100000、500000、1000000、2000000、3000000个或更多个基因座。错误优化的基因座可分布到至少1、2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20、25、30、35、40、45、50、60、70、80、90、100、200、300、400、500、600、700、800、900、1000、1500、2000、2500、3000、4000、5000、6000、7000、8000、9000、10000、30000、75000、100000、500000、1000000、2000000、3000000个或更多个多核苷酸或基因。

[0152] 可在使用或不使用错误校正的情形下达到所述错误率。可在整个文库中,或在文库的超过80%、85%、90%、93%、95%、96%、97%、98%、99%、99.5%、99.8%、99.9%、99.95%、99.98%、99.99%或更多中达到所述错误率。

[0153] 本文提供了可包含表面的结构,该表面支持在共同支持物上的可寻址位置处合成具有不同预定序列的多个多核苷酸。在一些情况下,装置为合成超过2,000、5,000、10,000、20,000、30,000、50,000、75,000、100,000、200,000、300,000、400,000、500,000、600,000、700,000、800,000、900,000、1,000,000、1,200,000、1,400,000、1,600,000、1,800,000、2,000,000、2,500,000、3,000,000、3,500,000、4,000,000、4,500,000、5,000,000、10,000,000个或更多个不同的多核苷酸提供支持。在一些情况下,该装置为合成超过2,000、5,000、10,000、20,000、30,000、50,000、75,000、100,000、200,000、300,000、400,000、500,000、600,000、700,000、800,000、900,000、1,000,000、1,200,000、1,400,000、1,600,000、1,800,000、2,000,000、2,500,000、3,000,000、3,500,000、4,000,000、4,500,000、5,000,000、10,000,000个或更多个编码不同序列的多核苷酸提供支持。在一些情况下,至少一部分多核苷酸具有相同的序列或被配置为用相同的序列合成。

[0154] 本文提供了用于制备和增长长度约为5、10、20、30、40、50、60、70、80、90、100、125、150、175、200、225、250、275、300、325、350、375、400、425、450、475、500、600、700、800、900、1000、1100、1200、1300、1400、1500、1600、1700、1800、1900或2000个碱基的多核苷酸的方法

和装置。在一些情况下,所形成的多核苷酸的长度约为5、10、20、30、40、50、60、70、80、90、100、125、150、175、200或225个碱基。多核苷酸的长度可以是至少5、10、20、30、40、50、60、70、80、90或100个碱基。多核苷酸的长度可以是10至225个碱基、12至100个碱基、20至150个碱基、20至130个碱基或30至100个碱基。

[0155] 在一些情况下,多核苷酸在基底的不同座位上合成,其中每个座位支持合成多核苷酸群体。在一些情况下,每个座位支持合成与在另一座位上增长的多核苷酸群体具有不同序列的多核苷酸群体。在一些情况下,装置的座位位于多个簇内。在一些情况下,装置包含至少10、500、1000、2000、3000、4000、5000、6000、7000、8000、9000、10000、11000、12000、13000、14000、15000、20000、30000、40000、50000个或更多个簇。在一些情况下,装置包含超过2,000、5,000、10,000、100,000、200,000、300,000、400,000、500,000、600,000、700,000、800,000、900,000、1,000,000、1,100,000、1,200,000、1,300,000、1,400,000、1,500,000、1,600,000、1,700,000、1,800,000、1,900,000、2,000,000、300,000、400,000、500,000、600,000、700,000、800,000、900,000、1,000,000、1,200,000、1,400,000、1,600,000、1,800,000、2,000,000、2,500,000、3,000,000、3,500,000、4,000,000、4,500,000、5,000,000或10,000,000个或更多个不同的座位。在一些情况下,装置包含约10,000个不同的座位。单簇内的座位的量在不同情况下是不同的。在一些情况下,每个簇包含1、2、3、4、5、6、7、8、9、10、20、30、40、50、60、70、80、90、100、120、130、150、200、300、400、500、1000个或更多个座位。在一些情况下,每个簇包含约50-500个座位。在一些情况下,每个簇包含约100-200个座位。在一些情况下,每个簇包含约100-150个座位。在一些情况下,每个簇包含约109、121、130或137个座位。在一些情况下,每个簇包含约19、20、61、64个或更多个座位。

[0156] 在装置上合成的不同多核苷酸的数目可取决于基底中可用的不同座位的数目。在一些情况下,装置的簇内的座位密度为至少或大约1个座位/ mm^2 、10个座位/ mm^2 、25个座位/ mm^2 、50个座位/ mm^2 、65个座位/ mm^2 、75个座位/ mm^2 、100个座位/ mm^2 、130个座位/ mm^2 、150个座位/ mm^2 、175个座位/ mm^2 、200个座位/ mm^2 、300个座位/ mm^2 、400个座位/ mm^2 、500个座位/ mm^2 、1,000个座位/ mm^2 或更大。在一些情况下,装置包含约10个座位/ mm^2 至约500个座位/ mm^2 、约25个座位/ mm^2 至约400个座位/ mm^2 、约50个座位/ mm^2 至约500个座位/ mm^2 、约100个座位/ mm^2 至约500个座位/ mm^2 、约150个座位/ mm^2 至约500个座位/ mm^2 、约10个座位/ mm^2 至约250个座位/ mm^2 、约50个座位/ mm^2 至约250个座位/ mm^2 、约10个座位/ mm^2 至约200个座位/ mm^2 或约50个座位/ mm^2 至约200个座位/ mm^2 。在一些情况下,簇内两个相邻座位中心的距离为约10 μm 至约500 μm 、约10 μm 至约200 μm 或约10 μm 至约100 μm 。在一些情况下,相邻座位的两个中心的距离为大于约10 μm 、20 μm 、30 μm 、40 μm 、50 μm 、60 μm 、70 μm 、80 μm 、90 μm 或100 μm 。在一些情况下,两个相邻座位的中心的距离为小于约200 μm 、150 μm 、100 μm 、80 μm 、70 μm 、60 μm 、50 μm 、40 μm 、30 μm 、20 μm 或10 μm 。在一些情况下,每个座位具有约0.5 μm 、1 μm 、2 μm 、3 μm 、4 μm 、5 μm 、6 μm 、7 μm 、8 μm 、9 μm 、10 μm 、20 μm 、30 μm 、40 μm 、50 μm 、60 μm 、70 μm 、80 μm 、90 μm 或100 μm 的宽度。在一些情况下,每个座位具有约0.5 μm 至100 μm 、约0.5 μm 至50 μm 、约10 μm 至75 μm 或约0.5 μm 至50 μm 的宽度。

[0157] 在一些情况下,装置内的簇密度为至少或大约1个簇/100 mm^2 、1个簇/10 mm^2 、1个簇/5 mm^2 、1个簇/4 mm^2 、1个簇/3 mm^2 、1个簇/2 mm^2 、1个簇/1 mm^2 、2个簇/1 mm^2 、3个簇/1 mm^2 、4个簇/1 mm^2 、5个簇/1 mm^2 、10个簇/1 mm^2 、50个簇/1 mm^2 或更大。在一些情况下,装置包含约1个簇/10 mm^2 至约10个簇/1 mm^2 。在一些情况下,两个相邻簇的中心的距离小于约50 μm 、100 μm 、

200um、500um、1000um或2000um或5000um。在一些情况下,两个相邻簇的中心的距离为约50um至约100um、约50um至约200um、约50um至约300um、约50um至约500um和约100um至约2000um。在一些情况下,两个相邻簇的中心的距离为约0.05mm至约50mm、约0.05mm至约10mm、约0.05mm至约5mm、约0.05mm至约4mm、约0.05mm至约3mm、约0.05mm至约2mm、约0.1mm至约10mm、约0.2mm至约10mm、约0.3mm至约10mm、约0.4mm至约10mm、约0.5mm至约10mm、约0.5mm至约5mm或约0.5mm至约2mm。在一些情况下,每个簇沿一个维度具有约0.5至2mm、约0.5至1mm或约1至2mm的直径或宽度。在一些情况下,每个簇沿一个维度具有约0.5、0.6、0.7、0.8、0.9、1、1.1、1.2、1.3、1.4、1.5、1.6、1.7、1.8、1.9或2mm的直径或宽度。在一些情况下,每个簇沿一个维度具有约0.5、0.6、0.7、0.8、0.9、1、1.1、1.15、1.2、1.3、1.4、1.5、1.6、1.7、1.8、1.9或2mm的内径或宽度。

[0158] 装置可以是大约标准96孔板的尺寸,例如约100至200mm乘以约50至150mm。在一些情况下,装置具有小于或等于约1000mm、500mm、450mm、400mm、300mm、250mm、200mm、150mm、100mm或50mm的直径。在一些情况下,装置的直径为约25mm至1000mm、约25mm至约800mm、约25mm至约600mm、约25mm至约500mm、约25mm至约400mm、约25mm至约300mm或约25mm至约200mm。装置尺寸的非限制性实例包括约300mm、200mm、150mm、130mm、100mm、76mm、51mm和25mm。在一些情况下,装置具有至少约100mm²、200mm²、500mm²、1,000mm²、2,000mm²、5,000mm²、10,000mm²、12,000mm²、15,000mm²、20,000mm²、30,000mm²、40,000mm²、50,000mm²或更大的平面表面积。在一些情况下,装置的厚度为约50mm至约2000mm、约50mm至约1000mm、约100mm至约1000mm、约200mm至约1000mm或约250mm至约1000mm。装置厚度的非限制性实例包括275mm、375mm、525mm、625mm、675mm、725mm、775mm和925mm。在一些情况下,装置的厚度随直径而变化,并取决于基底的组成。例如,包含硅之外的材料的装置具有与相同直径的硅装置不同的厚度。装置厚度可以取决于所用材料的机械强度,并且该装置必须厚到足以在操作期间支撑其自身重量而不会破裂。在一些情况下,结构包含多个本文所述的装置。

[0159] 表面材料

[0160] 本文提供了包含表面的装置,其中该表面被修饰用于支持在预定位置处的多核苷酸合成,并且具有低错误率、低遗漏率、高产率和高寡核苷酸呈现。在一些实施方案中,本文提供的用于多核苷酸合成的装置的表面由能够被修饰以支持从头多核苷酸合成反应的多种材料制成。在一些情况下,该装置具有足够的导电性,例如,能够跨整个装置或其一部分形成均匀的电场。本文所述的装置可包含柔性材料。示例性柔性材料包括但不限于改性尼龙、未改性的尼龙、硝酸纤维素和聚丙烯。本文所述的装置可包含刚性材料。示例性刚性材料包括但不限于玻璃、熔融石英、硅、二氧化硅、氮化硅、塑料(例如聚四氟乙烯、聚丙烯、聚苯乙烯、聚碳酸酯,及其掺合物)和金属(例如,金、铂)。本文公开的装置可由包含硅、聚苯乙烯、琼脂糖、葡聚糖、纤维素聚合物、聚丙烯酰胺、聚二甲基硅氧烷(PDMS)、玻璃或其任意组合的材料制成。在一些情况下,本文公开的装置使用此处所列材料或本领域已知的其它任何合适材料的组合制成。

[0161] 本文所述示例性材料的拉伸强度的列表提供如下:尼龙(70MPa)、硝酸纤维素(1.5MPa)、聚丙烯(40MPa)、硅(268MPa)、聚苯乙烯(40MPa)、琼脂糖(1-10MPa)、聚丙烯酰胺(1-10MPa)、聚二甲基硅氧烷(PDMS)(3.9-10.8MPa)。本文所述的固体支持物的拉伸强度可以是1至300、1至40、1至10、1至5或3至11MPa。本文所述的固体支持物的拉伸强度可以是约

1、1.5、2、3、4、5、6、7、8、9、10、11、20、25、40、50、60、70、80、90、100、150、200、250、270MPa或更大。在一些情况下,本文所述的装置包含用于多核苷酸合成的固体支持物,其为能够储存在连续环或卷轴中的柔性材料如带或柔性片的形式。

[0162] 杨氏模量衡量材料对弹性(可恢复的)载荷变形的抵抗力。本文所述示例性材料的刚度的杨氏模量列表提供如下:尼龙(3GPa)、硝酸纤维素(1.5GPa)、聚丙烯(2GPa)、硅(150GPa)、聚苯乙烯(3GPa)、琼脂糖(1-10GPa)、聚丙烯酰胺(1-10GPa)、聚二甲基硅氧烷(PDMS)(1-10GPa)。本文所述的固体支持物的杨氏模量可以是1至500、1至40、1至10、1至5或3至11GPa。本文所述的固体支持物的杨氏模量可以是约1、1.5、2、3、4、5、6、7、8、9、10、11、20、25、40、50、60、70、80、90、100、150、200、250、400、500GPa或更大。由于柔性与刚度之间的关系为彼此相反,因此柔性材料具有低杨氏模量并且在负载下其形状显著改变。在一些情况下,本文所述的固体支持物具有表面,该表面具有至少尼龙的柔性。

[0163] 在一些情况下,本文公开的装置包含二氧化硅基质和氧化硅表面层。或者,该装置可以具有氧化硅基质。本文提供的装置的表面可以是纹理化的,导致用于多核苷酸合成的总表面积增加。本文公开的装置可包含至少5%、10%、25%、50%、80%、90%、95%或99%的硅。本文公开的装置可以由绝缘体上硅(SOI)晶片制成。

[0164] 表面结构

[0165] 本文提供了包含凸起和/或凹陷特征的装置。具有这类特征的一个益处是用来支持多核苷酸合成的表面积增加。在一些情况下,具有凸起和/或凹陷特征的装置被称为三维基底。在一些情况下,三维装置包含一个或多个通道。在一些情况下,一个或多个座位包含通道。在一些情况下,通道可通过沉积装置如材料沉积装置进行试剂沉积。在一些情况下,试剂和/或流体收集在与一个或多个通道流体连通的较大的孔中。例如,装置包含对应于多个具有簇的座位的多个通道,并且所述多个通道与该簇的一个孔流体连通。在一些方法中,多核苷酸文库在簇的多个座位中合成。

[0166] 在一些情况下,所述结构被配置为允许用于表面上多核苷酸合成的受控制的流动和质量传递路径。在一些情况下,装置的构造允许在多核苷酸合成过程中质量传递路径、化学暴露次数和/或洗涤功效的受控且均匀的分布。在一些情况下,装置的构造允许增加扫描效率,例如通过提供足以用于增长多核苷酸的体积,使得由增长的多核苷酸所排除的体积占可用于或适合于增长多核苷酸的初始可用体积的不超过50%、45%、40%、35%、30%、25%、20%、15%、14%、13%、12%、11%、10%、9%、8%、7%、6%、5%、4%、3%、2%、1%或更少。在一些情况下,三维结构允许流体的受管控的流动,从而允许化学暴露的快速交换。

[0167] 本文提供了合成1fM、5fM、10fM、25fM、50fM、75fM、100fM、200fM、300fM、400fM、500fM、600fM、700fM、800fM、900fM、1pM、5pM、10pM、25pM、50pM、75pM、100pM、200pM、300pM、400pM、500pM、600pM、700pM、800pM、900pM或更多的量的DNA的方法。在一些情况下,多核苷酸文库可跨越基因的约1%、2%、3%、4%、5%、10%、15%、20%、30%、40%、50%、60%、70%、80%、90%、95%或100%的长度。基因可以变化最多约1%、2%、3%、4%、5%、10%、15%、20%、30%、40%、50%、60%、70%、80%、85%、90%、95%或100%。

[0168] 不同的多核苷酸可以共同编码基因的至少1%、2%、3%、4%、5%、10%、15%、20%、30%、40%、50%、60%、70%、80%、85%、90%、95%或100%的序列。在一些情况下,多核苷酸可以编码基因的50%、60%、70%、80%、85%、90%、95%或更多的序列。在一些情

况下,多核苷酸可以编码基因的80%、85%、90%、95%或更多的序列。

[0169] 在一些情况下,通过物理结构实现隔离。在一些情况下,通过表面的差异官能化以生成用于多核苷酸合成的活化和钝化区域来实现隔离。差异官能化还可通过在整个装置表面上交替呈现疏水性,从而造成可引起沉积的试剂结珠或润湿的水接触角效应来实现。采用较大的结构可减少飞溅和邻近斑点的试剂对不同的多核苷酸合成位置的交叉污染。在一些情况下,使用装置如多核苷酸合成仪将试剂沉积到不同的多核苷酸合成位置。具有三维特征的基底以允许以低错误率(例如,小于约1:500、1:1000、1:1500、1:2,000;1:3,000;1:5,000;或1:10,000)合成大量多核苷酸(例如,超过约10,000个)的方式配置。在一些情况下,装置包含密度为大约或大于约1、5、10、20、30、40、50、60、70、80、100、110、120、130、140、150、160、170、180、190、200、300、400或500个特征/ mm^2 的特征。

[0170] 装置的孔可具有与基底的另一个孔相同或不同的宽度、高度和/或容积。装置的通道可具有与基底的另一个通道相同或不同的宽度、高度和/或容积。在一些情况下,簇的宽度为约0.05mm至约50mm、约0.05mm至约10mm、约0.05mm至约5mm、约0.05mm至约4mm、约0.05mm至约3mm、约0.05mm至约2mm、约0.05mm至约1mm、约0.05mm至约0.5mm、约0.05mm至约0.1mm、约0.1mm至约10mm、约0.2mm至约10mm、约0.3mm至约10mm、约0.4mm至约10mm、约0.5mm至约10mm、约0.5mm至约5mm或约0.5mm至约2mm。在一些情况下,包含簇的孔的宽度为约0.05mm至约50mm、约0.05mm至约10mm、约0.05mm至约5mm、约0.05mm至约4mm、约0.05mm至约3mm、约0.05mm至约2mm、约0.05mm至约1mm、约0.05mm至约0.5mm、约0.05mm至约0.1mm、约0.1mm至约10mm、约0.2mm至约10mm、约0.3mm至约10mm、约0.4mm至约10mm、约0.5mm至约10mm、约0.5mm至约5mm或约0.5mm至约2mm。在一些情况下,簇的宽度为小于或约5mm、4mm、3mm、2mm、1mm、0.5mm、0.1mm、0.09mm、0.08mm、0.07mm、0.06mm或0.05mm。在一些情况下,簇的宽度约为1.0至约1.3mm。在一些情况下,簇的宽度约为1.150mm。在一些情况下,孔的宽度为小于或约5mm、4mm、3mm、2mm、1mm、0.5mm、0.1mm、0.09mm、0.08mm、0.07mm、0.06mm或0.05mm。在一些情况下,孔的宽度约为1.0至1.3mm。在一些情况下,孔的宽度约为1.150mm。在一些情况下,簇的宽度约为0.08mm。在一些情况下,孔的宽度约为0.08mm。簇的宽度可以指二维或三维基底内的簇。

[0171] 在一些情况下,孔的高度为约20um至约1000um、约50um至约1000um、约100um至约1000um、约200um至约1000um、约300um至约1000um、约400um至约1000um或约500um至约1000um。在一些情况下,孔的高度小于约1000um、小于约900um、小于约800um、小于约700um或小于约600um。

[0172] 在一些情况下,装置包含对应于簇内多个座位的多个通道,其中通道的高度或深度为约5um至约500um、约5um至约400um、约5um至约300um、约5um至约200um、约5um至约100um、约5um至约50um或约10um至约50um。在一些情况下,通道的高度小于100um、小于80um、小于60um、小于40um或小于20um。

[0173] 在一些情况下,通道、座位(例如,在基本上为平面的基底中)或通道和座位两者(例如,在其中座位对应于通道的三维装置中)的直径为约1um至约1000um、约1um至约500um、约1um至约200um、约1um至约100um、约5um至约100um或约10um至约100um,例如约90um、80um、70um、60um、50um、40um、30um、20um或10um。在一些情况下,通道、座位或通道和座位两者的直径小于约100um、90um、80um、70um、60um、50um、40um、30um、20um或10um。在一

些情况下,两个相邻通道、座位或通道和座位两者的中心的距离为约1 μm 至约500 μm 、约1 μm 至约200 μm 、约1 μm 至约100 μm 、约5 μm 至约200 μm 、约5 μm 至约100 μm 、约5 μm 至约50 μm 或约5 μm 至约30 μm ,例如约20 μm 。

[0174] 表面修饰

[0175] 在各种情况下,采用表面修饰通过加成工艺或减成工艺对表面进行化学和/或物理改变,以改变装置表面或装置表面的选定位点或区域的一种或多种化学和/或物理性质。例如,表面修饰包括但不限于:(1)改变表面的润湿性质;(2)对表面进行官能化,即,提供、修改或取代表面官能团;(3)对表面进行去官能化,即,移除表面官能团;(4)以其它方式例如通过刻蚀来改变表面的化学组成;(5)增大或减小表面粗糙度;(6)在表面上提供涂层,例如,展现出与表面的润湿性质不同的润湿性质的涂层;和/或(7)在表面上沉积微粒。

[0176] 在一些情况下,在表面顶部添加化学层(被称为粘附促进剂)有利于基底表面上的座位的结构化图案化。用于施加粘附促进剂的示例性表面包括但不限于玻璃、硅、二氧化硅和氮化硅。在一些情况下,该粘附促进剂是具有高表面能的化学品。在一些情况下,在基底的表面上沉积第二化学层。在一些情况下,第二化学层具有低表面能。在一些情况下,涂覆在表面上的化学层的表面能支持小液滴在表面上的定位。根据所选择的图案化布置,座位的接近度和/或在座位处的流体接触面积是可改变的。

[0177] 在一些情况下,(例如为了多核苷酸合成)多核苷酸或其它部分所沉积到的装置表面或解析座位是光滑的或基本上为平面的(例如,二维的),或者具有不规则性,诸如凸起或凹陷特征(例如,三维特征)。在一些情况下,用一个或多个不同的化合物层来修饰装置表面。感兴趣的此类修饰层包括但不限于无机层和有机层,如金属、金属氧化物,聚合物、有机小分子等。非限制性聚合物层包括肽、蛋白质、核酸或其模拟物(例如,肽核酸等)、多糖、磷脂、聚氨酯、聚酯、聚碳酸酯、聚脲、聚酰胺、聚乙烯胺、聚芳硫醚、聚硅氧烷、聚酰亚胺、聚乙酸酯,以及本文所述的或本领域已知的其它任何合适的化合物。在一些情况下,聚合物为杂聚物。在一些情况下,聚合物为均聚物。在一些情况下,聚合物包含官能部分或是缀合的。

[0178] 在一些情况下,使用增大和/或减小表面能的一个或多个部分对装置的解析座位进行官能化。在一些情况下,部分是化学惰性的。在一些情况下,部分被配置为支持所需的化学反应,例如在多核苷酸合成反应中的一个或多个过程。表面的表面能或疏水性是决定核苷酸附着到该表面上的亲和力的因素。在一些情况下,装置官能化方法可包括:(a)提供具有包含二氧化硅的表面的装置;和(b)使用本文所述的或本领域已知的合适的硅烷化剂(例如,有机官能烷氧基硅烷分子)对所述表面进行硅烷化。

[0179] 在一些情况下,所述有机官能烷氧基硅烷分子包括二甲基氯-十八烷基-硅烷、甲基二氯-十八烷基-硅烷、三氯-十八烷基-硅烷、三甲基-十八烷基-硅烷、三乙基-十八烷基-硅烷或其任意组合。在一些情况下,装置表面用聚乙烯/聚丙烯来官能化(通过 γ 辐射或铬酸氧化并还原成羟烷基表面来官能化)、包含高度交联的聚苯乙烯-二乙烯基苯(通过氯甲基化来衍生化,并胺化成苄胺官能表面)、尼龙(末端氨基己基基团是直接反应性的)或以还原的聚四氟乙烯来刻蚀。在通过引用整体并入本文的美国专利5474796中描述了其它方法和官能化剂。

[0180] 在一些情况下,装置表面通常经由存在于装置表面上的反应性亲水部分,在有效地将硅烷偶联至装置表面的反应条件下,使装置表面与含有硅烷混合物的衍生化组合物相

接触来进行官能化。硅烷化一般通过使用有机官能烷氧基硅烷分子自装配来覆盖表面。

[0181] 还可使用本领域当前已知的多种硅氧烷官能化试剂,例如用于降低或增大表面能。有机官能烷氧基硅烷可根据其有机官能来分类。

[0182] 本文提供了可包含能够与核苷偶联的试剂的图案化的装置。在一些情况下,装置可以涂覆有活性剂。在一些情况下,装置可以涂覆有钝化剂。包含在本文所述的涂层材料中的示例性活性剂包括但不限于N-(3-三乙氧基甲硅烷基丙基)-4-羟基丁酰胺(HAPS)、11-乙酰氧基十一烷基三乙氧基硅烷、正癸基三乙氧基硅烷、(3-氨丙基)三甲氧基硅烷、(3-氨丙基)三乙氧基硅烷、3-缩水甘油基氧基丙基三甲氧基硅烷(GOPS)、3-碘-丙基三甲氧基硅烷、丁基-醛-三甲氧基硅烷、二聚仲氨基烷基硅氧烷、(3-氨丙基)-二乙氧基-甲基硅烷、(3-氨丙基)二甲基-乙氧基硅烷和(3-氨丙基)-三甲氧基硅烷、(3-缩水甘油基氧基丙基)-二甲基-乙氧基硅烷、缩水甘油基氧基-三甲氧基硅烷、(3-巯基丙基)-三甲氧基硅烷,3-4环氧环己基-乙基三甲氧基硅烷以及(3-巯基丙基)-甲基-二甲氧基硅烷、烯丙基三氯氯硅烷、7-辛-1-烯基三氯氯硅烷或双(3-三甲氧基甲硅烷基丙基)胺。

