



## (12)发明专利申请

(10)申请公布号 CN 109564769 A

(43)申请公布日 2019.04.02

(21)申请号 201780048980.4

(22)申请日 2017.06.09

(30)优先权数据

62/348,786 2016.06.10 US

62/375,858 2016.08.16 US

(85)PCT国际申请进入国家阶段日

2019.02.03

(86)PCT国际申请的申请数据

PCT/US2017/036868 2017.06.09

(87)PCT国际申请的公布数据

W02017/214574 EN 2017.12.14

(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 詹姆斯·迪甘斯

(74)专利代理机构 北京安信方达知识产权代理有限公司 11262

代理人 贺淑东

(51)Int.Cl.

G16B 30/00(2019.01)

G16B 50/00(2019.01)

C12N 15/10(2006.01)

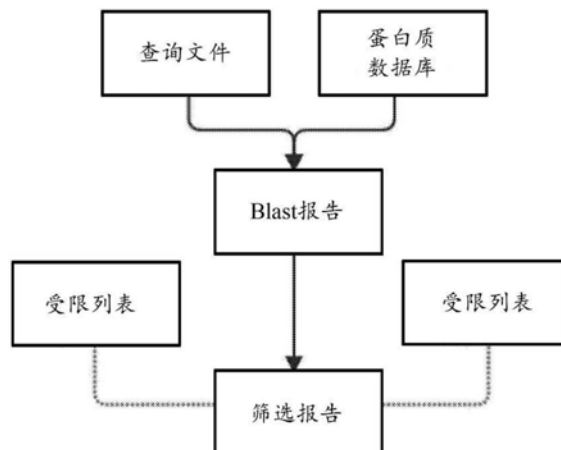
权利要求书3页 说明书19页 附图12页

(54)发明名称

用于自动注释和筛选生物序列的系统和方法

(57)摘要

本公开内容描述了用于基于团体知识和参与的有效生物安全的软件工具。本文描述的注释工具为合成生物学团体跟踪关于个体蛋白质与负面后果之间的联系的新兴科学提供了帮助。本文描述的筛选工具使得团体能够拓宽对生物安全性的兴趣和有效实践,从而使从业者和生物序列或构建体提供者有能力评估订单请求的安全性,而不用等到合成或甚至表达。此外,本文描述的筛选工具能够筛选同一订单或多个订单中的多核苷酸中与来自参考数据库的有害生物序列相关的序列。



1. 一种用于提供增强的多核苷酸合成的计算机化系统：
  - a) 用于托管数据库的服务器, 其中所述数据库适于表示有害生物序列的列表;
  - b) 网络连接; 以及
  - c) 包含用于通用计算机的指令的计算机可读介质, 其中所述计算机化系统被配置用于以包括以下步骤的方法操作:
    - i) 接收一个或多个设计指令, 其中所述设计指令包含多个生物序列, 其中每个所述生物序列的长度不超过500个碱基, 并且其中所述多个生物序列包括核酸或氨基酸序列;
    - ii) 自动确定所述多个生物序列中的至少两个生物序列是否共同对应于所述数据库中的有害生物序列的至少20%; 以及
    - iii) 如果检测到所述有害生物序列的至少20%, 则自动生成警报。
2. 根据权利要求1所述的系统, 其进一步包括其中如果没有生成警报, 则合成一个或多个序列。
3. 根据权利要求1所述的系统, 其进一步包括接收关于改变所述多个生物序列中对应于所述有害生物序列的至少20%的所述至少两个生物序列以去除所述有害生物序列的指令。
4. 根据权利要求1或3所述的系统, 其中在一个或多个时间点接收所述多个接收的设计指令。
5. 根据权利要求1至4中任一项所述的系统, 其中所述多个接收的设计指令来自不同的来源。
6. 根据权利要求5所述的系统, 其中所述多个接收的设计指令来自3个或更多个不同的来源。
7. 根据权利要求5所述的系统, 其中所述多个接收的设计指令来自5个或更多个不同的来源。
8. 根据权利要求5所述的系统, 其中所述多个接收的设计指令来自10个或更多个不同的来源。
9. 根据权利要求1至8中任一项所述的系统, 其中所述一个或多个生物序列的长度均不超过200个碱基。
10. 根据权利要求9所述的系统, 其中所述一个或多个生物序列的长度均不超过100个碱基。
11. 根据权利要求9所述的系统, 其中所述一个或多个生物序列的长度均不超过50个碱基。
12. 根据权利要求9所述的系统, 其中所述一个或多个生物序列的长度均不超过20个碱基。
13. 一种用于提供增强的多核苷酸合成的方法, 其包括:
  - a) 接收一个或多个设计指令, 其中所述设计指令包含多个生物序列, 其中每个所述生物序列的长度不超过500个碱基, 并且其中所述多个生物序列包括核酸或氨基酸序列;
  - b) 自动确定所述多个生物序列中的至少两个生物序列是否共同对应于数据库中的有害生物序列的至少20%; 以及
  - c) 如果检测到所述有害生物序列的至少20%, 则自动生成警报。