[0183] 包含在本文所述的涂层材料中的示例性钝化剂包括但不限于全氟辛基三氯硅烷;十三氟-1,1,2,2-四氢辛基三氯硅烷;1H,1H,2H,2H-氟辛基三乙氧基硅烷(FOS);三氯(1H,1H,2H,2H-全氟辛基)硅烷;叔丁基-[5-氟-4-(4,4,5,5-四甲基-1,3,2-二氧杂环戊硼烷-2-基)吡啶-1-基]-二甲基-硅烷;CYTOPTM;FluorinertTM;全氟辛基三氯硅烷(PFOTCS);全氟辛基二甲基氯硅烷(PFODCS);全氟癸基三乙氧基硅烷(PFDTES);五氟苯基-二甲基丙基氯-硅烷(PFP TES);全氟辛基三乙氧基硅烷;全氟辛基三甲氧基硅烷;辛基氯硅烷;二甲基氯-十八烷基-硅烷;甲基二氯-十八烷基-硅烷;三氯-十八烷基-硅烷;三甲基-十八烷基-硅烷;三乙基-十八烷基-硅烷;或十八烷基三氯硅烷。

[0184] 在一些情况下,官能化剂包括烃硅烷,如十八烷基三氯硅烷。在一些情况下,官能化剂包括11-乙酰氧基十一烷基三乙氧基硅烷、正癸基三乙氧基硅烷、(3-氨丙基)三甲氧基硅烷、(3-氨丙基)三乙氧基硅烷、缩水甘油基氧基丙基/三甲氧基硅烷和N-(3-三乙氧基甲硅烷基丙基)-4-羟基丁酰胺。

[0185] 多核苷酸合成

[0186] 用于多核苷酸合成的本公开的方法可包括涉及亚磷酰胺化学法的过程。在一些情况下,多核苷酸合成包括将碱基与亚磷酰胺偶联。多核苷酸合成可包括通过在偶联条件下沉积亚磷酰胺来偶联碱基,其中相同的碱基任选地与亚磷酰胺沉积超过一次,即双偶联。多核苷酸合成可包括未反应位点的加帽。在一些情况下,加帽是可选的。多核苷酸合成还可包括氧化或氧化步骤或多个氧化步骤。多核苷酸合成可包括解封闭、脱三苯甲基化和硫化。在一些情况下,多核苷酸合成包括氧化或硫化。在一些情况下,在多核苷酸合成反应期间的一个步骤或每个步骤之间,例如使用四唑或乙腈来洗涤所述装置。亚磷酰胺合成方法中任一步骤的时间范围可小于约2min、1min、50sec、40sec、30sec、20sec和10sec。

[0187] 使用亚磷酰胺方法的多核苷酸合成可包括随后将亚磷酰胺构件(例如,核苷亚磷酰胺)添加至增长的多核苷酸链以形成亚磷酸三酯键。亚磷酰胺多核苷酸合成沿3'至5'方向进行。亚磷酰胺多核苷酸合成允许在每个合成循环中将一个核苷酸受控添加至增长的多核苷酸链。在一些情况下,每个合成循环包括偶联步骤。亚磷酰胺偶联包括在活化的核苷亚磷酰胺与结合至基底的核苷之间(例如通过连接体)形成亚磷酸三酯键。在一些情况下,将

核苷亚磷酰胺提供给活化的装置。在一些情况下,将核苷亚磷酰胺提供给具有活化剂的装置。在一些情况下,核苷亚磷酰胺以相对于与基底结合的核苷1.5、2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20、25、30、35、40、50、60、70、80、90、100倍或更多倍的过量来提供给装置。在一些情况下,核苷亚磷酰胺的添加在无水环境中(例如,在无水乙腈中)进行。添加核苷亚磷酰胺后,任选地洗涤该装置。在一些情况下,偶联步骤重复一次或额外多次,任选地在向基底添加核苷亚磷酰胺之间进行洗涤步骤。在一些情况下,本文使用的多核苷酸合成方法包括1、2、3个或更多个连续的偶联步骤。在许多情况下,在偶联之前,与装置结合的核苷通过去除保护基团来脱保护,其中该保护基团起到防止聚合的作用。常见的保护基团为4,4'-二甲氧基三苯甲基(DMT)。

[0188] 偶联后,亚磷酰胺多核苷酸合成方法任选地包括加帽步骤。在加帽步骤中,用加帽剂处理增长的多核苷酸。加帽步骤可用来在偶联后封闭未反应的与基底结合的5'-OH基团以防止进一步链延伸,从而防止形成具有内部碱基缺失的多核苷酸。此外,用1H-四唑活化的亚磷酰胺可以在很小的程度上与鸟苷的O6位置反应。不受理论的束缚,在用I₂/水氧化后,该副产物(可能经由O6-N7迁移)可经历脱嘌呤。无嘌呤位点可终止在多核苷酸的最终脱保护过程中被切割,从而降低全长产物的产率。O6修饰可通过在用I₂/水氧化之前用加帽试剂处理而去除。在一些情况下,与没有加帽的合成相比,在多核苷酸合成过程中包括加帽步骤会降低错误率。作为实例,加帽步骤包括用乙酸酐和1-甲基咪唑的混合物处理与基底结合的多核苷酸。在加帽步骤之后,任选地洗涤所述装置。

[0189] 在一些情况下,在添加核苷亚磷酰胺之后,并且任选地在加帽和一个或多个洗涤步骤之后,对与装置结合的增长的多核苷酸进行氧化。氧化步骤包括将亚磷酸三酯氧化成四配位的磷酸三酯——天然存在的磷酸二酯核苷间连接的受保护的前体。在一些情况下,增长的多核苷酸的氧化通过任选地在弱碱(例如,吡啶、二甲基吡啶、三甲吡啶)的存在下用碘和水处理来实现。氧化可在无水条件下采用例如叔丁基过氧化氢或(1S)-(+)-(10-樟脑磺酰基)-氧杂吡丙啶(CSO)进行。在一些方法中,在氧化之后进行加帽步骤。第二个加帽步骤允许装置干燥,因为可能持续存在的来自氧化的残余水可以抑制随后的偶联。氧化后,任选地洗涤装置和增长的多核苷酸。在一些情况下,氧化步骤用硫化步骤来代替,以获得多核苷酸硫代磷酸,其中任何加帽步骤均可在硫化之后进行。许多试剂能够进行有效的硫转移,包括但不限于3-(二甲基氨基亚甲基)氨基)-3H-1,2,4-二噻唑-3-硫酮、DDTT、3H-1,2-苯并二噻戊环-3-酮1,1-二氧化物(也被称为Beaucage试剂)和N,N,N',N'-四乙基秋兰姆二硫化物(TETD)。

[0190] 为了使后续核苷掺入循环通过偶联而发生,除去与装置结合的增长的多核苷酸的受保护的5'末端,使得伯羟基与下一个核苷亚磷酰胺反应。在一些情况下,保护基团为DMT,并且用在二氯甲烷中的三氯乙酸进行解封闭。进行延长时间的脱三苯甲基化或者使用比推荐的酸溶液更强的酸溶液进行脱三苯甲基化可导致与固体支持物结合的多核苷酸的脱嘌呤增加,并因此降低了所需全长产物的产率。本文所述的本公开的方法和组合物提供了受控的解封闭条件,从而限制不希望的脱嘌呤反应。在一些情况下,与装置结合的多核苷酸在解封闭后洗涤。在一些情况下,解封闭后的有效洗涤有助于以低错误率合成多核苷酸。

[0191] 多核苷酸合成方法一般包括一系列迭代的以下步骤:将受保护的单体施加至活化官能化的表面(例如,座位)以与活化的表面、连接体或与预先脱保护的单体连接;使所施加

的单体脱保护,使其可与随后施加的受保护的单体反应;以及施加另一种受保护的单体以供连接。一个或多个中间步骤包括氧化或硫化。在一些情况下,在一个或全部步骤之前或之后有一个或多个洗涤步骤。

[0192] 基于亚磷酰胺的多核苷酸合成方法包括一系列化学步骤。在一些情况下,合成方法的一个或多个步骤涉及试剂循环,其中该方法的一个或多个步骤包括向该装置施加对该步骤有用的试剂。例如,试剂通过一系列液相沉积和真空干燥步骤进行循环。对于包含诸如孔、微孔、通道等三维特征的基底,试剂任选地经由孔和/或通道穿过该装置的一个或多个区域。

[0193] 本文所述的方法和系统涉及用于合成多核苷酸的多核苷酸合成装置。该合成可以是平行的。例如,可以平行合成至少或大约至少2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20、21、22、23、24、25、30、35、40、45、50、100、150、200、250、300、350、400、450、500、550、600、650、700、750、800、850、900、1000、10000、50000、75000、100000个或更多个多核苷酸。可以平行合成的多核苷酸的总数可以是2-100000、3-50000、4-10000、5-1000、6-900、7-850、8-800、9-750、10-700、11-650、12-600、13-550、14-500、15-450、16-400、17-350、18-300、19-250、20-200、21-150、22-100、23-50、24-45、25-40、30-35个。本领域技术人员知晓,平行合成的多核苷酸的总数可处于由这些值中的任何值所限定的任何范围内,例如25-100。平行合成的多核苷酸的总数可处于由充当范围端点的任何值所限定的任何范围内。在装置内合成的多核苷酸的总摩尔质量或每种多核苷酸的摩尔质量可以是至少或至少约10、20、30、40、50、100、250、500、750、1000、2000、3000、4000、5000、6000、7000、8000、9000、10000、25000、50000、75000、100000皮摩尔或更大。每种多核苷酸的长度或装置内多核苷酸的平均长度可以是至少或大约至少10、15、20、25、30、35、40、45、50、100、150、200、300、400、500个或更多个核苷酸。每种多核苷酸的长度或装置内多核苷酸的平均长度可以是至多或大约至多500、400、300、200、150、100、50、45、35、30、25、20、19、18、17、16、15、14、13、12、11、10个或更少的核苷酸。每种多核苷酸的长度或装置内多核苷酸的平均长度可以处于10-500、9-400、11-300、12-200、13-150、14-100、15-50、16-45、17-40、18-35、19-25之间。本领域技术人员知晓,每种多核苷酸的长度或装置内多核苷酸的平均长度可处于由这些值中的任何值所限定的任何范围内,例如100-300。每种多核苷酸的长度或装置内多核苷酸的平均长度可处于由充当范围端点的任何值所限定的任何范围内。

[0194] 本文提供的在表面上合成多核苷酸的方法允许以较快的速度合成。作为实例,每小时合成至少3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20、21、22、23、24、25、26、27、28、29、30、35、40、45、50、55、60、70、80、90、100、125、150、175、200个或更多个核苷酸。核苷酸包括腺嘌呤、鸟嘌呤、胸腺嘧啶、胞嘧啶、尿苷构件,或其类似物/修饰形式。在一些情况下,多核苷酸文库在基底上平行合成。例如,包含大约或至少约100、1,000、10,000、30,000、75,000、100,000、1,000,000、2,000,000、3,000,000、4,000,000或5,000,000个解析座位的装置能够支持合成至少相同数目的不同的多核苷酸,其中编码不同序列的多核苷酸在解析座位上合成。在一些情况下,在少于约三个月、两个月、一个月、三周、15天、14天、13天、12天、11天、10天、9天、8天、7天、6天、5天、4天、3天、2天、24小时或更短的时间内,以本文所述的低错误率在装置上合成多核苷酸文库。在一些情况下,使用本文所述的基底和方法从以低错误率合成的多核苷酸文库装配的较大核酸在少于约三个月、两个月、一个月、三周、15

天、14天、13天、12天、11天、10天、9天、8天、7天、6天、5天、4天、3天、2天、24小时或更短的时间内制备。

[0195] 在一些情况下,本文所述的方法提供了生成包含在多个密码子位点处不同的变异核酸的核酸文库。在一些情况下,核酸可具有1个位点、2个位点、3个位点、4个位点、5个位点、6个位点、7个位点、8个位点、9个位点、10个位点、11个位点、12个位点、13个位点、14个位点、15个位点、16个位点、17个位点、18个位点、19个位点、20个位点、30个位点、40个位点、50个位点或更多个变异密码子位点。

[0196] 在一些情况下,变异密码子位点的一个或多个位点可以是相邻的。在一些情况下,变异密码子位点的一个或多个位点可以是不相邻的,并且由1、2、3、4、5、6、7、8、9、10个或更多个密码子隔开。

[0197] 在一些情况下,核酸可包含变异密码子位点的多个位点,其中所有变异密码子位点彼此相邻,形成一段变异密码子位点。在一些情况下,核酸可包含变异密码子位点的多个位点,其中所述变异密码子位点彼此均不相邻。在一些情况下,核酸可包含变异密码子位点的多个位点,其中一些变异密码子位点彼此相邻,形成一段变异密码子位点,而一些变异密码子位点彼此不相邻。

[0198] 参见附图,图15示出了用于从较短多核苷酸合成核酸(例如,基因)的示例性处理工作流程。该工作流程大致分为以下阶段:(1)从头合成单链多核苷酸文库,(2)连接多核苷酸以形成更大的片段,(3)错误校正,(4)质量控制,以及(5)运输。在从头合成之前,预先选择预期的核酸序列或一组核酸序列。例如,预先选择一组基因用于生成。

[0199] 一旦选择用于生成的大核酸,则针对从头合成来设计预定的多核苷酸文库。用于生成高密度多核苷酸阵列的各种合适的方法是已知的。在该工作流程示例中,提供了装置表面层1501。在该示例中,改变表面的化学性质,以改进多核苷酸合成过程。生成低表面能区域以排斥液体,同时生成高表面能区域以吸引液体。表面本身可以是平面表面的形式或者包含形状的变化,例如增加表面积突起或微孔。在该工作流程示例中,如在通过引用整体并入本文的国际专利申请公开W0/2015/021080中所公开的,所选择的高表面能分子发挥支持DNA化学过程的双重功能。

[0200] 多核苷酸阵列的原位制备在固体支持物上进行,并利用单核苷酸延伸过程平行延伸多个寡聚物。沉积装置如材料沉积装置被设计为以逐步方式释放试剂,使得多个多核苷酸平行地一次延伸一个残基,以生成具有预定核酸序列的寡聚物1502。在一些情况下,多核苷酸在该阶段从表面上切下。切割包括例如采用氨或甲胺的气体切割。

[0201] 将生成的多核苷酸文库放置于反应室中。在该示例性工作流程中,反应室(也被称为“纳米反应器”)为硅涂覆的孔,其含有PCR试剂并下降到多核苷酸文库1503上。在多核苷酸密封1504之前或之后,添加试剂以从基底释放多核苷酸。在该示例性工作流程中,多核苷酸在纳米反应器密封1505之后释放。一旦释放,单链多核苷酸的片段即发生杂交,以跨越整个长程DNA序列。部分杂交1505是可能的,因为每个合成的多核苷酸被设计为具有与群体中的至少一个其它多核苷酸重叠的一小部分。

[0202] 杂交后,开始PCA反应。在聚合酶循环过程中,多核苷酸与互补片段退火,并且用聚合酶补平缺口。根据哪些多核苷酸彼此发现,每个循环随机增加各个片段的长度。片段之间的互补性允许形成完整的大跨度的双链DNA 1506。

[0203] 在PCA完成之后,将纳米反应器与装置分开1507,并定位成与具有PCR引物的装置相互作用1508。密封后,纳米反应器经历PCR1509并扩增较大的核酸。在PCR之后1510,打开纳米室1511,添加错误校正试剂1512,将腔室密封1513并进行错误校正反应,以从双链PCR扩增产物中去除具有较差互补性的错配碱基对和/或链1514。打开并分离纳米反应器1515。错误校正产物接下来经历另外的处理步骤,如PCR和分子条形码化,随后包装1522以供运输1523。

[0204] 在一些情况下,采取质量控制措施。在错误校正之后,质量控制步骤包括例如与具有用于扩增错误校正产物的测序引物的晶片进行相互作用1516,将晶片密封到含有错误校正扩增产物的腔室中1517,并进行另一轮扩增1518。打开纳米反应器1519,合并产物1520并进行测序1521。在得到可接受的质量控制结果之后,包装的产物1522准许运输1523。

[0205] 在一些情况下,通过诸如图15中的工作流程生成的多核苷酸使用本文公开的重叠引物进行诱变。在一些情况下,通过在固体支持物上原位制备来生成引物文库,并利用单核苷酸延伸过程平行延伸多个寡聚物。沉积装置如材料沉积装置被设计为以逐步方式释放试剂,使得多个多核苷酸平行地一次延伸一个残基,以生成具有预定核酸序列的寡聚物1502。

[0206] 计算机系统

[0207] 本文所述的任何系统均可以可操作地连接至计算机,并且可以本地或远程地通过计算机进行自动化。在各种情况下,本公开的方法和系统可进一步包括计算机系统上的软件程序及其使用。因此,对于分配/抽真空/再填充功能的同步(如编排和同步材料沉积装置运动、分配动作和真空致动)的计算机化控制处于本公开内容的范围内。计算机系统可被编程为在用户指定的碱基序列与材料沉积装置的位置之间接合,以将正确的试剂递送至基底的指定区域。

[0208] 图16中示出的计算机系统1600可被理解为能够从介质1611和/或网络端口1605读取指令的逻辑设备,其可任选地连接至具有固定介质1612的服务器1609。诸如图16示出的系统可包括CPU 1601、磁盘驱动器1603、可选的输入设备如键盘1615和/或鼠标1616以及可选的监视器1607。可通过示出的通信媒介实现与本地或远程位置处的服务器的数据通信。通信媒介可包括传输和/或接收数据的任何手段。例如,通信媒介可以是网络连接、无线连接或因特网连接。这样的连接可提供经由万维网的通信。可以预期有关本公开的数据可经过这样的网络或连接而传输,以便由图16所示的用户方1622接收和/或审阅。

[0209] 图17是示出可与本公开的示例实例结合使用的计算机系统1700的第一示例架构的框图。如图17所示,该示例计算机系统可包括用于处理指令的处理器1702。处理器的非限制性实例包括: Intel Xeon™处理器、AMD Opteron™处理器、Samsung 32-位RISC ARM 1176JZ (F) -Sv1.0™处理器、ARM Cortex-A8 Samsung S5PC100™处理器、ARM Cortex-A8 Apple A4™处理器、Marvell PXA 930™处理器或功能上等效的处理器。多个执行线程可用于并行处理。在一些情况下,也可以使用多个处理器或具有多个核的处理器,无论是在单一计算机系统中,在群集中,还是通过包含多个计算机、蜂窝电话和/或个人数据助理设备的网络跨系统分布。

[0210] 如图17所示,高速缓冲存储器1704可连接至或并入处理器1702,以提供由处理器1702新近或频繁使用的指令或数据的高速存储器。处理器1702通过处理器总线1708连接至北桥1706。北桥1706通过存储器总线1712连接至随机存取存储器(RAM) 1710,并管理处理器

1702对RAM 1710的访问。北桥1706还通过芯片集总线1716连接至南桥1714。南桥1714又连接至外围总线1718。外围总线可以是例如PCI、PCI-X、PCI Express或其它外围总线。北桥和南桥通常被称为处理器芯片集,并管理在处理器、RAM与外围总线1718上的外围组件之间的数据传送。在一些备选的架构中,北桥的功能性可以并入处理器中,而不是使用单独的北桥芯片。在一些情况下,系统1700可包括附接至外围总线1718的加速器卡1722。加速器可包括现场可编程门阵列(FPGA)或用于加速某个处理的其它硬件。例如,加速器可用于适应性数据重建或用来评价在扩展集处理中使用的代数表达式。

[0211] 软件和数据存储在外部存储器1724中,并可加载至RAM 1710和/或高速缓冲存储器1704中,以供处理器使用。系统1700包括用于管理系统资源的操作系统;操作系统的非限制性实例包括:Linux、WindowsTM、MACOSTM、BlackBerry OSTM、iOSTM和其它功能上等效的操作系统,以及在操作系统顶部运行的、用于根据本公开的示例实例管理数据存储和优化的应用软件。在该实例中,系统1700还包括与外围总线连接的网络接口卡(NIC)1720和1721,以提供与外部存储如网络附加存储(NAS)和可用于分布式并行处理的其它计算机系统的网络接口。

[0212] 图18是显示了具有多个计算机系统1802a和1802b、多个蜂窝电话和个人数据助理1802c以及网络附加存储(NAS)1804a和1804b的网络1800的示图。在示例实例中,系统1802a、1802b和1802c可管理数据存储并优化对存储在网络附加存储(NAS)1804a和1804b中的数据的数据访问。数学模型可用于该数据,并使用跨计算机系统1802a和1802b和蜂窝电话以及个人数据助理系统1802c的分布式并行处理进行评价。计算机系统1802a和1802b和蜂窝电话以及个人数据助理系统1802c也可提供对存储在网络附加存储(NAS)1804a和1804b中的数据的适应性数据重建的并行处理。图18仅示出了一个实例,而多种多样的其它计算机架构和系统可与本公开的多个实例一起使用。例如,刀片式服务器可用来提供并行处理。处理器刀片可通过背板连接,以提供并行处理。存储还可通过单独的网络接口连接至背板或作为网络附加存储(NAS)。在一些示例实例中,处理器可维持单独的存储空间,并通过网络接口、背板或其它连接器传输数据以便由其它处理器并行处理。在其它情况下,部分或全部处理器可使用共享的虚拟地址存储空间。

[0213] 图19是根据示例实例使用共享虚拟地址存储空间的多处理器计算机系统1900的框图。该系统包括可访问共享的存储器子系统1904的多个处理器1902a-f。该系统中并入存储器子系统1904中的多个可编程硬件存储算法处理器(MAP)1906a-f。MAP 1906a-f中的每一个可包括存储器1908a-f和一个或多个现场可编程门阵列(FPGA)1910a-f。MAP提供可配置的功能单元,并且可向FPGA 1910a-f提供特定算法或算法的部分,以便与各自的处理器密切协调处理。例如,在示例实例中,MAP可用来评价与数据模型相关的代数表达式以及用来进行适应性数据重建。在该示例中,每个MAP可被用于这些目的的所有处理器全局访问。在一种配置中,每个MAP可使用直接存储器访问(DMA)来访问相关联的存储器1908a-f,使其独立于且异步于各自的微处理器1902a-f而执行任务。在这一配置中,MAP可将结果直接馈送至另一MAP以用于流水处理和并行执行算法。

[0214] 以上计算机架构和系统仅为实例,并且多种多样的其它计算机、蜂窝电话和个人数据助理架构和系统可与示例实例结合使用,包括使用通用处理器、协处理器、FPGA和其它可编程逻辑设备、芯片上系统(SOC)、专用集成电路(ASIC)和其它处理和逻辑元件的任何组

合的系统。在一些情况下,全部或部分计算机系统可用软件或硬件来实现。任何种类的数据存储介质可与示例实例结合使用,包括随机存取存储器、硬盘驱动器、闪存存储器、磁带驱动器、磁盘阵列、网络附加存储(NAS)和其它的本地或分布式数据存储设备和系统。

[0215] 在示例实例中,计算机系统可使用在任何上述或其它计算机架构和系统上执行的软件模块来实现。在其它实例中,该系统的功能可部分或完全地在固件、可编程逻辑设备如图19提到的现场可编程门阵列(FPGA)、芯片上系统(SOC)、专用集成电路(ASIC)或其它处理和逻辑元件中实现。例如,集处理器(Set Processor)和优化器可通过使用硬件加速器卡如图17所示的加速器卡1722用硬件加速方式实现。

[0216] 阐述以下实施例是为了向本领域技术人员更清楚地说明本文所公开的实施方案的原理和实践,而不应解释为限制任何请求保护的实施方案的范围。除非另有说明,否则所有份数和百分比均以重量计。

[0217] 实施例

[0218] 给出以下实施例是为了说明本公开的各个实施方案的目的,而不意味着以任何方式限制本公开内容。这些实施例以及目前代表优选实施方案的本文所述方法是示例性的,而非旨在限制本公开的范围。本领域技术人员将会想到其变化以及包含在由权利要求的范围所限定的本公开的精神之内的其它用途。

[0219] 实施例1:装置表面的官能化

[0220] 将装置进行官能化以支持多核苷酸文库的附着和合成。首先使用包含90% H_2SO_4 和10% H_2O_2 的水虎鱼溶液(piranha solution)将装置表面润湿清洗20分钟。将该装置在含有去离子水的数个烧杯中冲洗,在去离子水鹅颈旋塞下保持5min,并用 N_2 干燥。随后将该装置在 NH_4OH (1:100;3mL:300mL)中浸泡5min,使用手持式喷枪(handgun)用去离子水冲洗,在连续三个含有去离子水的烧杯中各浸泡1min,然后再使用手持式喷枪用去离子水冲洗。然后通过将装置表面暴露于 O_2 来等离子体清洗该装置。使用SAMCO PC-300仪器在下游模式下以250瓦进行 O_2 等离子体蚀刻1min。

[0221] 使用具有以下参数的YES-1224P气相沉积烘箱系统,用包含N-(3-三乙氧基甲硅烷基丙基)-4-羟基丁酰胺的溶液对清洁的装置表面进行活化官能化:0.5至1托,60min,70℃,135℃汽化器。使用Brewer Science 200X旋涂仪对装置表面进行抗蚀剂涂覆。将SPR™ 3612光致抗蚀剂以2500rpm旋涂在装置上40sec。该装置在Brewer热板上以90℃预烘30min。使用Karl Suss MA6掩模对准仪对装置进行光刻。将该装置暴露2.2sec并在MSF 26A中显影1min。剩余的显影剂用手持式喷枪冲洗,并将装置在水中浸泡5min。该装置在烘箱中以100℃烘烤30min,随后使用Nikon L200目视检查光刻缺陷。采用清洁工艺利用SAMCO PC-300仪器以250瓦进行 O_2 等离子体蚀刻1min来去除残余抗蚀剂。

[0222] 用与10μL轻质矿物油混合的100μL全氟辛基三氯硅烷溶液对装置表面进行钝化官能化。将该装置放置于腔室中,泵送10min,随后关闭通往泵的阀门并静置10min。使该腔室排气。该装置通过在70℃下在500mL NMP中进行两次5min浸泡并同时以最大功率(在Crest系统上的9)进行超声波处理来剥离抗蚀剂。然后将该装置在室温下在500mL异丙醇中浸泡5min,同时以最大功率进行超声波处理。将该装置浸入300mL的200标准酒精度(proof)的乙醇中并用 N_2 吹干。活化该官能化表面以充当多核苷酸合成的支持物。

[0223] 实施例2:50-聚体序列的合成

[0224] 将二维寡核苷酸合成装置组装至流动池中,其与流动池(Applied Biosystems (ABI394 DNA合成仪”)连接。该二维寡核苷酸合成装置用N-(3-三乙氧基甲硅烷基丙基)-4-羟基丁酰胺(Geldest)均匀地官能化,并用来使用本文所述的多核苷酸合成方法合成50bp的示例性多核苷酸(“50-聚体多核苷酸”)。

[0225] 所述50-聚体的序列如SEQ ID NO.:20所述。5’AGACAATCAACCATTTGGGGTGGACAGCC TTGACCTCTAGACTT CGGCAT##TTTTTTTTTTT3’(SEQ ID NO.:20),其中#表示胸苷-琥珀酰基己酰胺CED亚磷酰胺(来自ChemGenes的CLP-2244),它是允许在脱保护过程中从表面上释放多核苷酸的可切割的连接体。

[0226] 根据表4中的方案和ABI合成仪,使用标准DNA合成化学法(偶联、加帽、氧化和解封闭)完成合成。

[0227] 表4:合成方案

[0228]

表 4

[0229]

通用 DNA 合成 工艺名称	工艺步骤	时间(sec)
洗涤(乙腈洗液流)	乙腈系统冲洗	4
	乙腈至流动池	23
	N2 系统冲洗	4
	乙腈系统冲洗	4
DNA 碱基添加(亚磷酰胺+活化剂流)	活化剂歧管冲洗	2
	活化剂至流动池	6
	活化剂+亚磷酰胺至流动池	6
	活化剂至流动池	0.5
	活化剂+亚磷酰胺至流动池	5
	活化剂至流动池	0.5
	活化剂+亚磷酰胺至流动池	5
	活化剂至流动池	0.5
	活化剂+亚磷酰胺至流动池	5
	孵育 25sec	25
洗涤(乙腈洗液流)	乙腈系统冲洗	4
	乙腈至流动池	15
	N2 系统冲洗	4
	乙腈系统冲洗	4
DNA 碱基添加(亚磷酰胺+活化剂流)	活化剂歧管冲洗	2
	活化剂至流动池	5
	活化剂+亚磷酰胺至流动池	18
	孵育 25sec	25
洗涤(乙腈洗液流)	乙腈系统冲洗	4
	乙腈至流动池	15
	N2 系统冲洗	4
	乙腈系统冲洗	4

[0230]

表 4		
通用 DNA 合成 工艺名称	工艺步骤	时间(sec)
加帽(帽 A+B, 1:1, 流)	帽 A+B 至流动池	15
洗涤(乙腈洗液流)	乙腈系统冲洗	4
	乙腈至流动池	15
	乙腈系统冲洗	4
氧化(氧化剂流)	氧化剂至流动池	18
洗涤(乙腈洗液流)	乙腈系统冲洗	4
	N2 系统冲洗	4
	乙腈系统冲洗	4
	乙腈至流动池	15
	乙腈系统冲洗	4
	乙腈至流动池	15
	N2 系统冲洗	4
	乙腈系统冲洗	4
	乙腈至流动池	23
	N2 系统冲洗	4
	乙腈系统冲洗	4
解封闭(解封闭流)	解封闭剂至流动池	36
洗涤(乙腈洗液流)	乙腈系统冲洗	4
	N2 系统冲洗	4
	乙腈系统冲洗	4
	乙腈至流动池	18
	N2 系统冲洗	4.13
	乙腈系统冲洗	4.13
	乙腈至流动池	15

[0231] 亚磷酰胺/活化剂组合以类似于本体试剂通过流动池递送的方式进行递送。当在全部时间内保持环境被试剂“润湿”时,不进行干燥步骤。

[0232] 从ABI 394合成仪中去除限流器,以使得能够更快速流动。在没有限流器的情况下,酰胺类(amidites)(在ACN中0.1M)、活化剂(在ACN中的0.25M苯甲酰基硫基四唑(“BTT”;来自GlenResearch的30-3070-xx)) and Ox(在20%吡啶、10%水和70% THF中的0.02M I2)的流速大致为约100uL/sec,乙腈(“ACN”)和加帽试剂(帽A和帽B的1:1混合物,其中帽A是在THF/吡啶中的乙酸酐,帽B是在THF中的16%1-甲基咪唑(1-methylimidazole))的流速大致为约200uL/sec,而解封闭剂(在甲苯中的3%二氯乙酸)的流速大致为约300uL/sec(相比之

下,在有限流器的情况下,所有试剂的流速均为约50uL/sec)。观测完全排出氧化剂的时间,相应地调节化学品流动时间的时间选择,并在不同的化学品之间引入额外的ACN洗涤。在多核苷酸合成后,将芯片在75psi下在气态氨中脱保护过夜。将五滴水施加到表面上以回收多核苷酸。然后在BioAnalyzer小RNA芯片上分析所回收的多核苷酸(数据未示出)。

[0233] 实施例3:100-聚体序列的合成

[0234] 使用实施例2中描述的用于合成50-聚体序列的相同过程,在两个不同的硅芯片上合成100-聚体多核苷酸(“100-聚体多核苷酸”;5'CGGGATCCTTATCGTCATCGTCGTACAGATCCCGACCCATTTGCTG TCCACCAGTCATGCTAGCCATACCATGATGATGATGATGATGAGA ACCCCGCAT##TTTTTTTTTTT3',其中#表示胸苷-琥珀酰基己酰胺CED亚磷酰胺(来自ChemGenes的CLP-2244);SEQ ID NO.:21),第一个用N-(3-三乙氧基甲硅烷基丙基)-4-羟基丁酰胺均匀地官能化,而第二个用11-乙酰氧基十一烷基三乙氧基硅烷和正癸基三乙氧基硅烷的5/95混合物官能化,并在BioAnalyzer仪器上分析从表面提取的多核苷酸(数据未示出)。

[0235] 使用下列热循环程序,在50uL PCR混合物(25uL NEB Q5主混合物,2.5uL 10uM正向引物,2.5uL 10uM反向引物,1uL从表面提取的多核苷酸,用水加至50uL)中使用正向引物(5'ATGCGGGGTTCTCATCATC3';SEQ ID NO.:22)和反向引物(5'CGGGATCCTTATCGTCATCG3';SEQ ID NO.:23)进一步PCR扩增来自两个芯片的全部十个样品:

[0236] 98°C,30sec

[0237] 98°C,10sec;63°C,10sec;72°C,10sec;重复12个循环

[0238] 72°C,2min

[0239] PCR产物还在BioAnalyzer上运行(数据未示出),在100-聚体位置处显示出尖锐峰。然后,对PCR扩增的样品进行克隆,并进行Sanger测序。表5总结了从来自芯片1的斑点1-5采集的样品和从来自芯片2的斑点6-10采集的样品的Sanger测序结果。

[0240] 表5:测序结果

[0241]

斑点	错误率	循环效率
1	1/763bp	99.87%
2	1/824bp	99.88%
3	1/780bp	99.87%
4	1/429bp	99.77%
5	1/1525bp	99.93%
6	1/1615bp	99.94%
7	1/531bp	99.81%
8	1/1769bp	99.94%
9	1/854bp	99.88%
10	1/1451bp	99.93%

[0242] 因此,合成的多核苷酸的高质量和均匀性在具有不同表面化学的两个芯片上重现。总体上,89%,相当于被测序的262个100-聚体中的233个,是没有错误的完美序列。最后,表6总结了从来自斑点1-10的多核苷酸样品中获得的序列的错误特征。

[0243] 表6:错误特征

[0244]

样品编号/ 斑点号	OSA_0 046/1	OSA_0 047/2	OSA_0 048/3	OSA_0 049/4	OSA_0 050/5	OSA_0 051/6	OSA_0 052/7	OSA_0 053/8	OSA_0 054/9	OSA_00 55/10
全序列	32	32	32	32	32	32	32	32	32	32
测序质量	25/28	27/27	26/30	21/23	25/26	29/30	27/31	29/31	28/29	25/28
寡核苷酸 质量	23/25	25/27	22/26	18/21	24/25	25/29	22/27	28/29	26/28	20/25
ROI 匹配 计数	2500	2698	2561	2122	2499	2666	2625	2899	2798	2348
ROI 突变	2	2	1	3	1	0	2	1	2	1
ROI 多碱 基缺失	0	0	0	0	0	0	0	0	0	0
ROI 小插 入	1	0	0	0	0	0	0	0	0	0
ROI 单碱 基缺失	0	0	0	0	0	0	0	0	0	0
大缺失计 数	0	0	1	0	0	1	1	0	0	0
突变： G>A	2	2	1	2	1	0	2	1	2	1
突变： T>C	0	0	0	1	0	0	0	0	0	0
ROI 错误 计数	3	2	2	3	1	1	3	1	2	1
ROI 错误 率	错误： 约 1/834	错误： 约 1/1350	错误： 约 1/1282	错误： 约 1/708	错误： 约 1/2500	错误： 约 1/2667	错误： 约 1/876	错误： 约 1/2900	错误： 约 1/1400	错误：约 1/2349
ROI-引物 错误率	MP 错 误：约 1/763	MP 错 误：约 1/824	MP 错 误：约 1/780	MP 错 误：约 1/429	MP 错 误：约 1/1525	MP 错 误：约 1/1615	MP 错 误：约 1/531	MP 错 误：约 1/1769	MP 错 误：约 1/854	MP 错 误：约 1/1451

[0245] 实施例4:通过单位点、单位置诱变生成核酸文库

[0246] 从头合成多核苷酸引物,以用于用来生成模板核酸的核酸变体文库的一系列PCR反应,参见图4A-图4D。图4A中生成了四种类型的引物:外部5' 引物415、外部3' 引物430、内部5' 引物425和内部3' 引物420。内部5' 引物/第一多核苷酸420和内部3' 引物/第二多核苷酸425使用如表4中大体上概括的多核苷酸合成方法生成。内部5' 引物/第一多核苷酸420代表一组至多19个具有预定序列的引物,其中该组中的每个引物在序列的单个位点上与另一个引物在单个密码子处不同。

[0247] 在具有至少两个簇的装置上进行多核苷酸合成,每个簇具有121个可单独寻址的座位。

[0248] 内部5'引物425和内部3'引物420在单独的簇中合成。内部5'引物425复制121次,在单簇内的121个座位上延伸。对于内部3'引物420,变异序列的19个引物中的每一个在6个不同的座位上各自延伸,导致在114个不同座位上延伸114个多核苷酸。

[0249] 将合成的多核苷酸从装置表面上切下并转移到塑料小瓶中。如图4B所示,使用长核酸序列435、440的片段进行第一PCR反应以扩增模板核酸。如图4C-图4D所示,使用引物组合和第一PCR反应的产物作为模板进行第二PCR反应。第二PCR产物的分析在BioAnalyzer上进行,如图20的迹线所示。

[0250] 实施例5:包含96个不同组的单位位置变体的核酸文库的生成

[0251] 大体上如图4A所示和实施例2中所提到的,使用从头多核苷酸合成来生成四组引物。对于内部5'引物420,生成96个不同组的引物,每组引物靶向位于模板核酸的单个位点内的不同单个密码子。对于每组引物,生成19个不同的变体,每个变体在所述单个位点处包含编码不同氨基酸的密码子。大体上如图4A-图4D所示和实施例2中所述,使用所生成的引物进行两轮PCR。96组扩增产物在电泳图(图21)中可视化,其用来计算100%扩增成功率。

[0252] 实施例6:包含500个不同组的单位位置变体的核酸文库的生成

[0253] 大体上如图4A所示和实施例2中所提到的,使用从头多核苷酸合成来生成四组引物。对于内部5'引物420,生成500个不同组的引物,每组引物靶向位于模板核酸的单个位点内的不同单个密码子。对于每组引物,生成19个不同的变体,每个变体在所述单个位点处包含编码不同氨基酸的密码子。大体上如图4A所示和实施例2中所述,使用所生成的引物进行两轮PCR。电泳图显示了500组PCR产物中的每一组具有在不同单个位点处具有19个变体的核酸群体(数据未示出)。对该文库的全面测序分析显示出在预选密码子突变中大于99%的成功率(序列追踪和分析数据未示出)。

[0254] 实施例7:针对1个位置的单位点诱变引物

[0255] 表7中提供了针对黄色荧光蛋白的密码子变异设计的实例。在这种情况下,来自50-聚体序列的单个密码子改变19次。变异核酸序列用粗体字母表示。野生型引物序列为:ATGGTGAGCAAGGGCGAGGAGCTGTTACCGGGGTGGTGGCCAT(SEQ ID NO.:1)。在这种情况下,野生型密码子编码缬氨酸,在SEQ ID NO.:1中用下划线表示。因此,以下19个变体不包括编码缬氨酸的密码子。在备选实例中,如果要考虑所有三联体,那么将生成全部60个变体,包括野生型密码子的备选序列。

[0256] 表7. 变异序列

[0257]	SEQ ID NO.	变异序列	变异密码子
	2	atg TT TAGCAAGGGCGAGGAGCTGTTACACGGGGTGGTGCCCAT	F
	3	atg TTA AGCAAGGGCGAGGAGCTGTTACACGGGGTGGTGCCCAT	L
	4	atg ATT AGCAAGGGCGAGGAGCTGTTACACGGGGTGGTGCCCAT	I
	5	atg TCT AGCAAGGGCGAGGAGCTGTTACACGGGGTGGTGCCCAT	S
	6	atg CCT AGCAAGGGCGAGGAGCTGTTACACGGGGTGGTGCCCAT	P
	7	atg ACT AGCAAGGGCGAGGAGCTGTTACACGGGGTGGTGCCCAT	T
	8	atg GCT AGCAAGGGCGAGGAGCTGTTACACGGGGTGGTGCCCAT	A
	9	atg TAT AGCAAGGGCGAGGAGCTGTTACACGGGGTGGTGCCCAT	Y
[0258]	10	atg CAT AGCAAGGGCGAGGAGCTGTTACACGGGGTGGTGCCCAT	H
	11	atg CAA AGCAAGGGCGAGGAGCTGTTACACGGGGTGGTGCCCAT	Q
	12	atg AAT AGCAAGGGCGAGGAGCTGTTACACGGGGTGGTGCCCAT	N
	13	atg AAA AGCAAGGGCGAGGAGCTGTTACACGGGGTGGTGCCCAT	K
	14	atg GAT AGCAAGGGCGAGGAGCTGTTACACGGGGTGGTGCCCAT	D
	15	atg GAA AGCAAGGGCGAGGAGCTGTTACACGGGGTGGTGCCCAT	E
	16	atg TGT AGCAAGGGCGAGGAGCTGTTACACGGGGTGGTGCCCAT	C
	17	atg TGG AGCAAGGGCGAGGAGCTGTTACACGGGGTGGTGCCCAT	W
	18	atg CGT AGCAAGGGCGAGGAGCTGTTACACGGGGTGGTGCCCAT	R
	19	atg GGT AGCAAGGGCGAGGAGCTGTTACACGGGGTGGTGCCCAT	G

[0259] 实施例8:单位点、双位置核酸变体

[0260] 在与实施例2中所述的类似条件下进行从头多核苷酸合成。生成装置上的单簇,其在单个位点处含有针对2个连续密码子位置的核酸的合成预定变体,每个位置存在编码氨基酸的密码子。在这种布置中,对于每个核酸有3次重复的2个位置,生成19个变体/每个位置,导致合成114个核酸。

[0261] 实施例9:多位点、双位置核酸变体

[0262] 在与实施例2中所述的类似条件下进行从头多核苷酸合成。生成装置上的单簇,其含有针对2个非连续密码子位置的核酸的合成预定变体,每个位置存在编码氨基酸的密码子。在这种布置中,对于2个位置生成19个变体/每个位置。

[0263] 实施例10:单段、三位置核酸变体

[0264] 在与实施例2中所述的类似条件下进行从头多核苷酸合成。生成装置上的单簇,其含有针对3个连续密码子位置的参考核酸的合成预定变体。在3个连续密码子位置的布置中,对于每个核酸有2次重复的3个位置,生成19个变体/每个位置,并导致合成114个核酸。

[0265] 实施例11:多位点、三位置核酸变体

[0266] 在与实施例2中所述的类似条件下进行从头多核苷酸合成。生成装置上的

单簇,其含有针对至少3个非连续密码子位置的参考核酸的合成预定变体。在预定的区域内,编码3个组氨酸残基的密码子的位置发生改变。

[0267] 实施例12:多位点、多位置核酸变体

[0268] 在与实施例2中所述的条件类似的条件下进行从头多核苷酸合成。生成装置上的单簇,其含有针对1个或多个区段中的1个或多个密码子位置的参考核酸的合成预定变体。该文库中的五个位置发生改变。第一个位置编码在表达的蛋白质中得到50/50的K/R比的密码子;第二个位置编码在表达的蛋白质中得到50/25/25的V/L/S比的密码子,第三个位置编码在表达的蛋白质中得到50/25/25的Y/R/D比的密码子,第四个位置编码在表达的蛋白质中对于所有氨基酸得到相等比例的密码子,而第五个位置编码在表达的蛋白质中得到75/25的G/P比的密码子。

[0269] 实施例13:通过采样生成核酸文库

[0270] 为了生成具有预选分布的核酸群体,使用了计算技术。以下表8提供了示例性的预选分布,其中数字代表每个位置上每个氨基酸的期望百分比。如表9所示,首先计算了累积分布值,得到0.0至1.0的值。在诸如Excel的程序中,使用均匀随机数生成器为用作采样群体的500个核酸的10个氨基酸位置的每个位置创建介于0和1之间的值。例如,对于位置1,均匀随机值“0.95”将落入“S”桶中,因此表示氨基酸“S”。该技术被称为“轮盘赌”选择。从每个设计的寡核苷酸的10个离散分布中生成10个随机数;重复该过程500次,以生成500个核酸的样本群体。为了验证生成的样本群体,然后确定该群体中每个氨基酸在该位置出现的频率的总和,并以百分比表示。例如,计算出氨基酸C在500个核酸的样本中在位置1处出现的百分比。这些值代表群体中的近似分布。通过在群体中使用足够数量的核酸,样本分布接近于预选的分布。

[0271] 表8.氨基酸的预选分布

氨基酸	位置									
	1	2	3	4	5	6	7	8	9	10
% C	0.1	0.2	0.1	0.1	0.0	0.0	0.1	6.0	2.0	0.1
% A	13.7	2.4	4.0	8.9	4.8	7.1	5.1	6.0	5.1	13.1
% R	13.7	16.7	6.7	13.3	6.0	8.3	5.1	6.0	3.8	6.6
% N	1.1	2.4	2.7	4.4	2.4	3.6	6.3	6.0	3.8	4.9
% D	15.8	16.7	5.3	4.4	1.2	13.1	21.5	6.0	11.4	8.2
% Q	7.4	6.0	4.0	2.2	1.2	1.2	2.5	6.0	3.8	1.6
% E	4.2	3.6	1.3	6.7	2.4	13.1	5.1	6.0	11.4	3.3
% G	14.7	15.5	20.0	13.3	27.4	9.5	11.4	6.0	11.4	3.3
% H	1.1	1.2	1.3	2.2	1.2	2.4	2.5	6.0	3.8	3.3
% I	1.1	1.2	10.7	3.3	1.2	6.0	5.1	6.0	3.8	3.3
% L	7.4	2.4	2.7	4.4	3.6	4.8	6.3	6.0	3.8	3.3
% K	2.1	4.8	1.3	2.2	6.0	1.2	1.3	6.0	2.5	9.8
% M	2.1	1.2	1.3	2.2	2.4	1.2	1.3	6.0	10.1	2.0
% F	1.1	1.2	4.0	6.7	3.6	2.4	1.3	6.0	2.5	3.3
% P	7.4	6.0	8.0	12.2	3.6	4.8	2.5	6.0	10.1	16.4
% S	5.3	9.5	10.7	10.0	19.0	19.0	19.0	6.0	8.9	11.5
% T	2.1	9.5	16.0	3.3	14.3	2.4	3.8	6.0	3.8	6.6

[0273] 表9.累积归一化分布

氨基酸	位置									
	1	2	3	4	5	6	7	8	9	10
C	.00	.00	.00	.00	.00	.00	.00	.06	.02	.00
A	.14	.03	.04	.09	.05	.07	.05	.12	.07	.13
R	.27	.19	.11	.22	.11	.15	.10	.18	.11	.20
N	.28	.22	.13	.27	.13	.19	.17	.24	.14	.25
D	.44	.38	.19	.31	.14	.32	.38	.29	.26	.33
Q	.52	.44	.23	.33	.16	.33	.41	.35	.29	.34
E	.56	.48	.24	.40	.18	.46	.46	.41	.40	.38
G	.70	.63	.44	.53	.45	.56	.57	.47	.52	.41
H	.72	.64	.45	.56	.46	.58	.59	.53	.55	.44
I	.73	.66	.56	.59	.48	.64	.65	.59	.59	.47
L	.80	.68	.59	.63	.51	.69	.71	.65	.63	.51
K	.82	.73	.60	.66	.57	.70	.72	.71	.65	.60
M	.84	.74	.61	.68	.60	.71	.73	.76	.75	.62
F	.85	.75	.65	.74	.63	.74	.75	.82	.78	.66
P	.93	.81	.73	.87	.67	.79	.77	.88	.88	.82
S	.98	.91	.84	.97	.86	.98	.96	.94	.96	.93
T	1	1	1	1	1	1	1	1	1	1

[0276] 实施例14. 通过过滤采样生成核酸文库

[0277] 使用实施例13中描述的方法, 对群体进行重新采样以去除不期望的组合, 并将其从群体中滤除。例如, 在任何位置具有4个“H”(组氨酸)氨基酸的组合被认为不适合生物学目的。因此, 在这种情况下, 当生成第500个寡核苷酸作为“HHHCCHHCHH(SEQ ID NO:55)”时, 由于具有8个H, 因此该组合是不期望的。结果, 按照实施例13中描述的方法, 在其位置生成了另一种随机生成的组合。使用许多标准来生成预选的分布。例如, 生成群体, 以在任何位置处在每个寡核苷酸中包括至少一个“A”(丙氨酸)氨基酸。还生成了群体, 使得生成的组合都不具有彼此相邻的两个“M”(甲硫氨酸)氨基酸。因此, 进行随机采样直到满足预选的分布和特定标准。

[0278] 实施例15: 具有均匀分布的组合文库

[0279] 在与实施例2中所述的类似条件下进行从头多核苷酸合成。如实施例4-6和8-12所述生成核酸群体, 其编码在单位点或多位点处的密码子变异, 其中在每个位置处预先选择变体, 并且它们具有预选的分布。

[0280] 为了通过组合方法生成均匀的变体分布文库, 将变体文库的参考序列拆分成两部分。如本文所用的, 均匀的变体分布是指每种变体意欲以近似相等的量合成。拆分的一侧被称为5'侧, 拆分的第二侧被称为3'侧。为参考序列的每一侧设计并合成序列, 使得在退火时, 合成所需的核酸文库。对于具有与表10类似的变异的均匀文库, 5'侧的多样性为2548 (14x 14x 13)。在3'侧, 多样性为546 (3x 13x 14)。通过退火合成5'侧和3'侧, 导致总多样性为1,391,208 (2548x 546)。通过下一代测序分析这些变体(数据未示出)。

[0281] 表10. 均匀文库的变异

[0282]

	5'变体			3'变体		
N	N	N	N		N	N
P						
Q						
R	R	R	R		R	R
S	S	S	S		S	S
T						
V	V	V	V		V	V
W	W	W	W	W	W	W
Y	Y	Y	Y		Y	Y
A	A	A	A	A	A	A
C						
D						
E						
F	F	F	F		F	F
G	G	G	G		G	G
H	H	H	H		H	H
I	I	I				I
K	K	K	K	K	K	K
L	L	L	L		L	L
M	M	M	M		M	M
多样性	2548			546		
总多样性	1391208					

[0283] 实施例16:具有非均匀分布的组合文库

[0284] 在与实施例2中所述的类似条件下进行从头多核苷酸合成。如实施例4-6和8-12所述生成核酸群体,其编码在单位点或多位点处的密码子变异,其中在每个位置处预先选择变体,并且它们具有预选的分布。