14. 根据权利要求13所述的方法,其进一步包括其中如果没有生成警报,则合成所述一个或多个序列。

15. 根据权利要求13所述的方法,其进一步包括接收关于改变所述多个生物序列中对应于所述有害生物序列的至少20%的至少两个生物序列以去除所述有害生物序列的指令。

16. 一种用于提供增强的多核苷酸合成的计算机化系统:

a) 用于托管数据库的服务器,其中所述数据库适于表示序列列表;  
b) 网络连接;以及  
c) 包含用于通用计算机的指令的计算机可读介质,其中所述计算机化系统被配置用于以包括以下步骤的方法操作:

i) 接收一个或多个设计指令,其中所述设计指令包含多个生物序列,其中所述多个生物序列包括载体序列和多个另外的插入序列;

ii) 自动确定所述多个插入序列中的至少一个和所述载体是否共同对应于所述数据库中的有害生物序列的至少20%;以及

iii) 如果检测到所述有害生物序列的至少20%,则自动生成警报。

17. 根据权利要求16所述的系统,其中如果没有生成警报,则合成所述一个或多个生物序列。

18. 根据权利要求16所述的系统,其进一步包括接收关于改变对应于所述有害生物序列的至少20%的所述多个插入序列中的所述至少一个和所述载体以去除所述有害生物序列的指令。

19. 根据权利要求16至18中任一项所述的系统,其中在一个或多个时间点接收所述多个接收的设计指令。

20. 根据权利要求16至19中任一项所述的系统,其中从不同的来源接收所述多个接收的设计指令。

21. 根据权利要求20所述的系统,其中所述多个接收的设计指令来自3个或更多个不同的来源。

22. 根据权利要求20所述的系统,其中所述多个接收的设计指令来自5个或更多个不同的来源。

23. 根据权利要求20所述的系统,其中所述多个接收的设计指令来自10个或更多个不同的来源。

24. 根据权利要求16至23中任一项所述的系统,其中所述一个或多个生物序列的长度不超过200个碱基。

25. 根据权利要求24所述的系统,其中所述一个或多个生物序列的长度均不超过100个碱基。

26. 根据权利要求24所述的系统,其中所述一个或多个生物序列的长度均不超过50个碱基。

27. 根据权利要求24所述的系统,其中所述一个或多个生物序列的长度均不超过20个碱基。

28. 一种用于提供增强的多核苷酸合成的方法,其包括:

a) 接收一个或多个设计指令,其中所述设计指令包含多个生物序列,其中所述多个生

物序列是载体序列和多个另外的插入序列；

b) 自动确定所述多个插入序列中的至少一个和所述载体是否共同对应于数据库中的有害生物序列的至少20%；以及

c) 如果检测到所述有害生物序列的至少20%，则自动生成警报。

29. 根据权利要求28所述的方法，其中所述生物序列获自对实体核酸或蛋白质样品进行测序。

30. 根据权利要求28所述的方法，其接收关于改变对应于所述有害生物序列的至少20%的所述多个插入序列中的所述至少一个和所述载体以去除所述有害生物序列的指令。

31. 根据权利要求28至30中任一项所述的方法，其进一步包括其中如果没有生成警报，则合成所述一个或多个生物序列。

## 用于自动注释和筛选生物序列的系统和方法

### 交叉引用

[0001] 本申请要求于2016年6月10日提交的美国临时专利申请号62/348,786和于2016年8月16日提交的美国临时专利申请号62/375,858的权益,以上专利申请中的每一个通过引用以其全文并入。

### 背景技术

[0002] 我们关于能够对公共安全和/或环境构成潜在威胁的个体蛋白质和生物系统的总体知识正在非常快速地增长。然而,该知识广泛分布在不同的研究团体、机构、甚至期刊上。而对于给定蛋白质,缺乏专注于对其造成危害的可能性以及在何种情景下可能产生这种危害进行注释的集中信息源。因此,需要新的系统和方法来应对这一挑战。

### 发明内容

[0003] 本文提供了用于提供增强的多核苷酸合成的计算机化系统,其包含用于托管数据库的服务器,其中所述数据库适于表示有害生物序列的列表;网络连接;以及包含用于通用计算机的指令的计算机可读介质,其中所述计算机化系统被配置用于以包括以下步骤的方法操作:1)接收一个或多个设计指令,其中所述设计指令包含多个生物序列,其中每个所述生物序列的长度不超过500个碱基,并且其中所述多个生物序列包括核酸或氨基酸序列;2)自动确定所述多个生物序列中的至少两个生物序列是否共同对应于所述数据库中的有害生物序列的至少20%;以及3)如果检测到所述有害生物序列的至少20%,则自动生成警报。本文进一步提供了这样的计算机化系统,其进一步包括其中如果没有生成警报,则合成一个或多个序列。本文进一步提供了这样的计算机化系统,其进一步包括接收关于改变所述多个生物序列中对应于所述有害生物序列的至少20%的所述至少两个生物序列以去除所述有害生物序列的指令。本文进一步提供了这样的计算机化系统,其中在一个或多个时间点接收所述多个接收的设计指令。本文进一步提供了这样的计算机化系统,其中所述多个接收的设计指令来自3个或更多个不同的来源。本文进一步提供了这样的计算机化系统,其中所述多个接收的设计指令来自5个或更多个不同的来源。本文进一步提供了这样的计算机化系统,其中所述多个接收的设计指令来自10个或更多个不同的来源。本文进一步提供了这样的计算机化系统,其中所述一个或多个生物序列的长度均不超过200个碱基。本文进一步提供了这样的计算机化系统,其中所述一个或多个生物序列的长度均不超过100个碱基。本文进一步提供了这样的计算机化系统,其中所述一个或多个生物序列的长度均不超过50个碱基。本文进一步提供了这样的计算机化系统,其中所述一个或多个生物序列的长度均不超过20个碱基。

[0004] 本文提供了用于提供增强的多核苷酸合成的方法,其包括:1)接收一个或多个设计指令,其中所述设计指令包含多个生物序列,其中每个所述生物序列的长度不超过500个碱基,并且其中所述多个生物序列包括核酸或氨基酸序列;2)自动确定所述多个生物序列中的至少两个生物序列是否共同对应于数据库中的有害生物序列的至少20%;以及3)如果

检测到所述有害生物序列的至少20%，则自动生成警报。本文进一步提供了这样的方法，其进一步包括其中如果没有生成警报，则合成所述一个或多个序列。本文进一步提供了这样的方法，其进一步包括接收关于改变所述多个生物序列中对应于所述有害生物序列的至少20%的至少两个生物序列以去除有害生物序列的指令。