[0285] 还生成了具有非均匀变体分布的文库,该文库具有类似于表11中所示的预选分布。再次将参考序列拆分成两半,并为每个部分生成变体。拆分的一侧被称为5'侧,拆分的第二侧被称为3'侧。通过将该变体的理论置换频率相乘计算出5'变体和3'变体的预期概率。例如,对于序列NRS的5'变体,预期概率为0.0677% (9.9% x 7.6% x 9.0%)。对于5'变体和3'变体,某些变体具有相同的概率,并被分组在一起,即,分在相同的概率“箱元”中。因此,同一箱元中的所有变体具有相同的理论发生频率。对于总共1,391,208个理论变体,存在162个不同的概率,因此有162个不同的概率箱元。

[0286] 表11.变异分布

	5'变体			3'变体			
[0287]	N	9.9%	7.6%	8.7%	0.0%	8.7%	9.9%
	P	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	Q	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	R	9.9%	7.6%	8.7%	0.0%	8.7%	9.9%
	S	3.0%	9.0%	9.0%	0.0%	9.0%	3.0%
	T	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	V	9.9%	7.6%	8.7%	0.0%	8.7%	9.9%
	W	4.0%	4.0%	4.0%	20.0%	4.0%	4.0%
	Y	9.9%	7.6%	8.7%	0.0%	8.7%	9.9%
	A	4.0%	4.0%	4.0%	60.0%	4.0%	4.0%
	C	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	D	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	E	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	F	9.9%	7.6%	8.7%	0.0%	8.7%	9.9%
	G	9.9%	7.6%	8.7%	0.0%	8.7%	9.9%
	H	9.9%	7.6%	8.7%	0.0%	8.7%	9.9%
	I	9.9%	7.6%	0.0%	0.0%	0.0%	9.9%
	K	4.0%	4.0%	4.0%	20.0%	4.0%	4.0%
	L	3.0%	9.0%	9.0%	0.0%	9.0%	3.0%
	M	3.0%	9.0%	9.0%	0.0%	9.0%	3.0%

[0288] 然后进行下一代测序(NGS),以确定在所生成的变体中呈现出多少理论多样性。因为用 10^6 个读取(read)进行测序,所以仅观察到实际多样性的30%。因此,确定了以所需频率呈现的实际多样性的总和。

[0289] 呈现具有相同频率的变体数目的162个不同概率箱元用来分析NGS数据。对于162个不同的概率箱元,将来自NGS的读取按其预期出现概率进行分组(虚线),如图22所示。然后将观测到的频率(实线)与预期概率进行比较。对于162个箱元中的每个箱元,通过将变体总数除以该箱元中的变体数确定观测频率。对于每个箱元计算该值,并表示为平均计数,如图23所示。将这些值绘制为观测频率,并与预期概率进行比较,如图22所示。

[0290] 如图22中所示的变体观测频率(实线)与变体预期概率(虚线)的比较指示出观测到的多样性是否以期望的频率呈现。如图22所示,观测到的多样性与预期概率很好地匹配,并且呈现出理论多样性的超过99%。

[0291] 另外,观察了高频组合以及预定的低频组合。跨越39个碱基对的多样性区域的NGS读取中有89.9%具有正确的大小,并且估计126个碱基对的完整构建体中有超过70%是无插入和缺失的。参见图24,如单个峰所示,生成了全长片段的高百分比。

[0292] 实施例17:在8个位置中的每一个处包含144个单密码子变体和9072个双密码子变体的组合文库

[0293] 在与实施例2中所述的类似条件下进行从头多核苷酸合成。与实施例4-6和8-12类似地生成核酸群体。该核酸群体包含144个单密码子变体和9072个双密码子变体(多样性为9216),其中变体在8个位置处预先选择。

[0294] 然后进行下一代测序(NGS),以确定观察到的组合变体的分布。以大于 10^5 的读取覆盖率进行测序。如图25所示,通过NGS检测到所观察变体中的超过99%,它们具有均匀的分布。所观察变体中的超过90%是无插入和缺失的,并且检测到低于5%的脱靶序列。观察

到不到1%的野生型序列。

[0295] 实施例18:使用基于阵列的方法生成代表性变体文库

[0296] 使用与实施例1-3类似的基于阵列的方法从头合成变体文库。然后将使用基于阵列的方法生成的变体文库与使用基于PCR的方法生成的变体文库进行比较。

[0297] 构建变体文库后,对来自两个文库的集落进行采样并测序。数据在表12中示出。测序失败的数目(“测序失败的数目”)被确定为不可能进行测序的集落的数目。多样性百分比(多样性(%))由测序后获得的突变体数目与理论上可能的突变体预期数目之比确定。正确性百分比(“正确性(%))”由具有正确DNA序列的突变体数目与用于测序的突变体数目之比确定。从表12中可以看出,使用基于阵列的方法生成的变体文库显示出更高的“正确性”,与改善的多样性和质量相关。

[0298] 还通过采样在蛋白质水平上比较了这两个文库。使用基于阵列的方法生成的变体文库比使用基于PCR的方法生成的变体文库具有更具代表性的变体群体,其理论上预期的生成突变体数目增加。

[0299] 表12.变体文库数据

文库	具有缺失的 突变体的数 目	具有插入的 突变体的数 目	测序失 败的数 目	野生 型数 目	不同突变 体的数目	多样 性(%)	正确 性(%)
[0300] 基于 PCR 的文库	14	4	18	4	109	42.5	86.4
基于阵列 的文库	12	4	2	2	164	64.1	93.2

[0301] 实施例19:密码子分配方案

[0302] 使用密码子分配设计了多核苷酸文库。密码子分配用来确定在每个位点处设计的密码子序列。

[0303] 针对具有如表13中列出的野生型(WT)氨基酸序列和WT DNA序列的人肿瘤蛋白p53(TP53),生成了密码子变异。当生成密码子变异时,将要设计的变异密码子序列基于以上表3的密码子分配。具体而言,当从野生型氨基酸生成变异氨基酸时,从表3中列出的密码子序列中按从左至右的优先顺序选择编码该变异氨基酸的变异密码子序列。

[0304] 参见表13,在该肽的位置2处的野生型氨基酸是“F”(粗体)。为了在位置2处生成变异,设计了野生型序列的变体,其中“F”被改变为其它19种氨基酸中的任何一种。然后使用根据表3的密码子分配来确定设计哪个变异密码子序列以在该位置处生成变异氨基酸。为了生成其中“F”变为“A”的变体,根据表3首先选择的变异密码子序列是“GCT”,而不是“GCA”、“GCC”或“GCG”,它们全都编码“A”。表14列出了位置2处“F”的所有可能的变异氨基酸,以及设计哪个变异密码子序列来生成变异氨基酸。

[0305] 表13.用于变异的序列

[0306]

SEQ ID NO	氨基酸 或 DNA	序列
33	氨基酸	MFCQLAKTCPVQLWVDSTPPPGTRVRAMAIYKQSQHMTEVV RRCPHHERCSDSDGLAPPQHILIRVEGNLRVEYLDDRNTFRHSV VVPYEPPEVGSDCTTIHYNMCMNSSCMGGMNRRPILTIITLED SGNLLGRNSFEVRVCACPGRRRTEENLRKKGEPHHELPPGS TKRALPNNTSSSPQPKKKPLDGEYFTLQIRGRERFEMFRELNEA LELKDAQAGKEPGGSRAHSSHLKSKKGQSTSRHKKLMFKTEG PDS
34	DNA	TGAGGCCAGGAGATGGAGGCTGCAGTGAGCTGTGATCACAC CACTGTGCTCCAGCCTGAGTGACAGAGCAAGACCCCTATCTC AAAAAAAAAAAAAAAAAGAAAGCTCCTGAGGTGTAGAC GCCAACTCTCTCTAGCTCGCTAGTGGGTTCAGGAGGTGCTT ACGCATGTTTGTCTTTGCTGCCGTCTTCCAGTTGCTTTATC TGTTCACTTGTGCCCTGACTTTCAACTCTGTCTCCTTCCTCTT CCTACAGTACTCCCCTGCCCTCAACAAGATGTTTTGCCAACT GGCCAAGACCTGCCCTGTGCAGCTGTGGGTGATTCCACAC CCCCGCCCGGCACCCGCGTCCGCGCCATGGCCATCTACAAG CAGTCACAGCACATGACGGAGGTTGTGAGGCGCTGCCCCCA CCATGAGCGCTGCTCAGATAGCGATGGTCTGGCCCCCTCCTC AGCATCTTATCCGAGTGGAAGGAAATTTGCGTGTGGAGTAT TTGGATGACAGAAACACTTTTCGACATAGTGTGGTGGTGCC CTATGAGCCGCCTGAGGTTGGCTCTGACTGTACCACCATCC ACTACAACACTACATGTGTAACAGTTCCTGCATGGGCGGCATG AACCGGAGGCCCATCCTCACCATCATCACACTGGAAGACTC CAGTGGTAATCTACTGGGACGGAACAGCTTTGAGGTGCGTG TTTGTGCCTGTCTTGGGAGAGACCGGCGCACAGAGGAAGAG AATCTCCGCAAGAAAGGGGAGCCTCACCACGAGCTGCCCCC AGGGAGCACTAAGCGAGCACTGCCCAACAACACCAGCTCCT CTCCCCAGCCAAAGAAGAAACCACTGGATGGAGAATATTTT ACCCTTCAGATCCGTGGGCGTGAGCGCTTCGAGATGTTCCG AGAGCTGAATGAGGCCTTGGAACCTCAAGGATGCCCAGGCTG GGAAGGAGCCAGGGGGGAGCAGGGGCTCACTCCAGCCACCT

[0307]

		GAAGTCCAAAAAGGGTCAGTCTACCTCCCGCCATAAAAAAC TCATGTTCAAGACAGAAGGGCCTGACTCAGACTGACATTCT CCACTTCTTGTTCCCACTGACAGCCTCCACCCCCATCTCT CCCTCCCTGCCATTTTGGGTTTTGGGTCTTTGAACCCTTGCT TGCAATAGGTGTGCGTCAGAAGCACCCAGGACTTCCATTTG CTTTGTCCCGGGGCTCCACTGAACAAGTTGGCCTGCACTGGT GTTTTGTGTGGGGAGGAGGATGGGGAGTAGGACATACCAG CTTAGATTTTAAGGTTTTTACTGTGAGGGATGTTTGGGAGAT GTAAGAAATGTTCTTGCAGTTAAGGGTTAGTTTACAATCAG CCACATTCTAGGTAGGGGCCCACTTCACCGTACTAACCAGG GAAGCTGTCCCTCACTGTTGAATTTTCTCTAACTTCAAGGCC CATATCTGTGAAATGCTGGCATTTCACCTACCTCACAGAGT GCATTGTGAGGGTTAATGAAATAATGTACATCTGGCCTTGA AACCACCTTTTATTACATGGGGTCTAGAAGTTGACCCCTTG AGGGTGCTTGTTCCCTCTCCCTGTTGGTTCGGTGGGTGGTAG TTTCTACAGTTGGGCAGCTGGTTAGGTAGAGGGAGTTGTCA AGTCTCTGCTGGCCAGCCAAACCCTGTCTGACAACCTCTTG GTGAACCTTAGTACCTAAAAGGAAATCTCACCCCATCCAC ACCCTGGAGGATTTTATCTCTTGTATATGATGATCTGGATCC ACCAAGACTTGTTTTATGCTCAGGGTCAATTTCTTTTTCTTT TTTTTTTTTTTTTTCTTTTCTTTGAGACTGGGTCTCGCTTG TTGCCAGGCTGGAGTGGAGTGGCGTGATCTTGGCTTACTG CAGCCTTTGCCTCCCGGCTCGAGCAGTCCTGCCTCAGCCTC CGGAGTAGCTGGGACCACAGGTTTCATGCCACCATGGCCAGC CAACTTTTGCATGTTTTGTAGAGATGGGGTCTCACAGTGTTG CCCAGGCTGGTCTCAAACCTCCTGGGCTCAGGCGATCCACCT GTCTCAGCCTCCCAGAGTGCTGGGATTACAATTGTGAGCCA CCACGTCCAGCTGGAAGGGTCAACATCTTTTACATTCTGCA AGCACATCTGCATTTTACCCCAACCTTCCCTCCTTCTCCCT TTTTATATCCCATTTTATATCGATCTCTTATTTTACAATAAA ACTTTGCTGCCACCTGTGTGTCTGAGGGGTG
--	--	--

[0308] 表14. 变异氨基酸

[0309]

SEQ ID NO	野生 型氨 基酸	野生 型密 码子	变异 氨基 酸	变异 密码 子	DNA 位置	氨基 酸位 置	示例子序列(变异密码子以小写示 出)
35	F	TTT	A	GCT	282	2	CCCCTGCCCTCAACAAGATGg ctTGCCAAGTGGCCAA
36	F	TTT	C	TGC	282	2	CCCCTGCCCTCAACAAGATGtg cTGCCAAGTGGCCAA
37	F	TTT	D	GAT	282	2	CCCCTGCCCTCAACAAGATGg atTGCCAAGTGGCCAA
38	F	TTT	E	GAG	282	2	CCCCTGCCCTCAACAAGATGg agTGCCAAGTGGCCAA
39	F	TTT	F	TTC	282	2	CCCCTGCCCTCAACAAGATGtt cTGCCAAGTGGCCAA
40	F	TTT	G	GGT	282	2	CCCCTGCCCTCAACAAGATGg gtTGCCAAGTGGCCAA

[0310]

41	F	TTT	H	CAC	282	2	CCCCTGCCCTCAACAAGATGc acTGCCAACTGGCCAA
42	F	TTT	I	ATC	282	2	CCCCTGCCCTCAACAAGATGat cTGCCAACTGGCCAA
43	F	TTT	K	AAG	282	2	CCCCTGCCCTCAACAAGATGa agTGCCAACTGGCCAA
44	F	TTT	L	CTG	282	2	CCCCTGCCCTCAACAAGATGct gTGCCAACTGGCCAA
45	F	TTT	M	ATG	282	2	CCCCTGCCCTCAACAAGATGat gTGCCAACTGGCCAA
46	F	TTT	N	AAC	282	2	CCCCTGCCCTCAACAAGATGa acTGCCAACTGGCCAA
47	F	TTT	P	CCT	282	2	CCCCTGCCCTCAACAAGATGc ctTGCCAACTGGCCAA
48	F	TTT	Q	CAG	282	2	CCCCTGCCCTCAACAAGATGc agTGCCAACTGGCCAA
49	F	TTT	R	AGA	282	2	CCCCTGCCCTCAACAAGATGa gaTGCCAACTGGCCAA
50	F	TTT	S	AGC	282	2	CCCCTGCCCTCAACAAGATGa gcTGCCAACTGGCCAA
51	F	TTT	T	ACC	282	2	CCCCTGCCCTCAACAAGATGa ccTGCCAACTGGCCAA
52	F	TTT	V	GTG	282	2	CCCCTGCCCTCAACAAGATGgt gTGCCAACTGGCCAA
53	F	TTT	W	TGG	282	2	CCCCTGCCCTCAACAAGATGtg gTGCCAACTGGCCAA
54	F	TTT	Y	TAC	282	2	CCCCTGCCCTCAACAAGATGta cTGCCAACTGGCCAA

[0311] 实施例20:具有多个变异位点的CDR中的一段

[0312] 如实施例4-6和8-12所述生成核酸文库,其编码在单位点或多位点处的密码子变异,其中在每个位置处预先选择变体。该变异区编码CDR的至少一部分。参见,例如,图12。合成的核酸从装置表面上释放下来,并用作引物以生成核酸文库,该核酸文库在细胞中表达以生成变异蛋白质文库。评估变异抗体对表位的结合亲和力的增加。

[0313] 实施例21:变异抗体文库的生成

[0314] 如以上实施例所述生成核酸文库。为编码图12的代表性CDR的核酸生成变体文库。对代表性CDR进行修饰,其中CDR区包含多个用于变异的位置,如图13所见。如图13所示,选择了不同数目的密码子变体和变体的位置。在图13中,可以创建的变体文库的多样性为1, 152。下一代测序分析表明,预期变体存在于正确的部分和正确的位置。

[0315] 实施例22:用于表达多样化肽的模块化质粒组件

[0316] 如实施例4-6和8-12所述生成核酸文库,其编码在构成表达构建体盒的部分的每个单独区域的单个位点或多个位点处的密码子变异,如图14中所示。为了生成表达两个构建体的盒,合成编码第一启动子1410、第一开放阅读框1420、第一终止子1430、第二启动子1440、第二开放阅读框1450或第二终止子序列1460的变异序列的至少一部分的变异核酸。如前述实施例中所述,在数轮扩增后,生成了具有1,024个表达构建体的文库。

[0317] 实施例23:多位点、单位位置变体

[0318] 如实施例4-6和8-12所述生成核酸文库,其编码在编码核酸至少一部分的区域中的单个位点或多个位点处的密码子变异。生成核酸变体文库,其中该文库由多位点、单位位置

变体组成。参见例如图8B。

[0319] 实施例24:变体文库合成

[0320] 在与实施例2中所述的类似条件下进行从头多核苷酸合成。从头合成至少约30,000个不同的多核苷酸,其中每个不同的多核苷酸均编码氨基酸序列的不同密码子变体。所合成的至少30,000个不同多核苷酸与上述至少约30,000个不同多核苷酸中的每一个的预定序列相比具有小于1/1000个碱基的总错误率。该文库用于长核酸的PCR诱变,并且形成至少约30,000个不同的变异多核苷酸。

[0321] 实施例25:基于簇的变体文库合成

[0322] 在与实施例2中所述的类似条件下进行从头多核苷酸合成。生成装置上的单簇,其含有针对2个密码子位置的参考核酸的合成预定变体。在2个连续密码子位置的布置中,对于每个核酸有2次重复的2个位置,生成19个变体/每个位置,并导致合成38个核酸。每个变异序列的长度为40个碱基。在相同的簇中,生成另外的非变异核酸序列,其中所述另外的非变异核酸和变异核酸共同编码基因的编码序列的38个变体。每个核酸均具有至少一个与另一个核酸互补的区域。通过气态氨切割来释放该簇中的核酸。包含水的大头针(pin)与该簇接触,挑取核酸,并将核酸移动至小瓶中。该小瓶还含有用于聚合酶循环装配(PCA)反应的DNA聚合酶试剂。使核酸退火,通过延伸反应补平缺口,并形成所得到的双链DNA分子,从而形成变异核酸文库。任选地对变异核酸文库进行限制酶切割,然后将其连接到表达载体中。

[0323] 实施例26:针对蛋白质结合亲和力的变化筛查变异核酸文库

[0324] 如实施例13-16所述生成多个表达载体。在该实施例中,表达载体为HIS标记的细菌表达载体。将载体文库电穿孔到细菌细胞中,然后选择克隆用于表达并纯化HIS标记的变异蛋白质。针对与靶分子的结合亲和力的变化筛选变异蛋白质。

[0325] 通过诸如使用金属亲和色谱法(IMAC)的方法检查亲和性,其中使用金属离子涂覆的树脂(例如,IDA-琼脂糖或NTA-琼脂糖)来分离HIS-标记的蛋白质。由于组氨酸残基串在特定缓冲液条件下与几种类型的固定化金属离子(包括镍、钴和铜)结合,所以可以纯化并检测表达的His-标记的蛋白质。结合/洗涤缓冲液的一个实例由含有10-25mM咪唑的Tris-缓冲盐水(TBS) pH 7.2组成。从IMAC柱中洗脱并回收所捕获的HIS-标记的蛋白质用高浓度的咪唑(至少200mM)(洗脱剂)、低pH(例如,0.1M甘氨酸-HCl, pH 2.5)或过量的强螯合剂(例如,EDTA)来完成。

[0326] 或者,抗HIS-标记抗体可商购获得,用于涉及HIS-标记的蛋白质的测定方法中,如分离HIS-标记的蛋白质的下拉测定或检测HIS-标记的蛋白质的免疫印迹测定。

[0327] 实施例27:针对细胞粘附和迁移调节剂的活性变化筛查变异核酸文库

[0328] 将如实施例13-16所述生成的变异核酸文库插入GFP-标记的哺乳动物表达载体中。将从文库中分离的克隆瞬时转染到哺乳动物细胞中。或者,从含有表达构建体的细胞中表达并分离蛋白质,然后将该蛋白质递送至细胞用于进一步测量。进行免疫荧光测定以评估GFP标记的变异表达产物的细胞定位的变化。进行FACS测定以评估与GFP标记的变异蛋白质表达产物的非变体形式相互作用的跨膜蛋白的构象状态变化。进行伤口愈合试验以评估表达GFP标记的变异蛋白质的细胞侵入通过在细胞培养皿上刮擦形成的空间的能力的变化。使用荧光光源和照相机对表达GFP标记的蛋白质的细胞进行鉴定及追踪。

[0329] 实施例28:针对抑制病毒进展的肽筛查变异核酸文库

[0330] 将如实施例13-16所述生成的变异核酸文库插入FLAG-标记的哺乳动物表达载体中,并且该变异核酸文库编码肽序列。原代哺乳动物细胞从患有病毒病症的受试者中获得。或者,用病毒感染来自健康受试者的原代细胞。将细胞接种到一系列微孔皿上。将从变体文库中分离的克隆瞬时转染至细胞中。或者,从含有表达构建体的细胞中表达并分离蛋白质,然后将该蛋白质递送至细胞用于进一步测量。进行细胞存活试验以评估受感染的细胞与变异肽相关的存活增强。示例性病毒包括但不限于禽流感、寨卡病毒(zika virus)、汉坦病毒、丙型肝炎和天花。

[0331] 一个示例性试验是中性红细胞毒性测定,其使用中性红染料,当添加至细胞中时,由于中性红的轻度阳离子性质,其扩散穿过质膜并积聚在酸性溶酶体区室中。病毒诱导的细胞变性导致膜破碎和溶酶体ATP驱动的质子移位活性的丧失。细胞内中性红的随后减少可以采用分光光度法以多孔板形式进行评估。表达变异肽的细胞通过信号增加颜色测定中细胞内中性红的增加来评分。针对抑制病毒诱导的细胞变性的肽来评估细胞。

[0332] 实施例29:筛选提高或降低细胞代谢活性的变异蛋白质

[0333] 为了鉴定导致细胞代谢活性变化的表达产物,如实施例13-16所述生成多种表达载体。在该实施例中,将表达载体转移(例如,通过转染或转导)至接种在一系列微孔皿上的细胞中。然后针对代谢活性的一种或多种变化筛选细胞。或者,从含有表达构建体的细胞中表达并分离蛋白质,然后将该蛋白质递送至细胞用于测定代谢活性。任选地,在筛选一种或多种代谢活性变化之前,用毒素处理用于测定代谢活性的细胞。所施用的示例性毒素包括但不限于肉毒杆菌毒素(包括免疫学类型:A、B、C1、C2、D、E、F和G)、葡萄球菌肠毒素B、鼠疫耶尔森氏菌(*Yersinia pestis*)、丙型肝炎、芥子剂、重金属、氰化物、内毒素、炭疽杆菌(*Bacillus anthracis*)、寨卡病毒、禽流感、除草剂、杀虫剂、汞、有机磷酸酯和蓖麻毒素。

[0334] 基础能量需求来源于代谢底物(例如,葡萄糖)的氧化,其通过涉及有氧三羧酸(TCA)或Kreb循环的氧化磷酸化或无氧糖酵解来进行。当糖酵解是能量的主要来源时,细胞的代谢活性可通过监测细胞分泌酸性代谢产物(例如,乳酸盐和 CO_2)的速率来估计。在有氧代谢的情况下,细胞外氧的消耗和氧化自由基的产生反映了细胞的能量需求。细胞内氧化还原电势可通过NADH和 NAD^+ 的自发荧光测量来测量。由细胞释放的能量(例如,热量)的量由代谢过程中产生和/或消耗的物质分析值得出,其在正常设定下可由消耗的氧气量(例如,4.8kcal/1 O_2)预测。热产生与氧利用之间的偶联可能受到毒素的干扰。直接微量热法测量热隔离的样品的温度升高。因此,当与耗氧量测量相结合时,量热法可用来检测毒素的解偶联活性。

[0335] 用于测量代谢活性各种标志物的变化的各种方法和装置是本领域中已知的。例如,在通过引用整体并入本文的美国专利7,704,745中讨论了这类方法、装置和标志物。简言之,记录每个细胞群体的任何以下特征的测量值:葡萄糖、乳酸盐、 CO_2 、NADH与 NAD^+ 之比、热量、 O_2 消耗量和自由基产生。筛选的细胞可包括肝细胞、巨噬细胞或神经母细胞瘤细胞。筛选的细胞可以是细胞系、来自受试者的原代细胞或来自模型系统(例如,小鼠模型)的细胞。

[0336] 各种技术可用于测量单细胞或位于多孔板的腔室内的细胞群体的氧气消耗速率。例如,包含细胞的腔室可具有记录温度、电流或荧光变化的传感器,以及耦合到每个腔室以

监测荧光的光学系统,例如,光纤耦合的光学系统。在该实施例中,每个腔室均具有用于照射光源的窗口以激发腔室内的分子。纤维耦合的光学系统可检测自发荧光,以测量细胞内NADH/NAD比例和电压以及钙敏感染料,以确定跨膜电势和细胞内钙。另外,也检测CO₂和/或O₂敏感的荧光染料信号的变化。

[0337] 实施例30:针对癌细胞的选择性靶向筛查变异核酸文库

[0338] 将如实施例13-16所述生成的变异核酸文库插入FLAG-标记的哺乳动物表达载体,并且该变异核酸文库编码肽序列。将从变体文库中分离的克隆分别瞬时转染至癌细胞和非癌细胞中。对癌细胞和非癌细胞均进行细胞存活和细胞死亡试验,每种细胞表达由变异核酸编码的变异肽。评估细胞与变异肽相关的选择性癌细胞杀伤。癌细胞任选地是来自被诊断为患有癌症的受试者的癌细胞系或原代癌细胞。在来自被诊断为患有癌症的受试者的原代癌细胞的情况下,任选地选择在筛选试验中鉴定出的变异肽以供施用于受试者。或者,从含有蛋白质表达构建体的细胞中表达并分离蛋白质,然后将该蛋白质递送至癌细胞和非癌细胞以供进一步测量。

[0339] 实施例31:组合文库的生成

[0340] 在实施例2中大体描述的条件下进行从头多核苷酸合成。如实施例4-6和8-12所述生成核酸群体,其编码在单位点或多位点处的密码子变异,其中在每个位置处预先选择变体。通过将第一群体的核酸与第二群体的核酸组合来生成组合文库。如图1所示,将4种核酸的群体110与4种核酸的另一群体120组合以产生16种组合。

[0341] 通过平端连接使核酸退火。在1.5ml小瓶中将一种核酸的50ng DNA与另一种核酸的50ng DNA混合。接下来,添加1μL的T4 DNA连接酶(New England BioLabs)以及20μL连接缓冲液和20μL无核酸酶水。然后将反应混合物孵育。孵育后,通过测序分析连接产物。

[0342] 实施例32:通过采样生成组合文库

[0343] 在实施例2中大体描述的条件下进行从头多核苷酸合成。如实施例4-6和8-12所述生成核酸群体,其编码在单位点或多位点处的密码子变异,其中在每个位置处预先选择变体。

[0344] 参见图26A,通过实施例13-16中所述的类似方法,以预选的分布产生具有非均匀变体分布的文库。图像中每个图案化的部分代表在每个位置(A1、A2、A3、B1、B2和B3)处具有不同预选分布的4种不同氨基酸中的1种。黑色圆圈代表每个位置内的随机选择。参见图26B,独立生成了针对A的5个随机生成的样本和针对B的5个随机生成的样本。然后,例如通过平端连接,将A处的5个随机生成的样本和B处的5个随机生成的样本退火在一起,如图26C所示。这产生25种组合($n^2=5^2$)。参见图26D,统计学比较证明,所得到的分布与预选分布相匹配。

[0345] 实施例33:组合抗体文库的生成

[0346] 如以上实施例所述生成核酸文库。针对编码以下CDR区的核酸生成了变体文库:单CDR区,如图27A所示;两个CDR区,如图27B所示;或多个CDR区,如图27C所示。

[0347] 还生成了以下变异抗体文库,其包含单个或多个重链和轻链支架中的变体,如图28A所示,或者单个或多个框架中的变体,如图28B所示。

[0348] 虽然本文已经示出并描述了本发明的优选实施方案,但对于本领域技术人员明显的是,这些实施方案仅通过示例的方式提供。本领域技术人员在不脱离本发明的情况下将

会想到许多变化、改变和替代。应当理解,可在实施本发明时采用本文所述本发明实施方案的各种替代方案。旨在以所附权利要求限定本发明的范围,并且由此涵盖这些权利要求范围内的方法和结构及其等同物。

[0349] 本申请提供了以下内容:

[0350] 1). 一种合成变异核酸文库的方法,其包括:

[0351] a. 提供编码至少500个多核苷酸序列的预定序列,其中所述至少500个多核苷酸序列具有预选的密码子分布;

[0352] b. 合成编码所述至少500个多核苷酸序列的多个多核苷酸;

[0353] c. 测定由所述多个多核苷酸编码的核酸或基于所述多个多核苷酸翻译的蛋白质的活性;以及

[0354] d. 从步骤(c)的测定中收集结果,其中所述收集包括收集与阴性或无效结果相关的预定序列的结果。

[0355] 2). 根据1)所述的方法,其中步骤(d)包括收集至少80%的所述预定序列的结果。

[0356] 3). 根据1)所述的方法,其中步骤(d)包括收集至少90%的所述预定序列的结果。

[0357] 4). 根据1)所述的方法,其中步骤(d)包括收集至少100%的所述预定序列的结果。

[0358] 5). 根据1)所述的方法,其中呈现出预测多样性的至少70%。

[0359] 6). 根据1)所述的方法,其中呈现出预测多样性的至少90%。

[0360] 7). 根据1)所述的方法,其中呈现出预测多样性的至少95%。

[0361] 8). 根据1)所述的方法,其中所述至少500个多核苷酸序列中的至少80%具有正确的大小。

[0362] 9). 根据1)所述的方法,其中所述至少500个多核苷酸序列中的至少80%各自以所述文库中每个所述多核苷酸序列的平均频率的2倍以内的量存在于所述变异核酸文库中。

[0363] 10). 根据1)所述的方法,其进一步包括从步骤(c)的测定中收集与增强或降低的活性相关的预定序列的结果。

[0364] 11). 根据1)所述的方法,其中所述活性是细胞活性。

[0365] 12). 根据11)所述的方法,其中所述细胞活性包括增殖、生长、粘附、死亡、迁移、能量产生、氧利用、代谢活性、细胞信号传导、对自由基损伤的响应或其任意组合。

[0366] 13). 根据1)所述的方法,其中所述变异核酸文库编码变异基因或其片段的序列。

[0367] 14). 根据1)所述的方法,其中所述变异核酸文库编码抗体、酶或肽的至少一部分。

[0368] 15). 一种生成核酸组合文库的方法,该方法包括:

[0369] a. 设计预定的序列,其编码:

[0370] (i) 第一多个多核苷酸,其中所述第一多个多核苷酸中的每个多核苷酸编码与单个参考序列相比的变异序列,和

[0371] (ii) 第二多个多核苷酸,其中所述第二多个多核苷酸中的每个多核苷酸编码与单个参考序列相比的变异序列;

[0372] b. 合成所述第一多个多核苷酸和所述第二多个多核苷酸;以及

[0373] c. 混合所述第一多个多核苷酸和所述第二多个多核苷酸以形成核酸的组合文库,其中呈现出预测多样性的至少70%。

[0374] 16). 根据15)所述的方法,其中所述组合文库是非饱和组合文库。

- [0375] 17). 根据15)所述的方法,其中所述组合文库是饱和组合文库。
- [0376] 18). 根据15)所述的方法,其中合成了至少10,000个多核苷酸。
- [0377] 19). 根据16)所述的方法,其中用于生成所述非饱和组合文库的多核苷酸的总数比用于生成饱和组合文库的多核苷酸的总数少至少25%。
- [0378] 20). 根据15)所述的方法,其中至少80%的变体具有正确的大小。
- [0379] 21). 根据15)所述的方法,其中呈现出预测多样性的至少90%。
- [0380] 22). 根据15)所述的方法,其中呈现出预测多样性的至少95%。
- [0381] 23). 根据15)所述的方法,其中所述组合文库编码第一参考序列或第二参考序列。
- [0382] 24). 根据15)所述的方法,其中所述组合文库在翻译时编码蛋白质文库。
- [0383] 25). 根据15)所述的方法,其中将所述组合文库的核酸插入载体中。
- [0384] 26). 根据15)所述的方法,其进一步包括使用所述组合文库作为PCR诱变反应的引物来进行核酸的PCR诱变。
- [0385] 27). 根据15)所述的方法,其中所述组合文库编码变异基因或其片段的序列。
- [0386] 28). 根据15)所述的方法,其中所述组合文库编码抗体、酶或肽的至少一部分。
- [0387] 29). 根据28)所述的方法,其中所述组合文库编码所述抗体的可变区或恒定区的至少一部分。
- [0388] 30). 根据28)所述的方法,其中所述组合文库编码所述抗体的至少一个CDR区。
- [0389] 31). 根据28)所述的方法,其中所述组合文库编码在所述抗体的重链上的CDR1、CDR2和CDR3以及在其轻链上的CDR1、CDR2和CDR3。
- [0390] 32). 一种合成变异核酸文库的方法,其包括:
- [0391] a. 提供编码多个多核苷酸的预定序列,其中所述多核苷酸编码与单个参考序列相比具有变异序列的多个密码子;
- [0392] b. 为预定核酸参考序列中预选位置处的密码子选择分布值;
- [0393] c. 提供机器指令以随机生成一组具有与所选分布值相匹配的分布值的核酸序列,其中该组核酸序列少于生成饱和密码子变体文库所需的核酸序列的量;以及
- [0394] d. 合成具有预选的分布的变异核酸文库,其中呈现出预测多样性的至少70%。
- [0395] 33). 根据32)所述的方法,其中至少80%的变体具有正确的大小。
- [0396] 34). 根据32)所述的方法,其中呈现出预测多样性的至少90%。
- [0397] 35). 根据32)所述的方法,其中呈现出预测多样性的至少95%。
- [0398] 36). 根据32)所述的方法,其中所述变异核酸文库在翻译时编码蛋白质文库。
- [0399] 37). 根据32)所述的方法,其中将所述变异核酸文库的核酸插入载体中。
- [0400] 38). 根据32)所述的方法,其进一步包括使用所述变异核酸文库作为PCR诱变反应的引物来进行核酸的PCR诱变。
- [0401] 39). 根据32)所述的方法,其中使用密码子分配来确定具有变异序列的所述多个密码子中的每个密码子。
- [0402] 40). 根据39)所述的方法,其中所述密码子分配基于生物体中密码子序列的频率。
- [0403] 41). 根据40)所述的方法,其中所述生物体是动物、植物、真菌、原生生物、古菌和细菌中的至少一种。
- [0404] 42). 根据39)所述的方法,其中所述密码子分配基于所述密码子序列的多样性。

- [0405] 43). 一种合成变异核酸文库的方法,其包括:
- [0406] a. 提供编码多个多核苷酸的预定序列,其中所述多核苷酸编码与单个参考序列相比具有变异序列的密码子;
- [0407] b. 将所述多个多核苷酸分成多核苷酸的5' 片段和多核苷酸的3' 片段;
- [0408] c. 为预定核酸参考序列中预选位置处的密码子选择分布值;
- [0409] d. 提供机器指令以随机生成一组具有与所选分布值相匹配的分布值的核酸,其中该组核酸少于生成饱和核酸文库所需的核酸的量;
- [0410] e. 合成多核苷酸的5' 片段和多核苷酸的3' 片段;以及
- [0411] f. 混合多核苷酸的5' 片段和多核苷酸的3' 片段以形成变异核酸文库,其中呈现出预测多样性的至少70%。
- [0412] 44). 根据43)所述的方法,其中合成了至少10,000个多核苷酸。
- [0413] 45). 根据43)所述的方法,其中至少80%的变体具有正确的大小。
- [0414] 46). 根据43)所述的方法,其中呈现出预测多样性的至少90%。
- [0415] 47). 根据43)所述的方法,其中呈现出预测多样性的至少95%。
- [0416] 48). 根据43)所述的方法,其中将所述多个多核苷酸分成多于一个5' 片段和多于一个3' 片段中的至少一种。
- [0417] 49). 根据43)所述的方法,其中所述变异核酸文库在翻译时编码蛋白质文库。
- [0418] 50). 根据43)所述的方法,其中将所述变异核酸文库的核酸插入载体中。
- [0419] 51). 根据43)所述的方法,其进一步包括使用所述变异核酸文库作为PCR诱变反应的引物来进行核酸的PCR诱变。
- [0420] 52). 根据43)所述的方法,其进一步包括鉴定具有增强或降低的活性的变异序列。
- [0421] 53). 根据52)所述的方法,其中所述活性是细胞活性。
- [0422] 54). 根据53)所述的方法,其中所述细胞活性包括增殖、生长、粘附、死亡、迁移、能量产生、氧利用、代谢活性、细胞信号传导、对自由基损伤的响应或其任意组合。
- [0423] 55). 根据43)所述的方法,其中所述变异核酸文库编码变异基因或其片段的序列。
- [0424] 56). 根据43)所述的方法,其中所述变异核酸文库编码抗体、酶或肽的至少一部分。
- [0425] 57). 根据56)所述的方法,其中所述变异核酸文库编码所述抗体的可变区或恒定区的至少一部分。
- [0426] 58). 根据56)所述的方法,其中所述变异核酸文库编码所述抗体的至少一个CDR区。
- [0427] 59). 根据56)所述的方法,其中所述变异核酸文库编码在所述抗体的重链上的CDR1、CDR2和CDR3以及在其轻链上的CDR1、CDR2和CDR3。
- [0428] 60). 根据43)所述的方法,其中在所述变异核酸文库中合成的不同序列的数目在50至1,000,000的范围内。
- [0429] 61). 根据43)所述的方法,其中在所述变异核酸文库中合成的不同序列的数目在500至25000的范围内。
- [0430] 62). 根据43)所述的方法,其中在所述变异核酸文库中合成的不同序列的数目在1000至15000的范围内。

[0431] 63). 根据43)所述的方法,其进一步包括使用所述变异核酸文库作为PCR诱变反应的引物来进行核酸的PCR诱变。

[0432] 64). 根据43)所述的方法,其中使用密码子分配来确定具有变异序列的密码子。

[0433] 65). 根据64)所述的方法,其中所述密码子分配基于生物体中密码子序列的频率。

[0434] 66). 根据65)所述的方法,其中所述生物体是动物、植物、真菌、原生生物、古菌和细菌中的至少一种。

[0435] 67). 根据64)所述的方法,其中所述密码子分配基于所述密码子序列的多样性。

[0436] 68). 一种生成核酸组合文库的方法,该方法包括:

[0437] a. 提供预定的序列,其编码:

[0438] (i) 第一多个多核苷酸,其中所述第一多个多核苷酸中的每个多核苷酸编码与单个参考序列相比的变异序列,和

[0439] (ii) 第二多个多核苷酸,其中所述第二多个多核苷酸中的每个多核苷酸编码与单个参考序列相比的变异序列;

[0440] b. 提供具有表面的结构;

[0441] c. 合成所述第一多个多核苷酸,其中所述第一多个多核苷酸中的每个多核苷酸从所述表面延伸;

[0442] d. 合成所述第二多个多核苷酸,其中所述第二多个多核苷酸中的每个多核苷酸从所述表面延伸;

[0443] e. 从所述表面释放所述第一多个多核苷酸和所述第二多个多核苷酸;以及

[0444] f. 混合所述第一多个多核苷酸和所述第二多个多核苷酸以形成核酸的组合文库,其中呈现出预测多样性的至少70%。

[0445] 69). 根据68)所述的方法,其中呈现出预测多样性的至少90%。

[0446] 70). 根据68)所述的方法,其中呈现出预测多样性的至少95%。

[0447] 71). 一种合成变异核酸文库的方法,其包括:

[0448] a. 设计编码多个多核苷酸的预定序列,其中所述多核苷酸编码与单个参考序列相比具有变异序列的多个密码子;

[0449] b. 合成所述多个多核苷酸以生成变异核酸文库,其中呈现出预测多样性的至少70%;