[0005] 本文提供了用于提供增强的多核苷酸合成的计算机化系统，其包含用于托管数据库的服务器，其中所述数据库适于表示序列列表；网络连接；以及包含用于通用计算机的指令的计算机可读介质，其中所述计算机化系统被配置用于以包括以下步骤的方法操作：1) 接收一个或多个设计指令，其中所述设计指令包含多个生物序列，其中所述多个生物序列是载体序列和多个另外的插入序列；2) 自动确定所述多个插入序列中的至少一个和所述载体是否共同对应于所述数据库中的有害生物序列的至少20%；以及3) 如果检测到所述有害生物序列的至少20%，则自动生成警报。本文进一步提供了这样的计算机化系统，其中所述生物序列获自对实体核酸样品进行测序。本文进一步提供了这样的计算机化系统，其进一步包括其中如果没有生成警报，则合成所述一个或多个生物序列。本文进一步提供了这样的计算机化系统，其进一步包括接收关于改变对应于所述有害生物序列的至少20%的所述多个插入序列中的所述至少一个和所述载体以去除所述有害生物序列的指令。本文进一步提供了用于提供增强的多核苷酸合成的计算机化系统，其中在一个或多个时间点接收所述多个接收的设计指令。本文进一步提供了这样的计算机化系统，其中从不同的来源接收所述多个接收的设计指令。本文进一步提供了这样的计算机化系统，其中所述多个接收的设计指令来自3个或更多个不同的来源。本文进一步提供了这样的计算机化系统，其中所述多个接收的设计指令来自5个或更多个不同的来源。本文进一步提供了这样的计算机化系统，其中所述多个接收的设计指令来自10个或更多个不同的来源。本文进一步提供了这样的计算机化系统，其中所述一个或多个生物序列的长度均不超过200个碱基。本文进一步提供了这样的计算机化系统，其中所述一个或多个生物序列的长度均不超过100个碱基。本文进一步提供了这样的计算机化系统，其中所述一个或多个生物序列的长度均不超过50个碱基。本文进一步提供了这样的计算机化系统，其中所述一个或多个生物序列的长度均不超过20个碱基。

[0006] 本文提供了用于提供增强的多核苷酸合成的方法，其包括：1) 接收一个或多个设计指令，其中所述设计指令包含多个生物序列，其中所述多个生物序列是载体序列以及多个另外的插入序列；2) 自动确定所述多个插入序列中的至少一个和所述载体是否共同对应于数据库中的有害生物序列的至少20%；以及3) 如果检测到所述有害生物序列的至少20%，则自动生成警报。本文进一步提供了这样的方法，其中所述生物序列获自对实体核酸或蛋白质样品进行测序。本文进一步提供了这样的方法，其进一步包括其中如果没有生成警报，则合成一个或多个生物序列。本文进一步提供了这样的方法，其接收关于改变对应于所述有害生物序列的至少20%的所述多个插入序列中的所述至少一个和所述载体以去除所述有害生物序列的指令。

援引并入

[0007] 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并入本文，其程度如同特别地且单独地指出每个单独的出版物、专利或专利申请通过引用以其全文并入。

## 附图说明

[0008] 在所附权利要求书中具体阐述了本公开内容的技术特征。通过参考以下对利用本公开内容原理的说明性实施方案加以阐述的详细描述和附图,将会获得对本公开内容的特征和优点的更好的理解,在这些附图中:

[0009] 图1图示了用户界面,其包括蛋白质序列和相关物种、宿主、病原体、危害途径、后果和蛋白质类型信息。还包括序列登录号、相同蛋白质的列表、指向具有序列记录的数据库的链接以及指向类似蛋白质的链接。