[0450] c. 表达所述变异核酸文库;以及

[0451] d. 评价与变异核酸文库相关的活性。

[0452] 72). 根据71)所述的方法,其中呈现出预测多样性的至少90%。

[0453] 73). 根据71)所述的方法,其中呈现出预测多样性的至少95%。

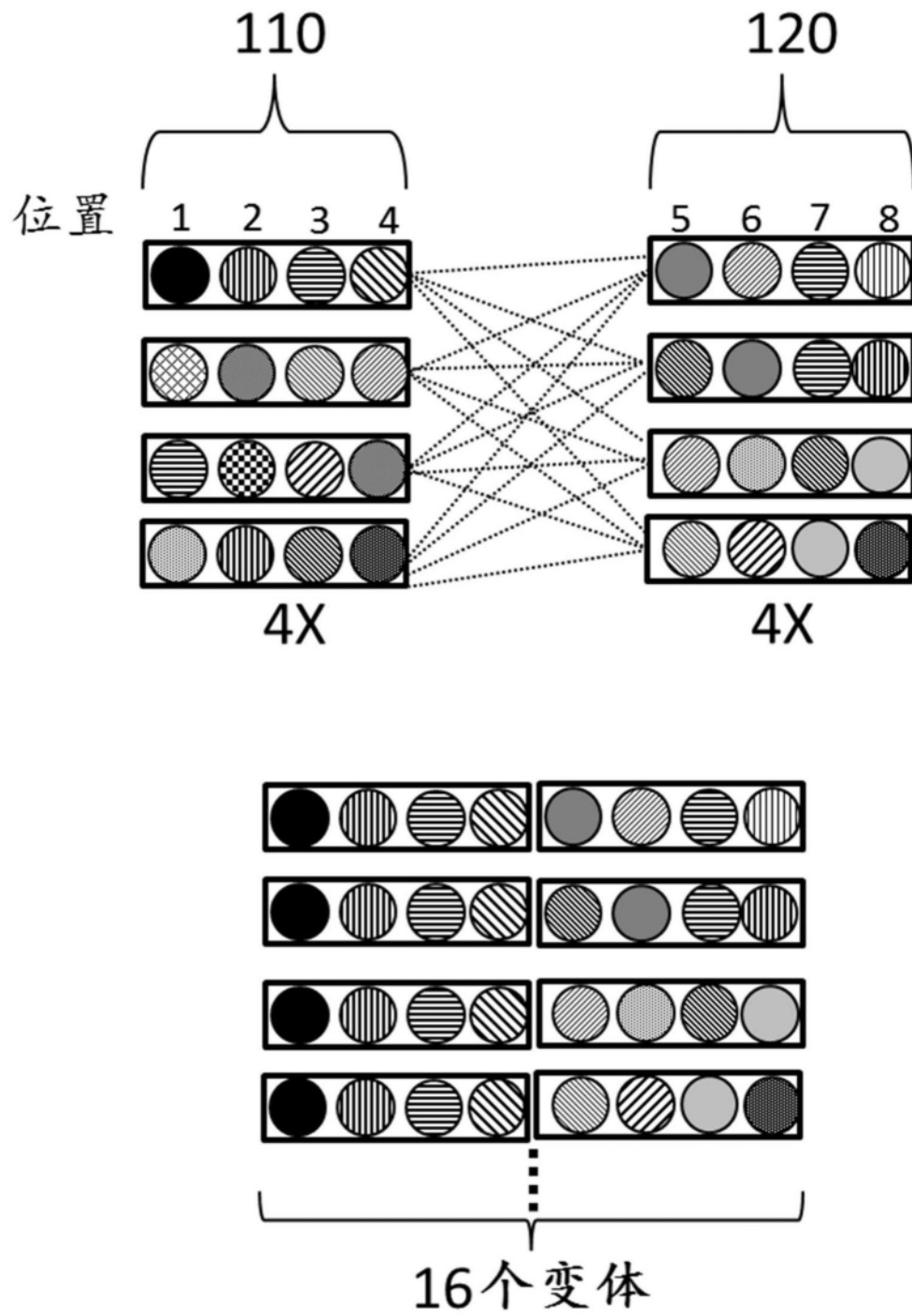


图1

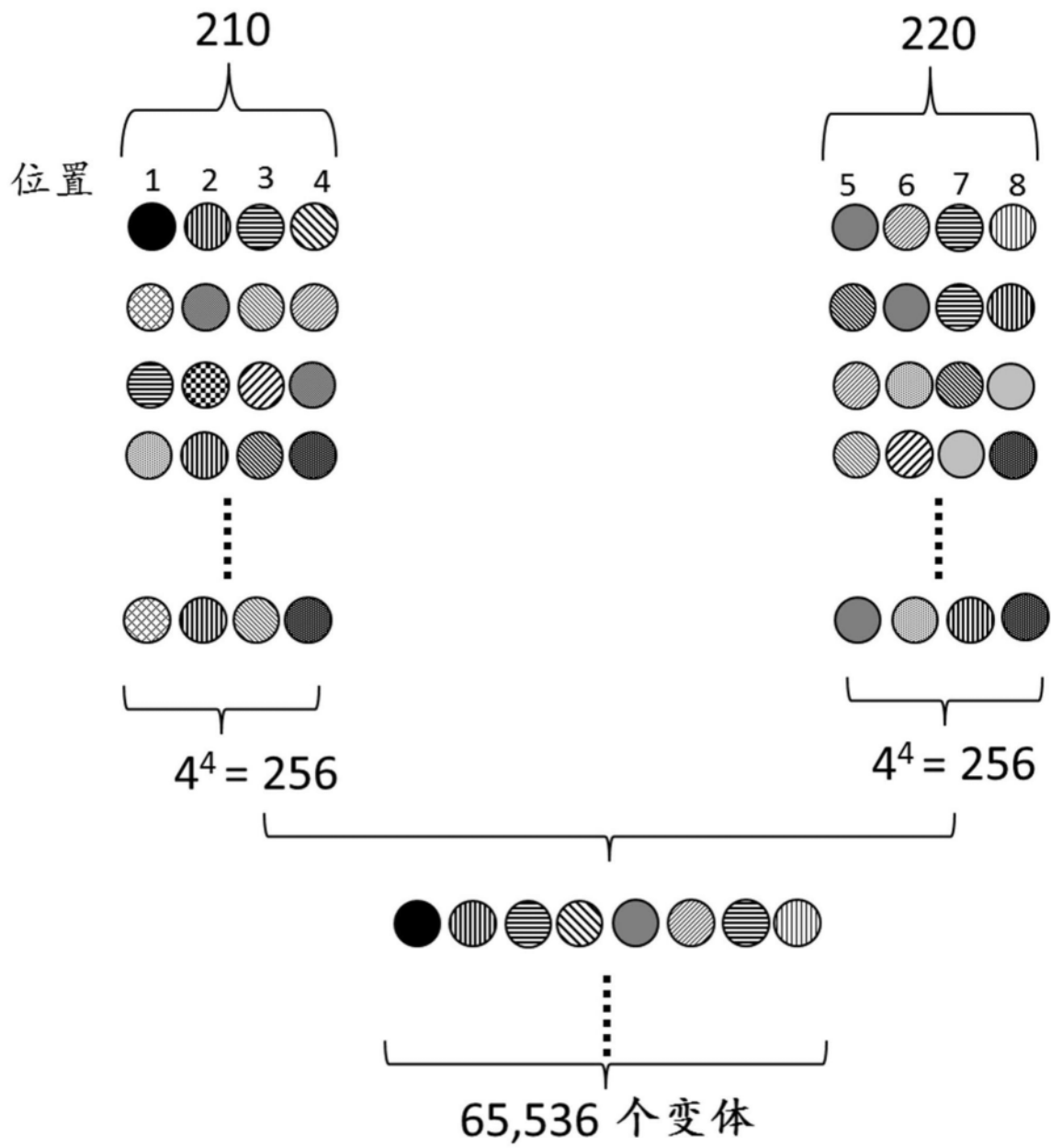


图2

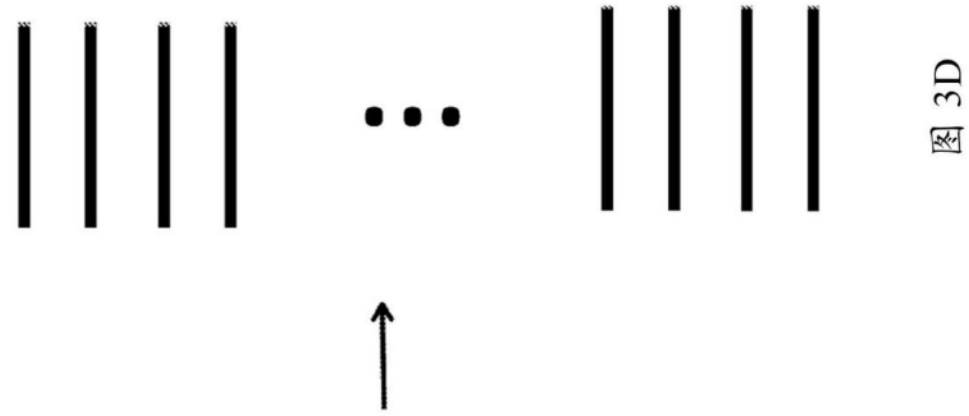


图 3D

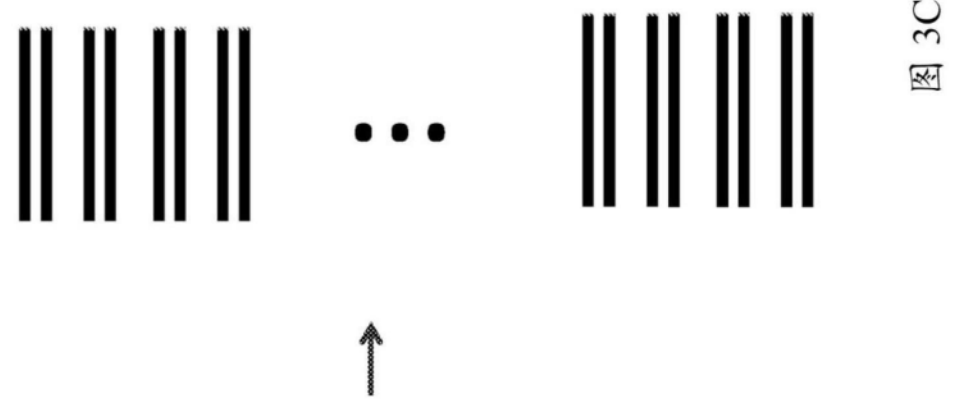


图 3C

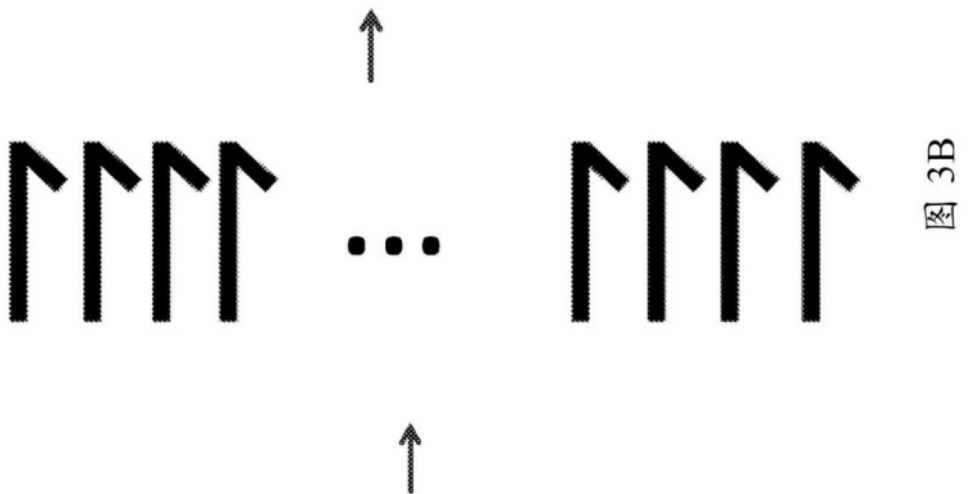


图 3B

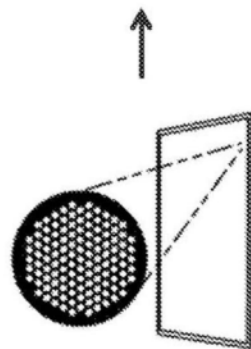


图 3A

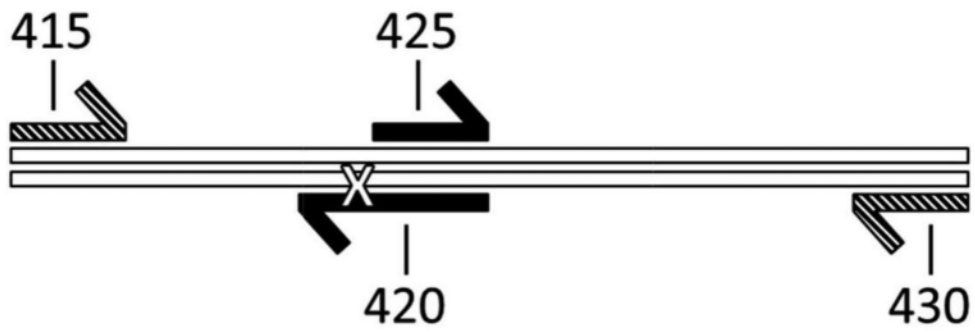


图 4A

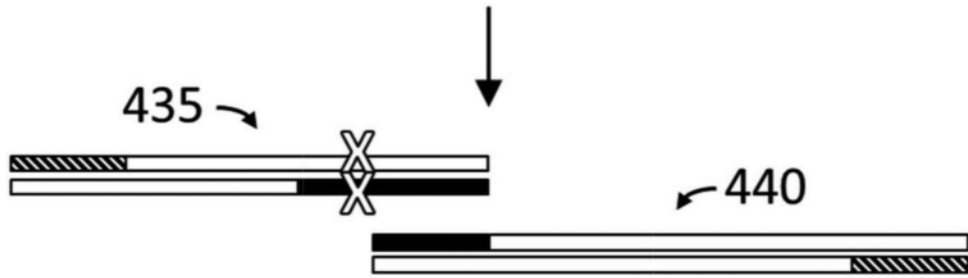


图 4B

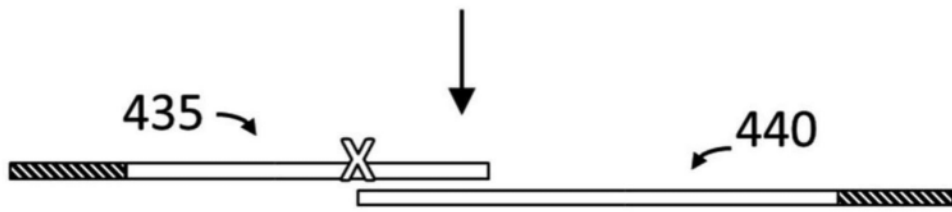


图 4C

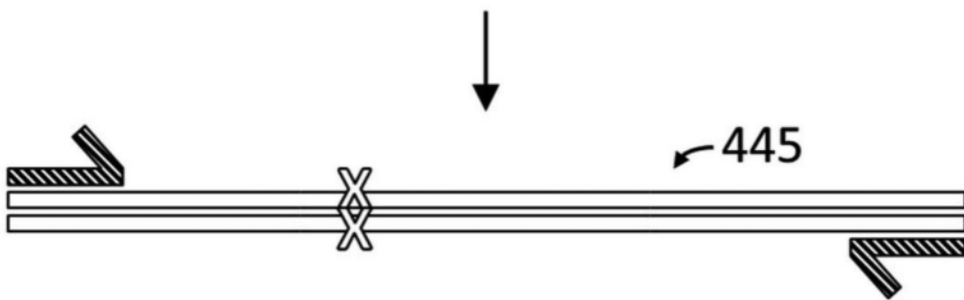


图 4D

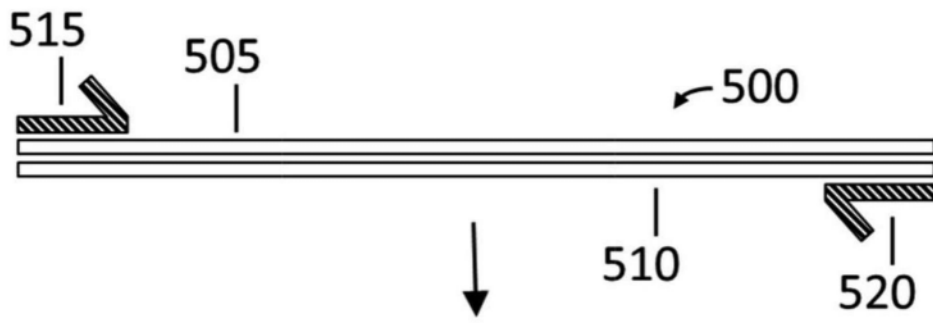


图 5A



图 5B

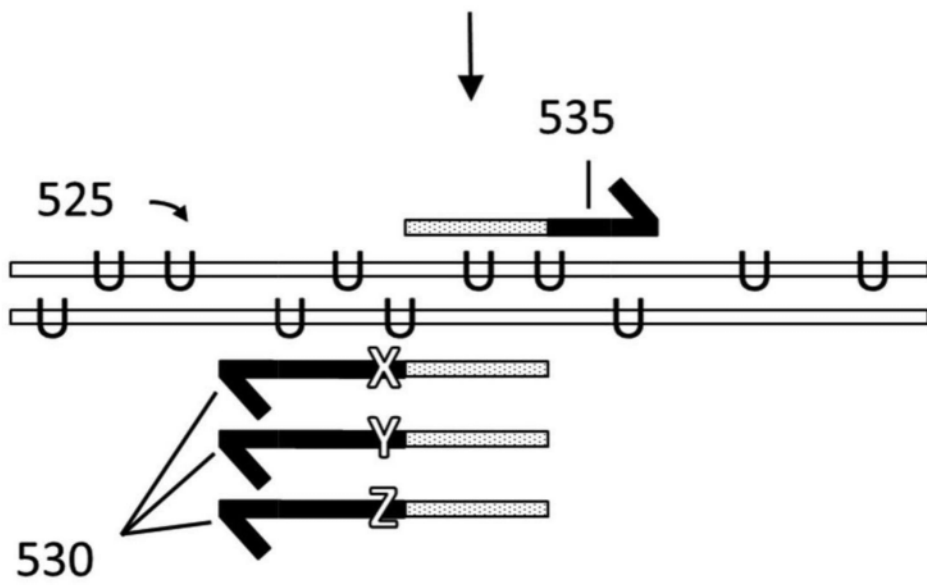


图 5C

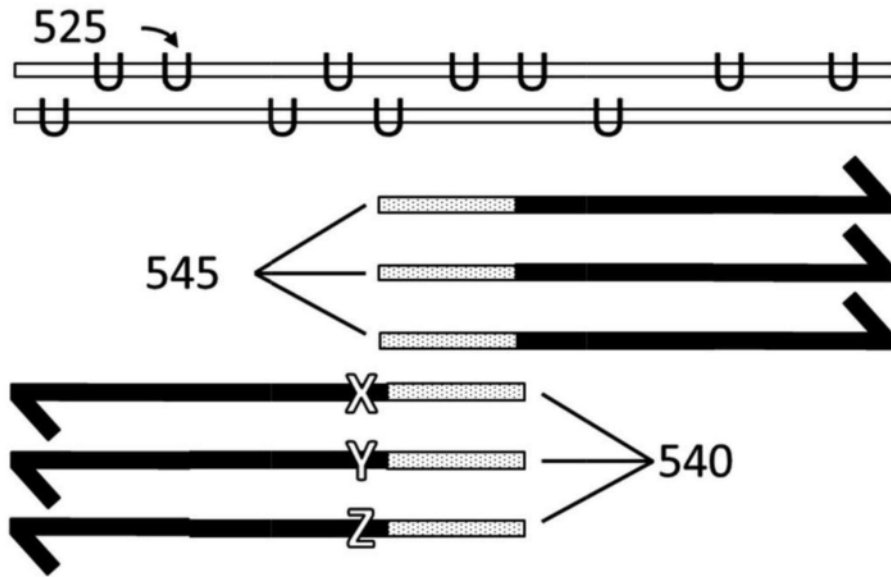


图 5D

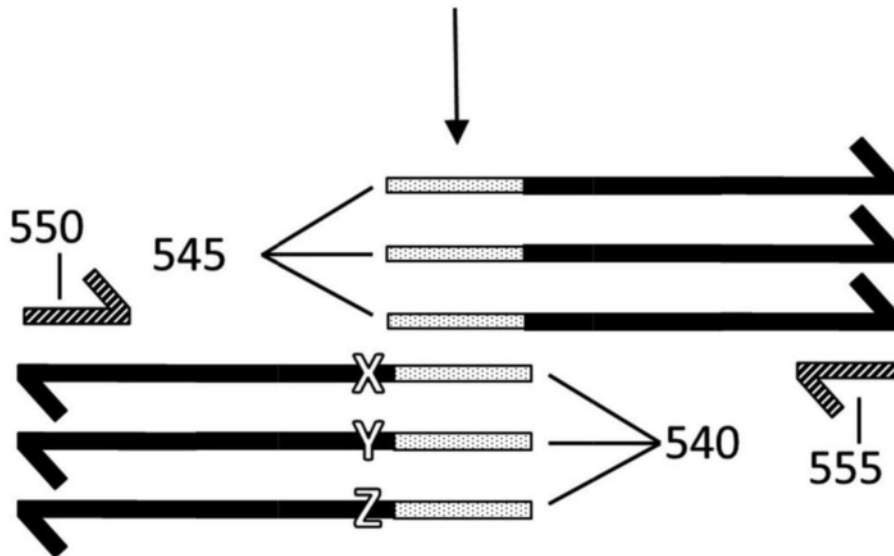


图 5E

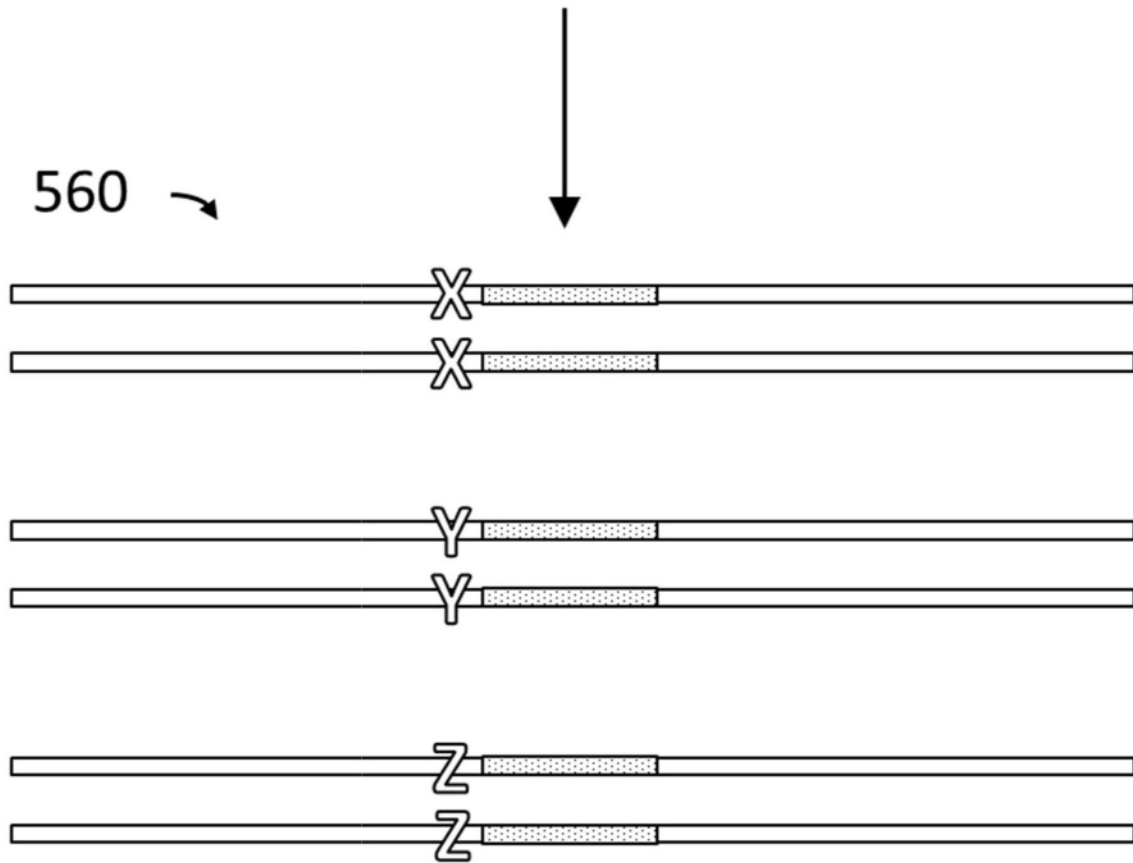


图 5F

图5F

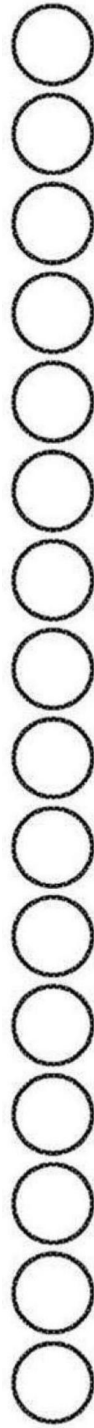


图6A



图6B



图6C



图6D



图6E

A W I K R E Q

图 7A



X	W	I	K	R	E	Q
A	X	I	K	R	E	Q
A	W	X	K	R	E	Q
A	W	I	X	R	E	Q
A	W	I	K	X	E	Q
A	W	I	K	R	X	Q
A	W	I	K	R	E	X

X = 任意密码子

图 7B



图 8A

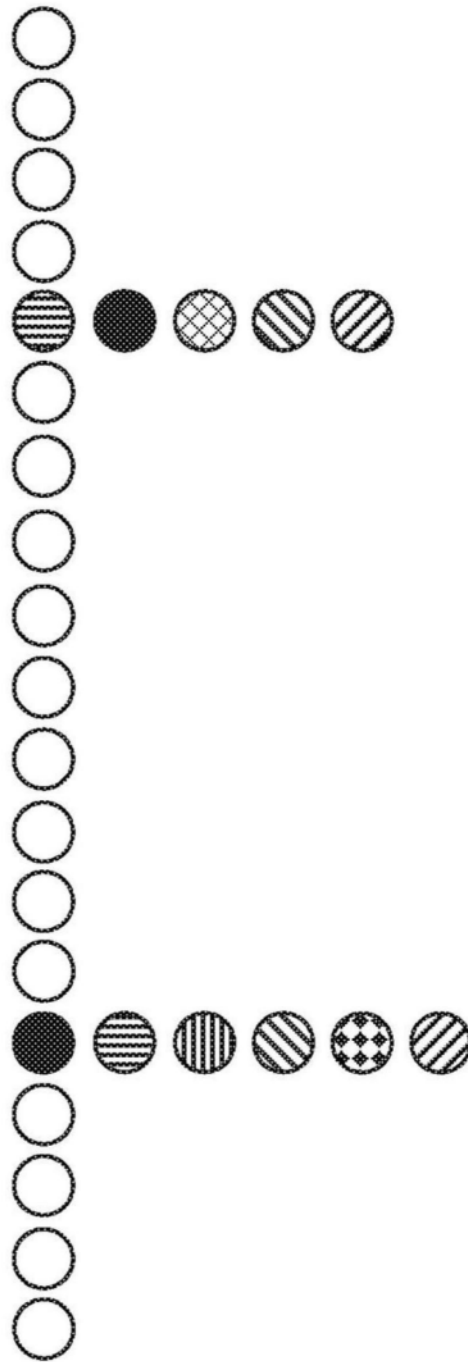


图 8B

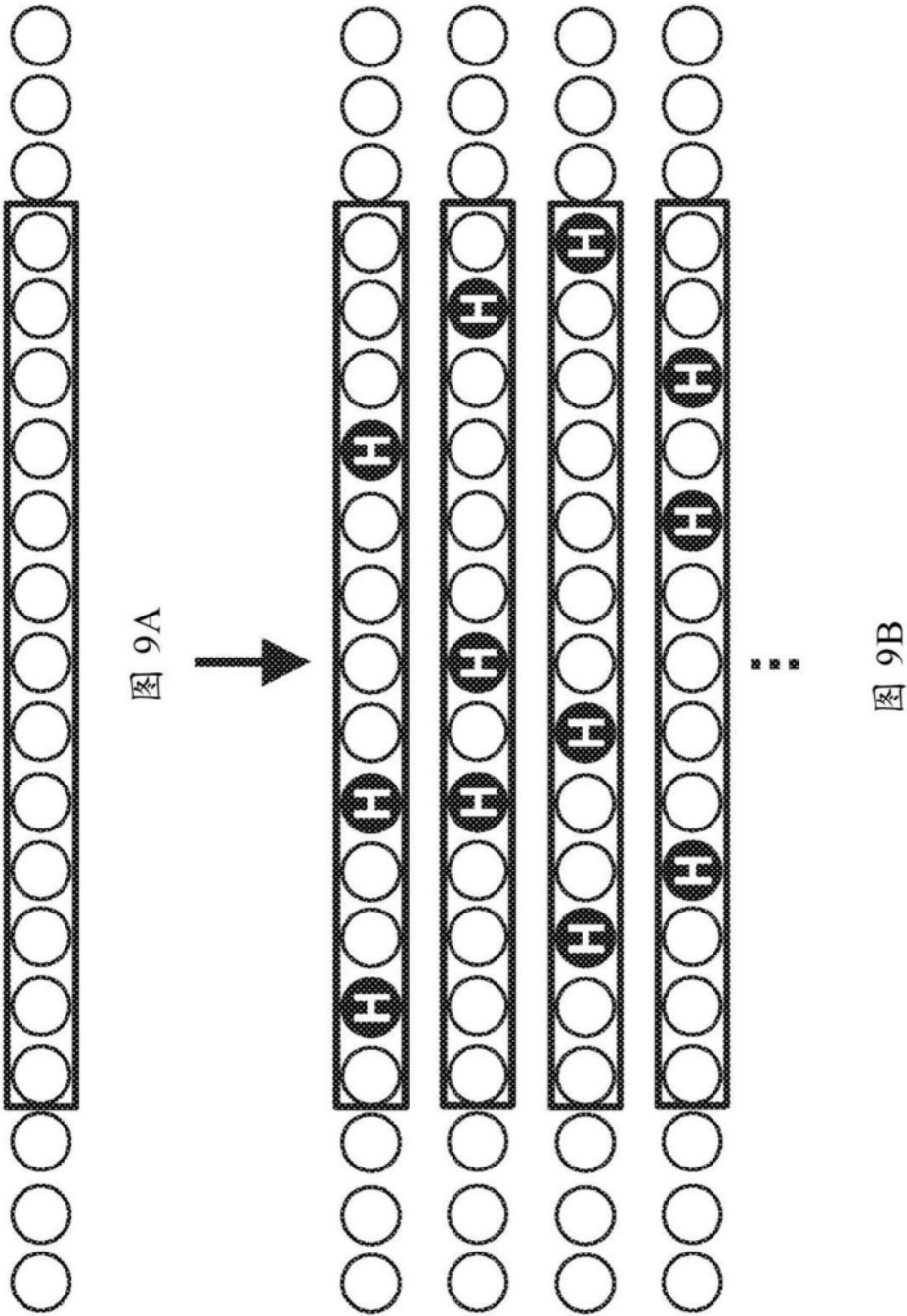
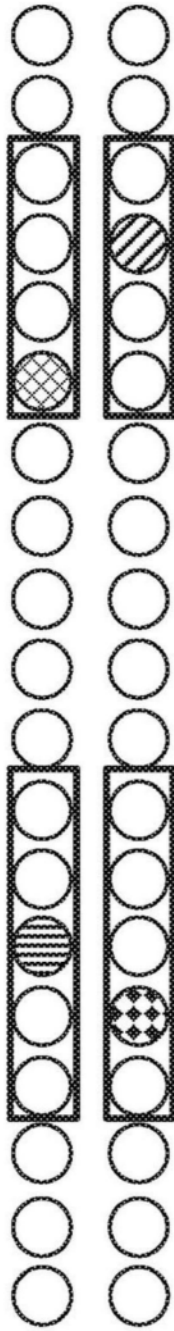
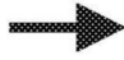




图 10A



...

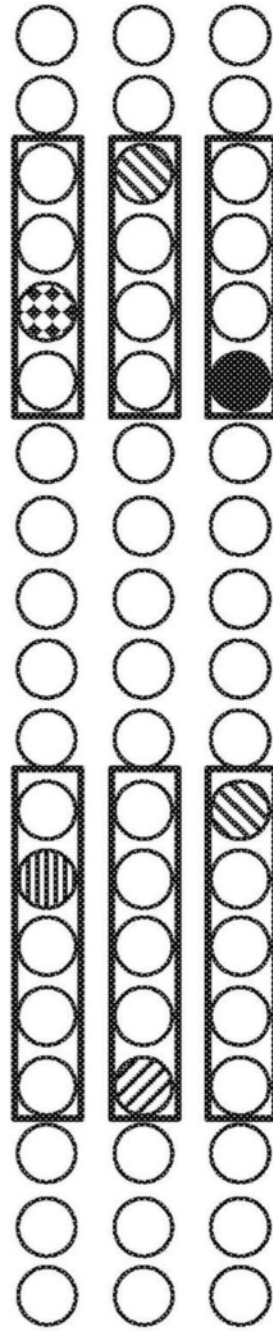


图 10B



图11A

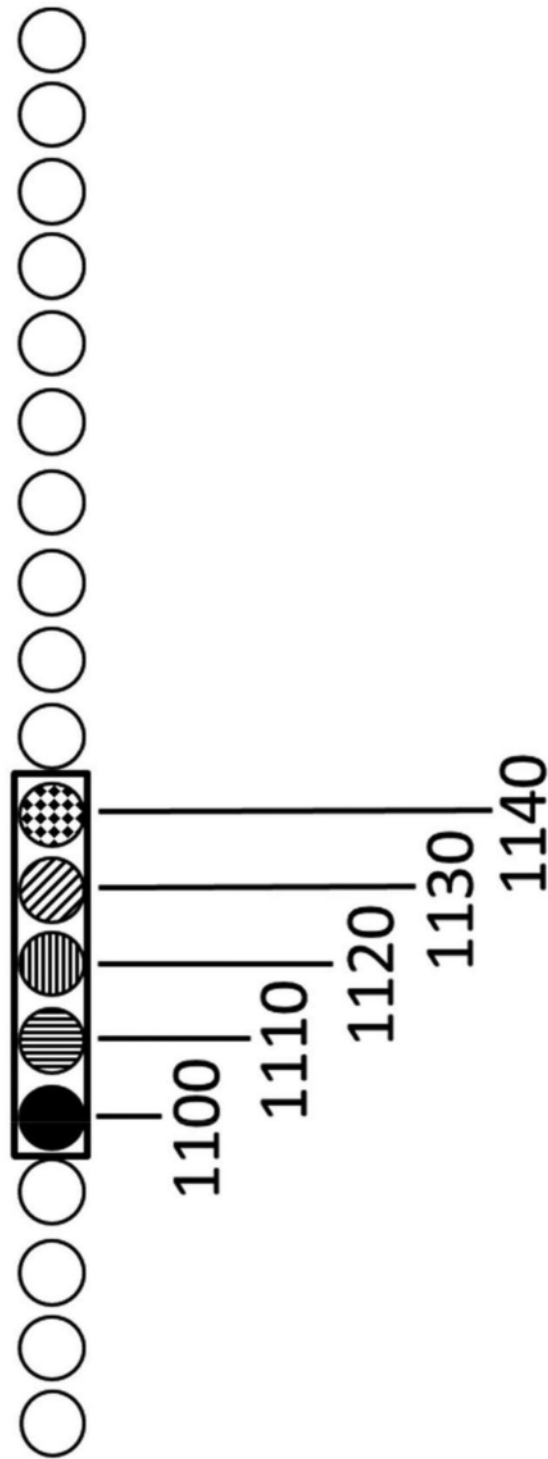


图11B

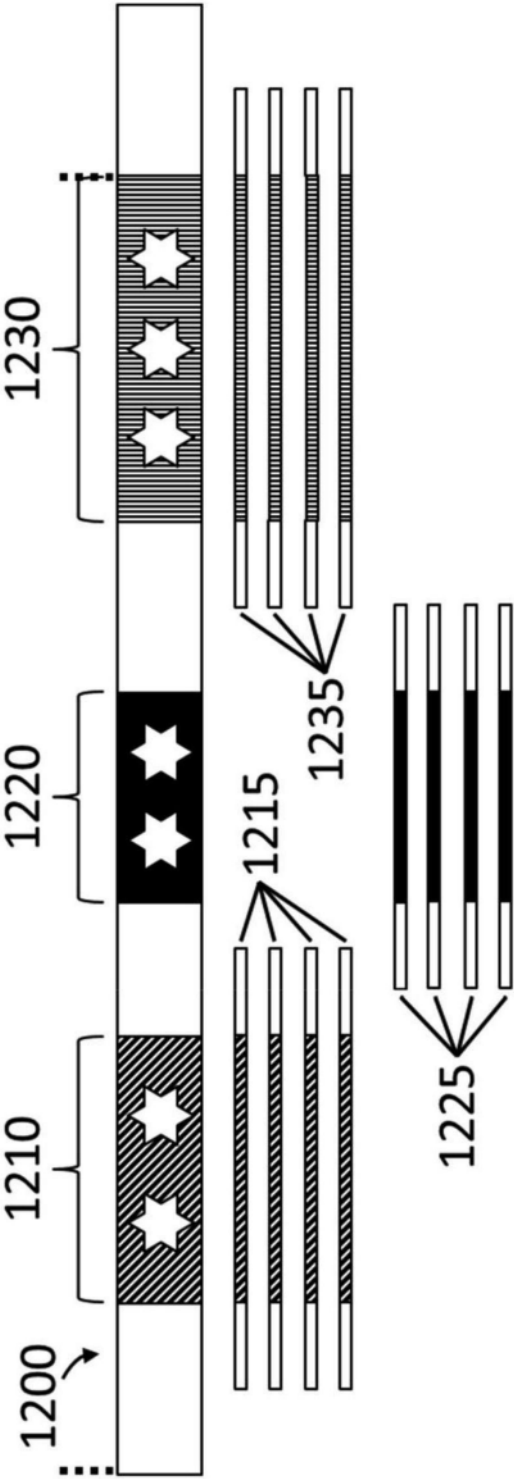


图12

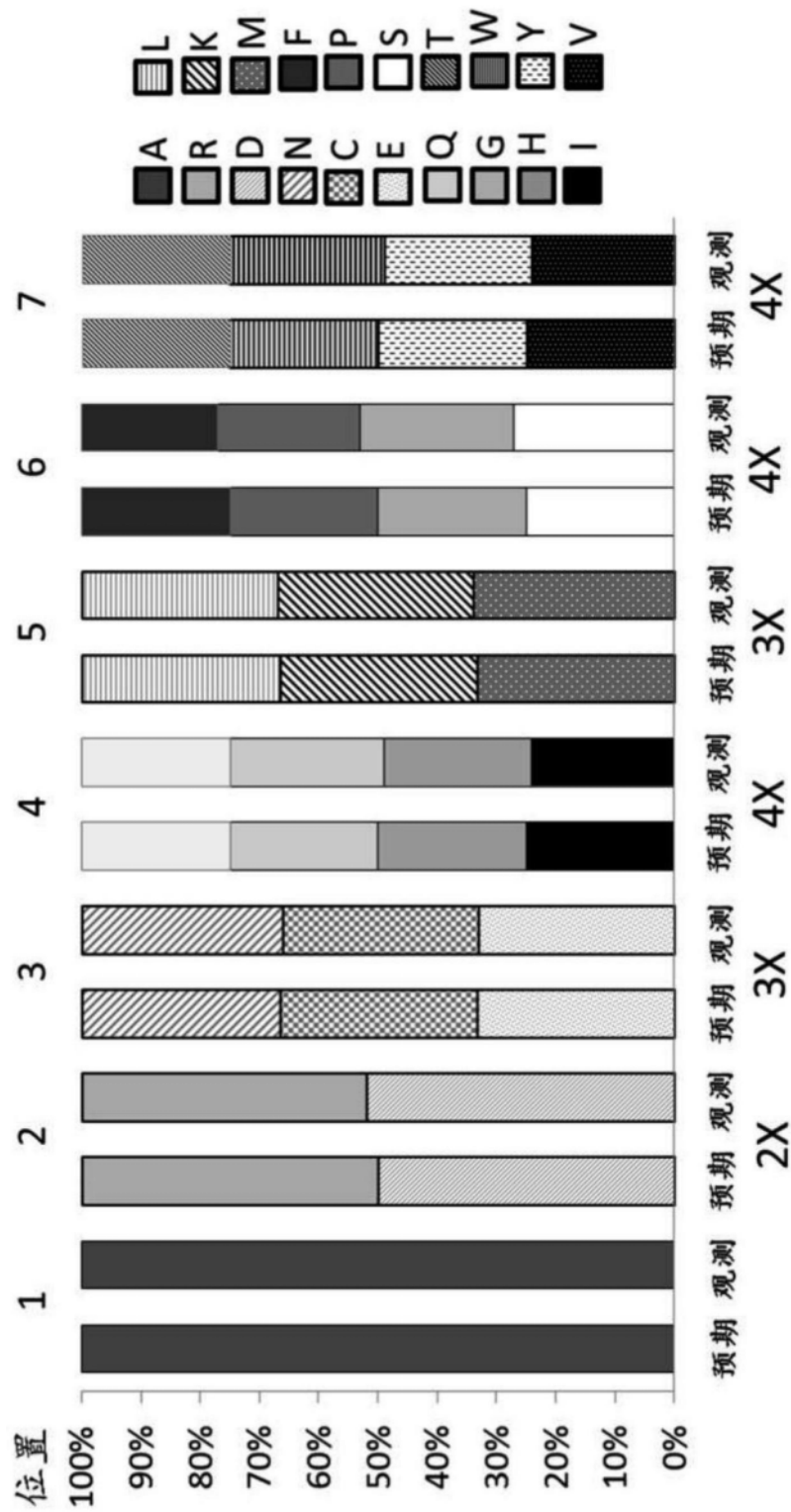


图13

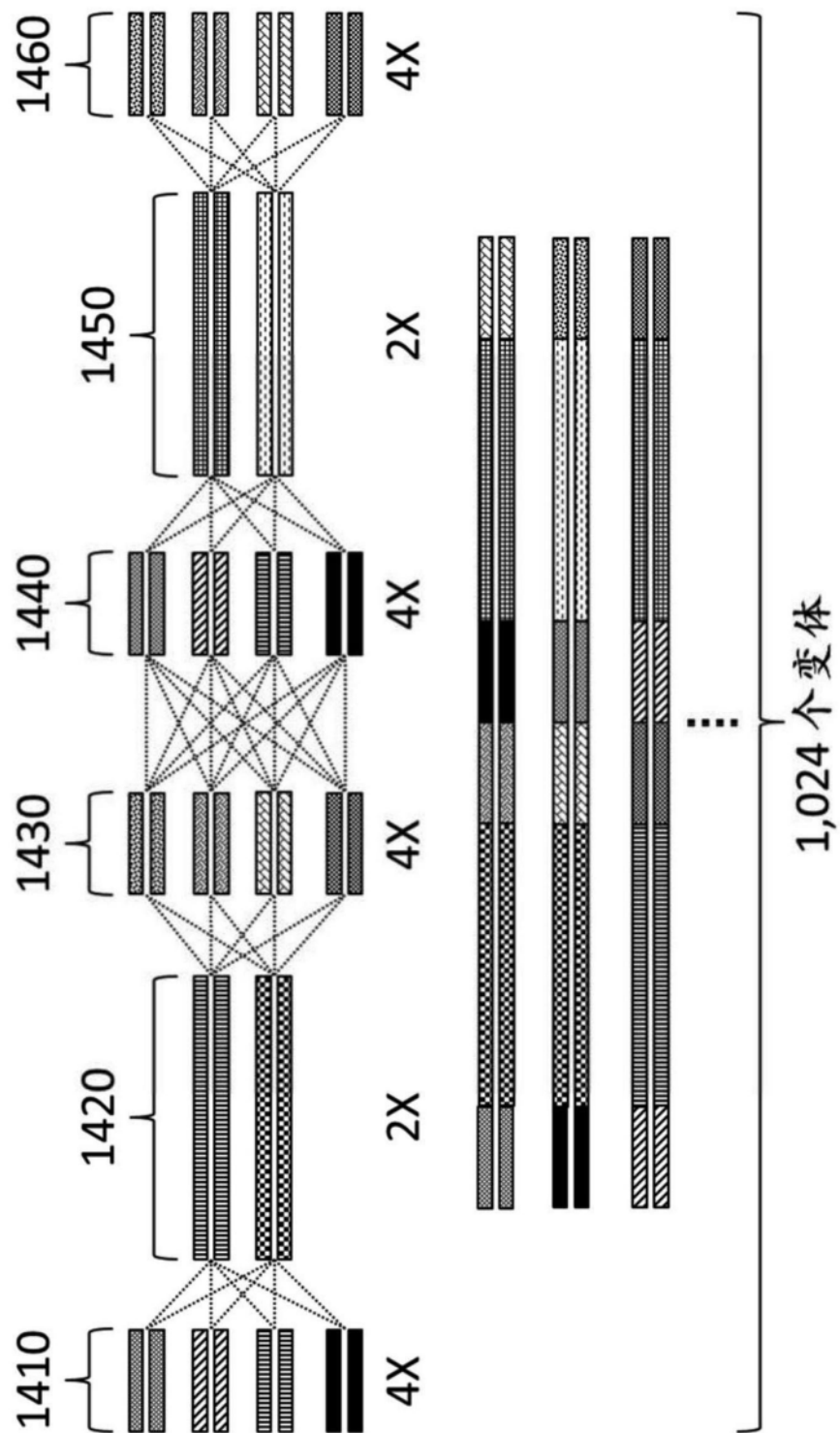


图14

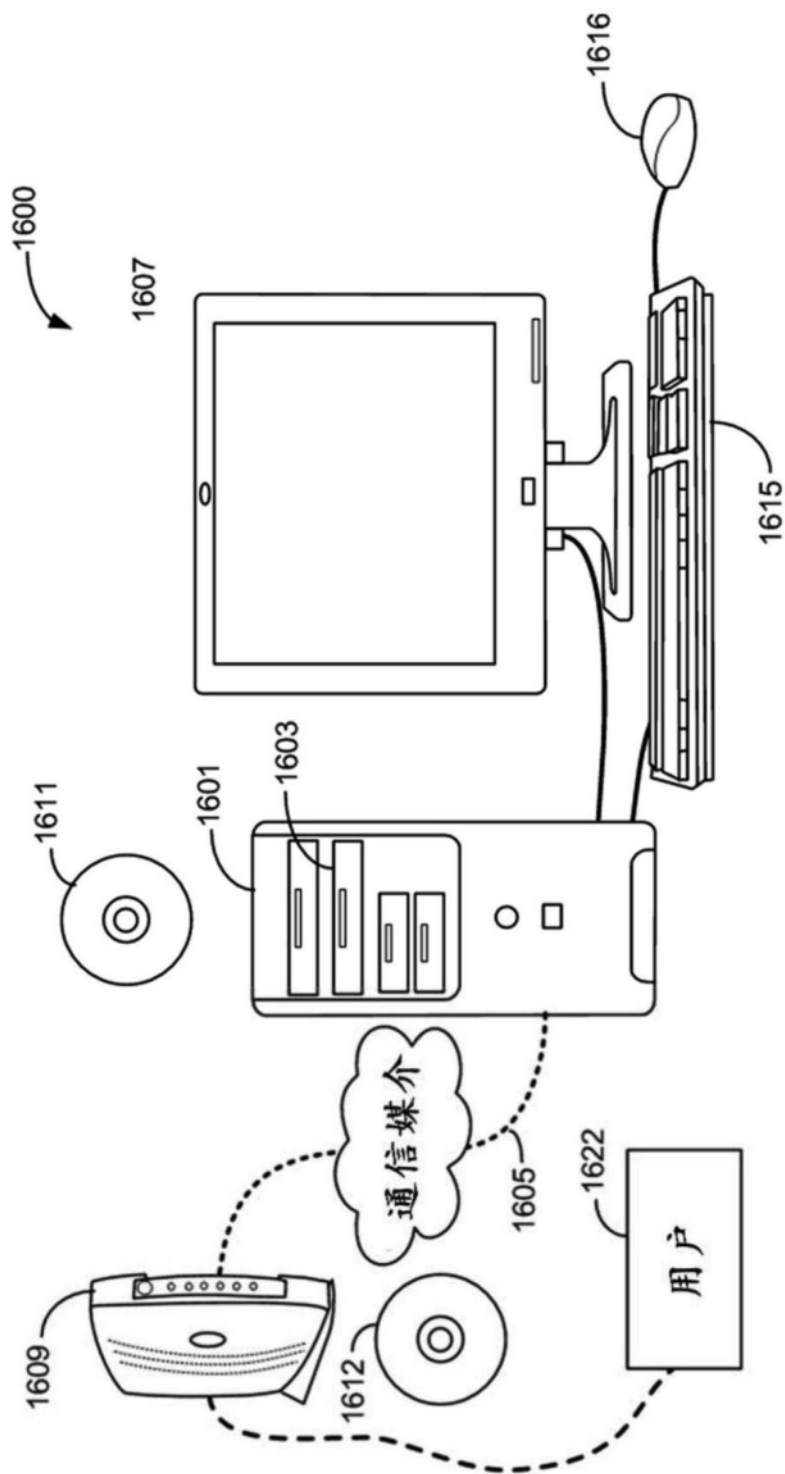


图16

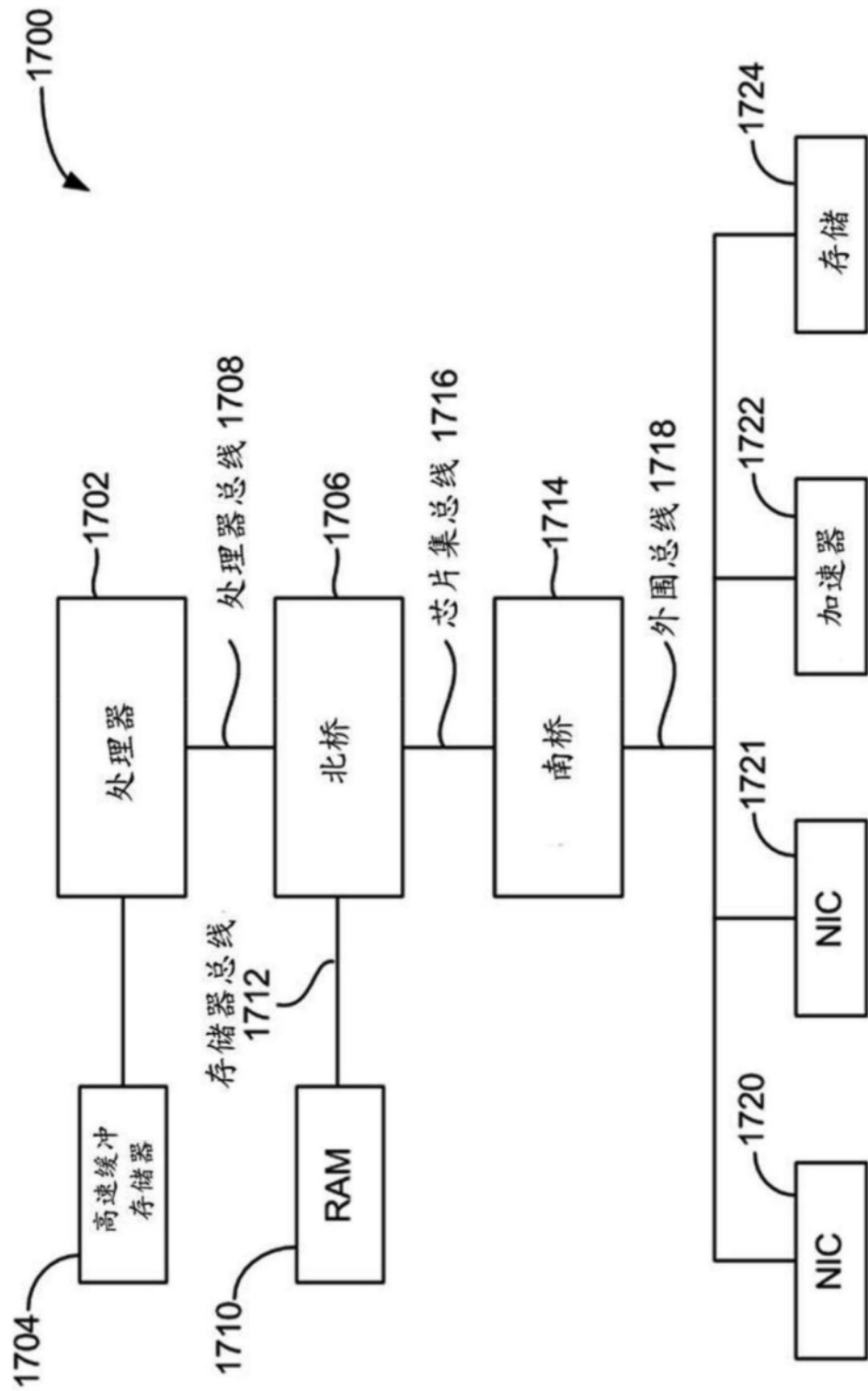


图17

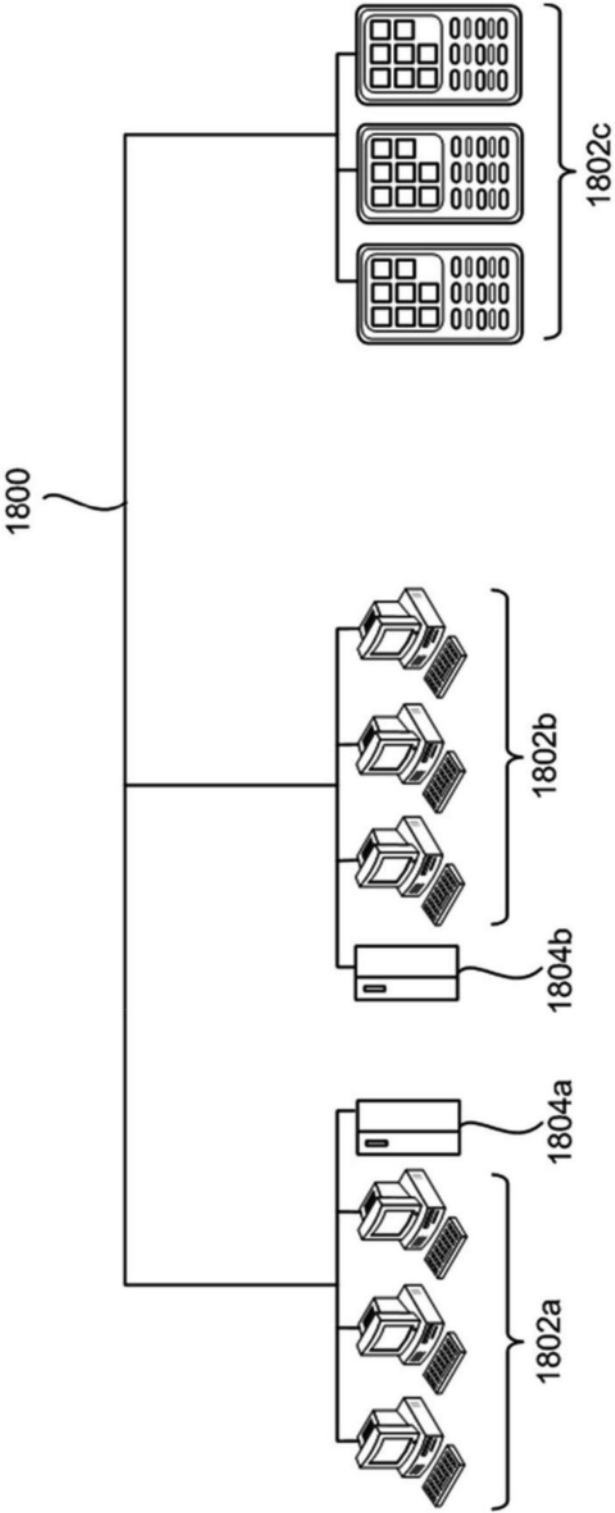


图18

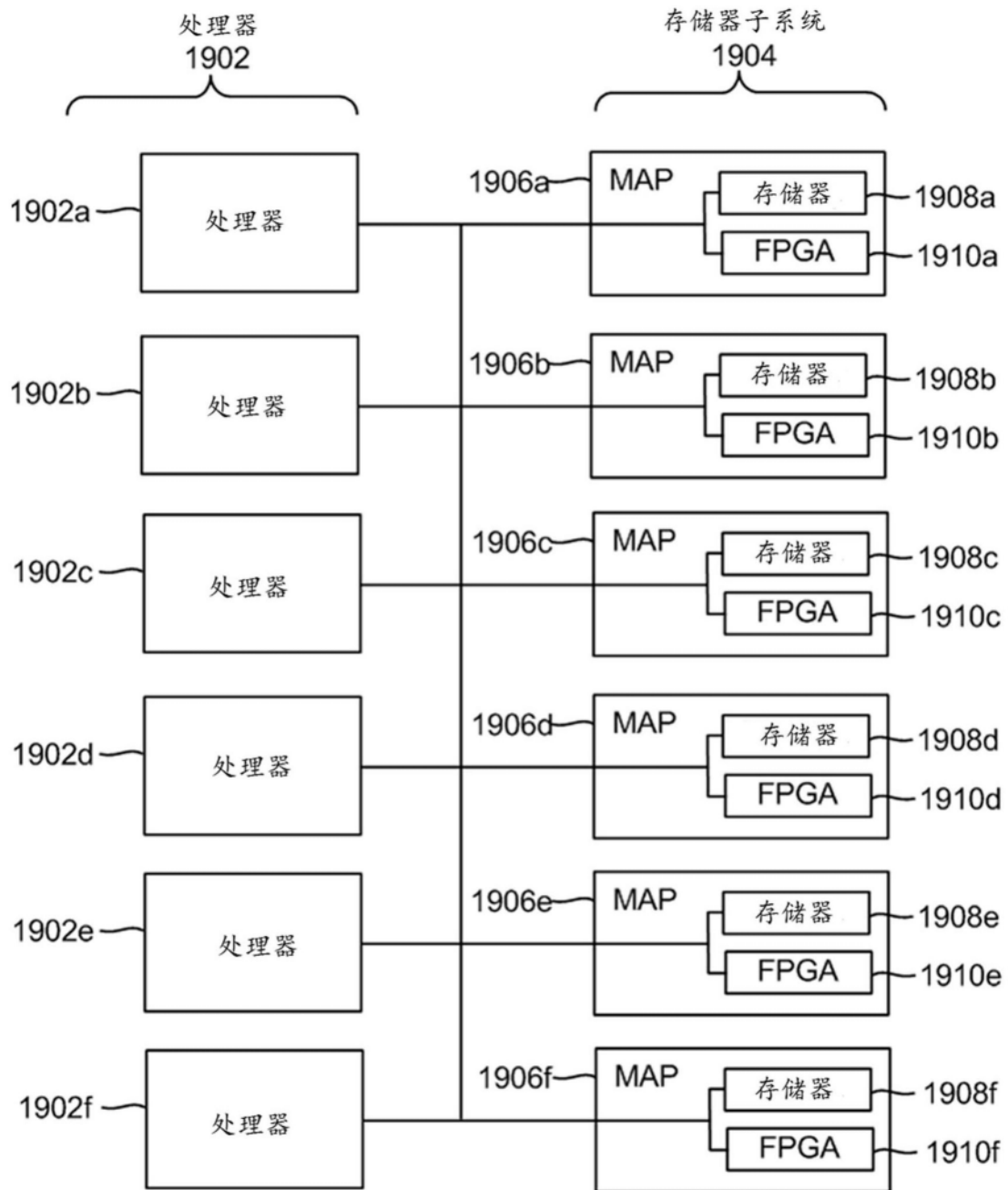


图19

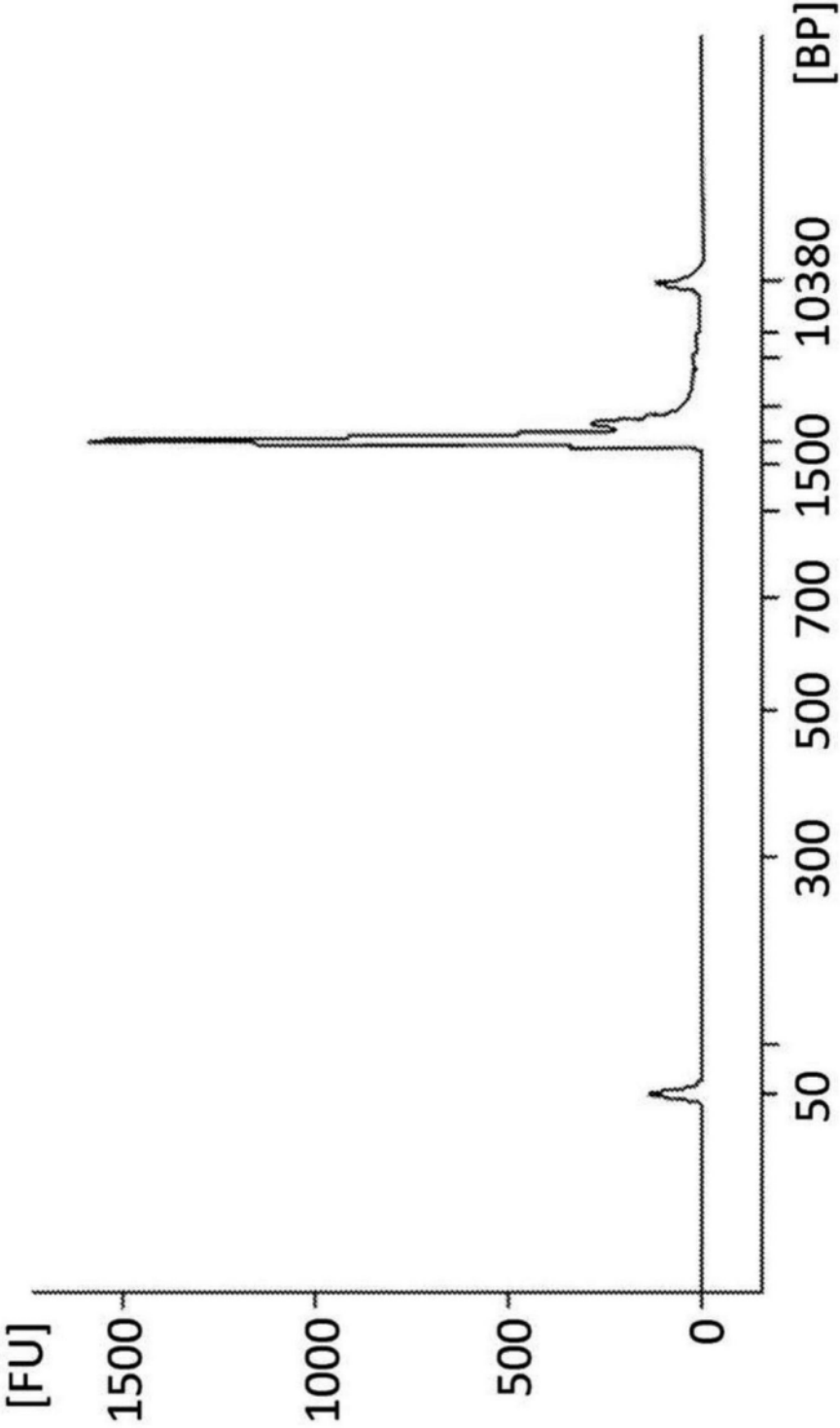


图20

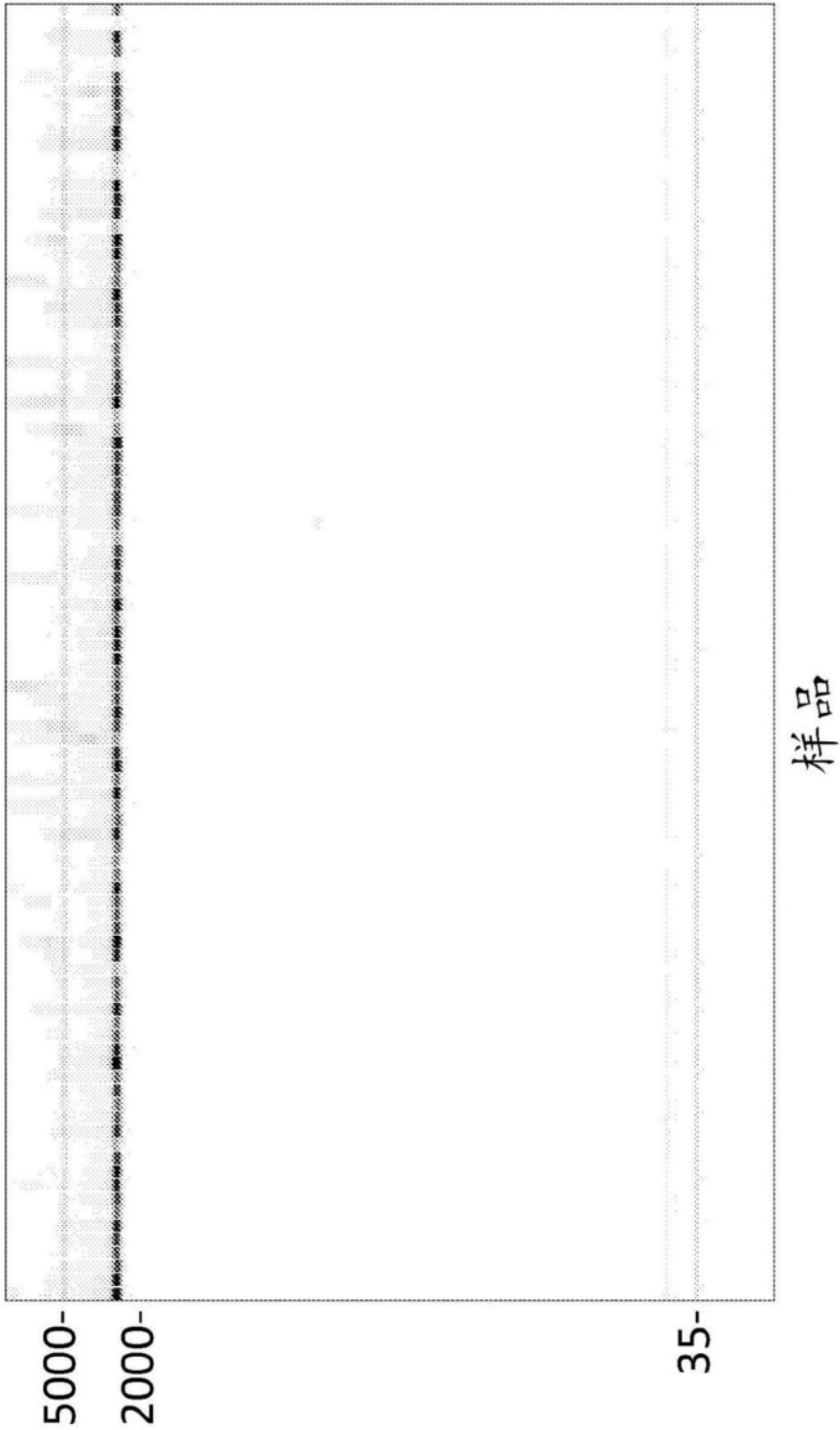


图21

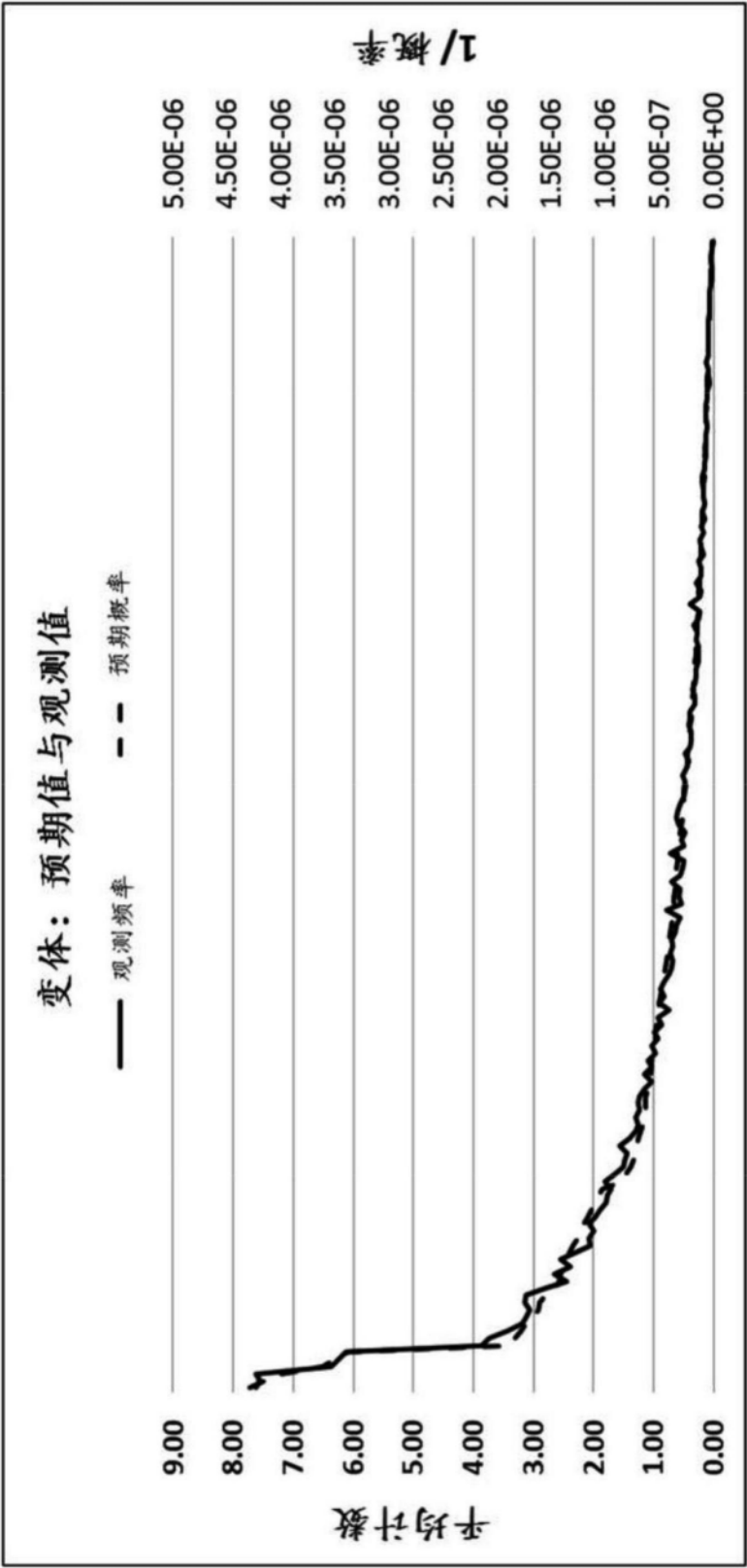


图22



图23

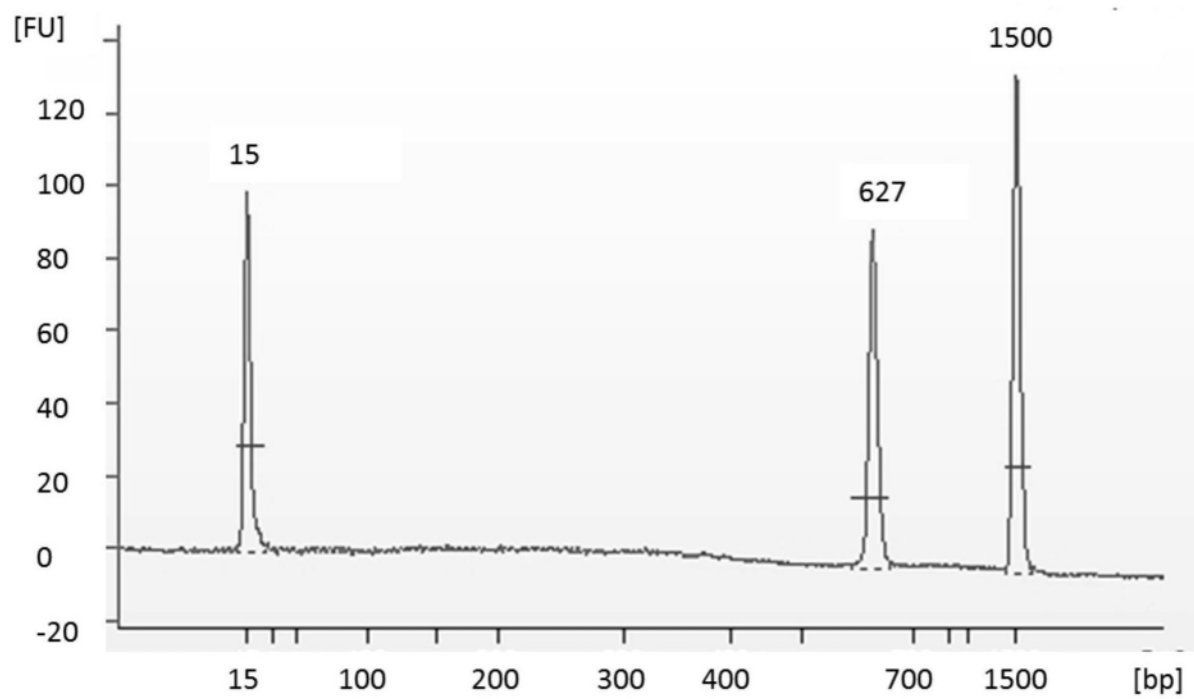


图24

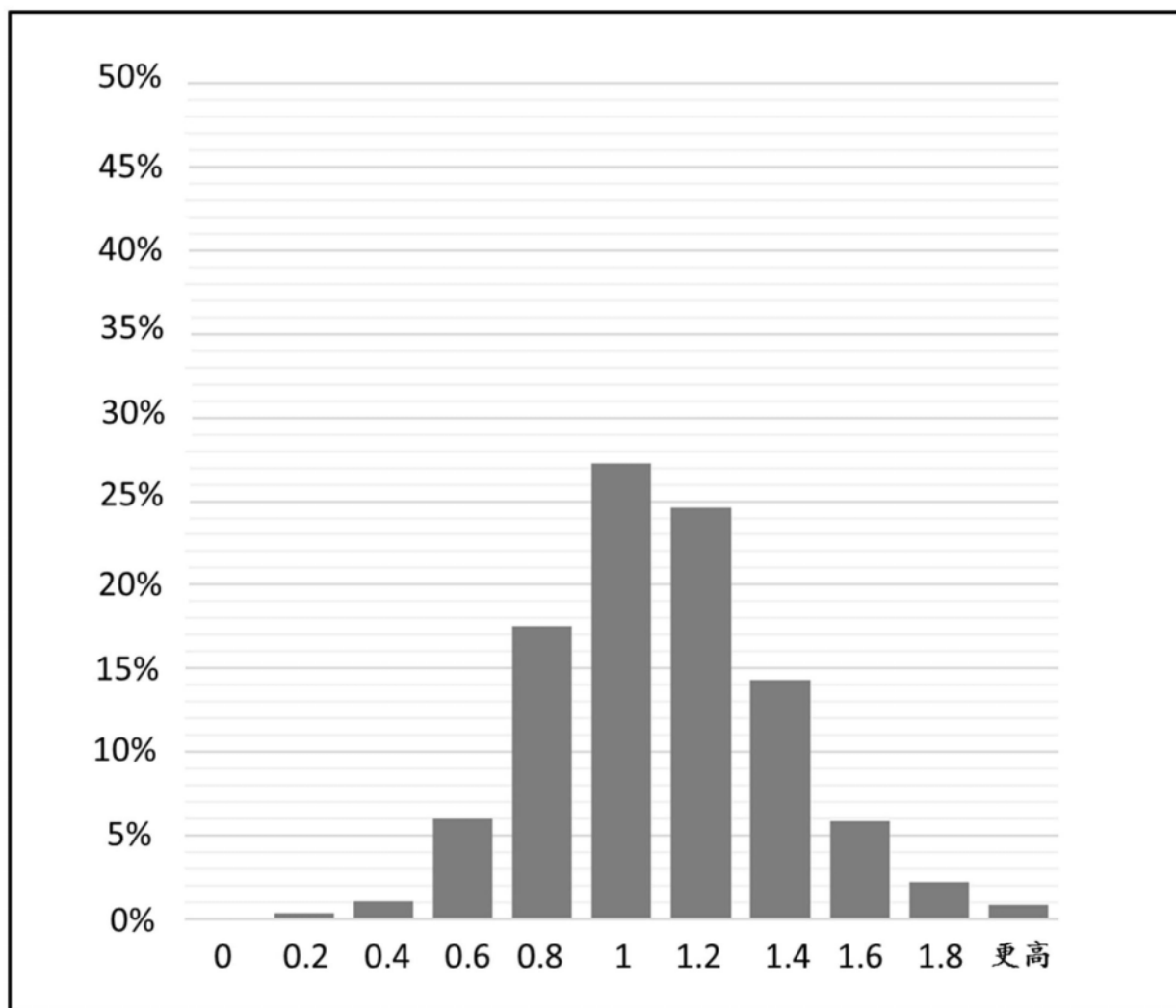


图25

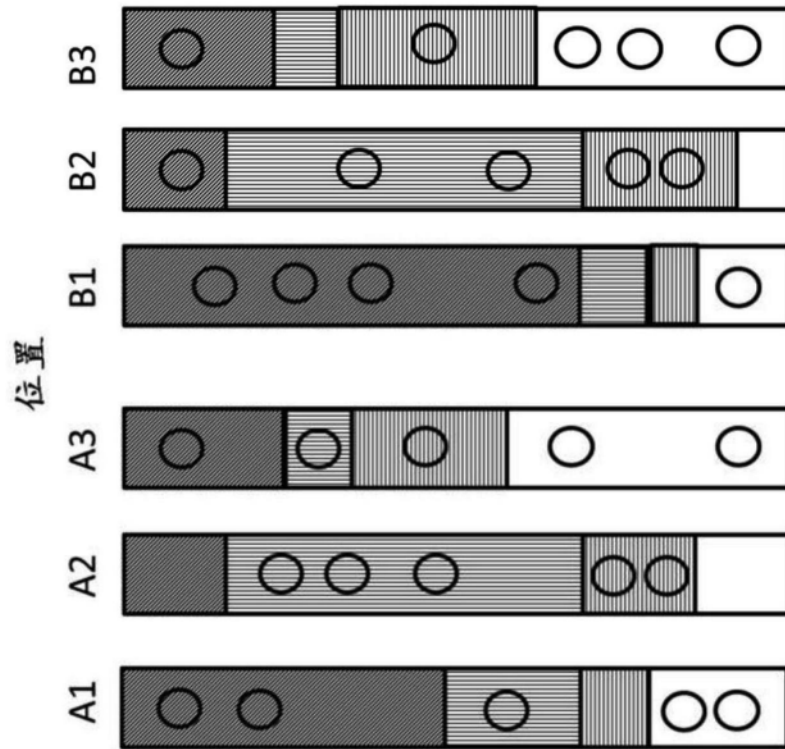


图26A

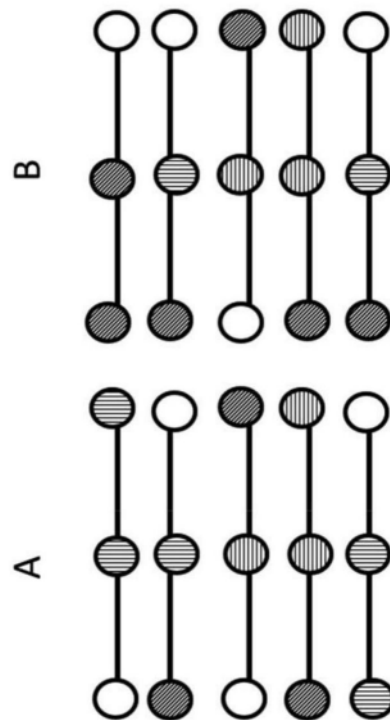


图26B

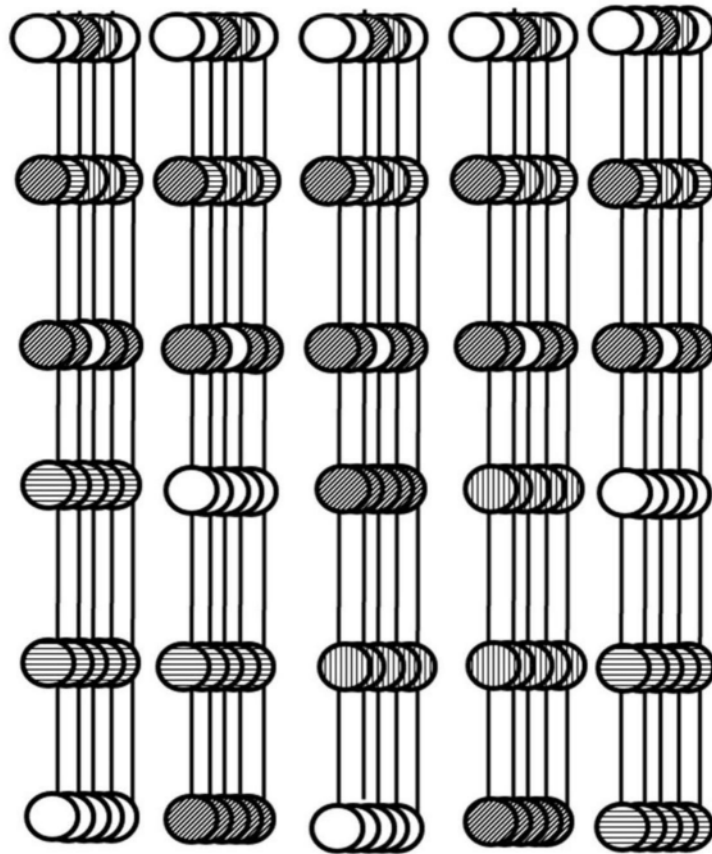


图26C

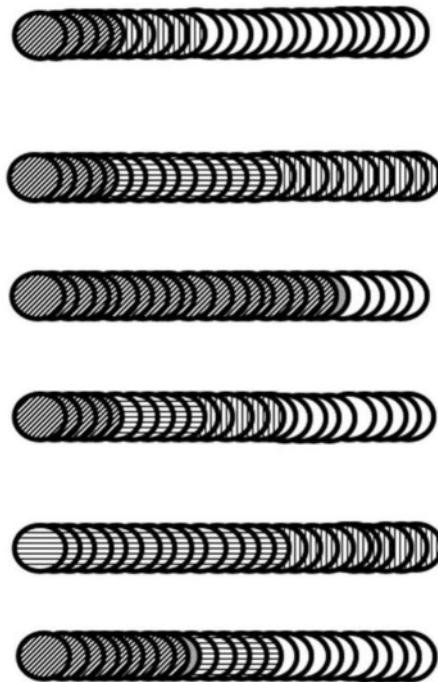


图26D



图27A



图27B



图27C

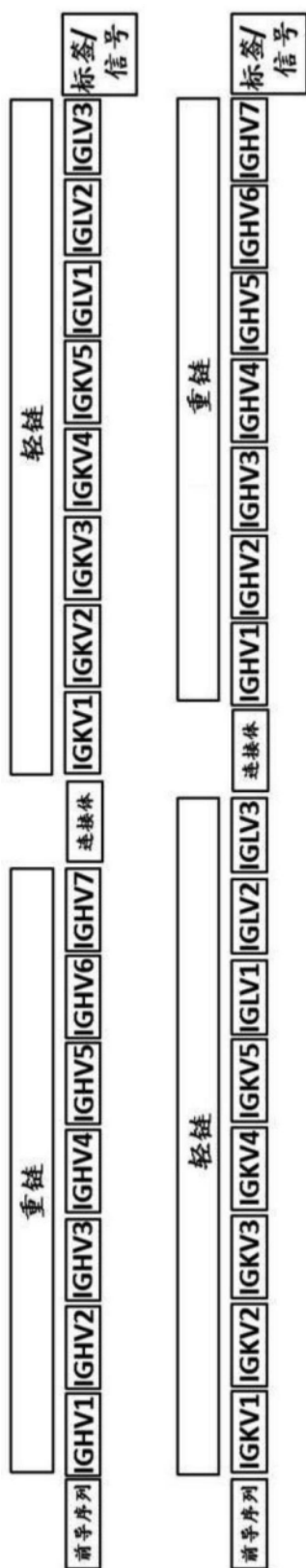


图28A

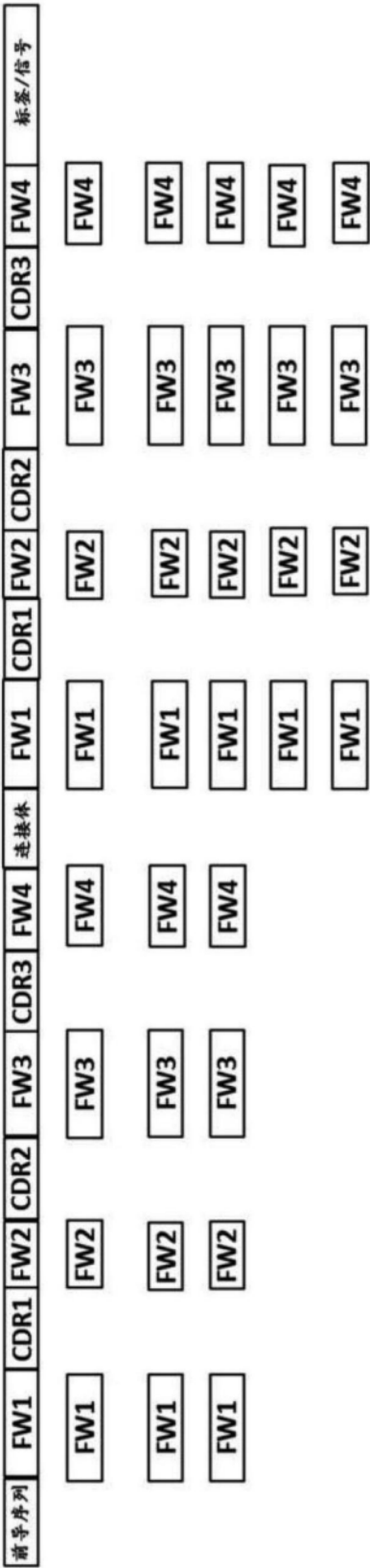


图28B