[0010] 图2图示了用户界面,其包括蛋白质变体和示例性蛋白质“血凝素神经氨酸酶-新城疫病毒”的部分列表。

[0011] 图3A描绘了流程图,其包括来自查询文件、蛋白质数据库、blast报告、受限列表(有害序列列表)和筛选报告的信息。

[0012] 图3B描绘了流程图,其包括各种形式的输入(核酸物质、核酸或蛋白质序列)、判定(受限列表、非受限列表、专家审查)和输出(发出警报)。

[0013] 图4图示了用户界面,其包括用于在筛选中搜索的数据库列表。包括角色(role)、类型、名称、描述、添加日期和活动状态列。

[0014] 图5图示了用户界面,其包括序列提交筛选。包括名称、数据库、描述和FASTA文件的表单条目,以及“提交”按钮。数据库表单有下拉列,其在点击子类别时显示,包括“Seqshield”、“nr”和“个人数据库”。

[0015] 图6图示了用户界面,其包括筛选状态的概要。

[0016] 图7图示了用户界面,其包括用于选择所筛选的“未经审查”、“关注”或“未关注”序列的下拉菜单。

[0017] 图8图示了计算机系统。

[0018] 图9图示了计算机系统。

[0019] 图10是图示计算机系统的架构的框图。

[0020] 图11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)。

[0021] 图12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。

## 具体实施方式

[0022] 随着合成生物学中设计能力的快速增长,现在有可能通常使用严重突变的序列来创建大量构建体,该严重突变的序列不与其最初衍生自的参考序列直接相似。同时,在理解致病性背后的过程(在各种宿主和生物情景中)方面的科学进步正迅速产生关于蛋白质序列的新知识,这些蛋白质序列以依赖于情景的方式可能对人、特定植物或动物或更广泛的自然环境造成危害。

[0023] 有道德且负责任的合成生物学家可能在无意中创建出能够造成危害的构建体,但是直到在生命系统中对合成设计进行例示之前无法预测或理解到该能力。由于仅从一级序列来预测功能是不可行的,因此这些科学家将通过获取对以下的访问来得到良好服务:1)关于哪些序列可能导致危害以及调控状态的元数据存储库,以及2)用于针对该元数据检查DNA或蛋白质序列并向用户警告任何潜在问题的有效筛选系统。此外,能够满足这些需求的

筛选系统本身必须能够自动化,以便无缝地适应高通量设计/构建/检验工作流程。本公开内容提供了软件工具以解决关于致病性的公众可获得的基因水平元数据的缺乏以及用于有效筛选的开源工具的缺乏。

[0024] 定义

[0025] 虽然本文已经示出并描述了多个实施方案,但是对于本领域技术人员显而易见的是,这些实施方案仅以示例的方式提供。本领域技术人员在不脱离本文公开的装置设备、系统和方法的情况下可想到多种变化、改变和替换。应当理解,可采用本文所述实施方案的各种替代方案。

[0026] 除非另外定义,否则本文使用的所有技术术语都具有与本公开内容所属领域的普通技术人员通常理解的含义相同的含义。如在本说明书和所附权利要求中所使用的,除非上下文另有明确说明,否则单数形式“一个”、“一种”和“该”包括复数的引用物。除非另有说明,否则本文中对“或”的任何提及旨在包含“和/或”。

[0027] 除非特别说明或从上下文中明显看出,否则如本文所使用的,关于数字或数字范围的术语“约”应理解为所述数字及其数字 $\pm 10\%$ ,或者对于范围列出的值,低于所列下限的 $10\%$ 且高于所列上限的 $10\%$ 。

序列注释

[0028] 关于任何单个序列造成某些类型危害的能力的知识可能极其分散。各个研究者团体关注于致病性的许多不同方面,包括生物体浸润宿主细胞、劫持宿主细胞机器、躲避宿主免疫系统甚至增强宿主免疫应答的能力。示例性的有害生物序列包括编码致病序列的那些序列,如有害并且来自病毒、细菌或寄生虫来源的那些序列。有害生物序列可包括已知具有致病作用的野生型序列的突变形式。有害生物序列包括在转录或翻译后产生有害序列产物的序列,或充当有害序列产物的前体。有害生物序列包括编码有害蛋白质的序列。

[0029] 在其他方面,本公开内容提供了基于Mediawiki的用户界面,该用户界面允许用户提交序列以及基于标签的致病性角色注释。可以鼓励用户为每个序列提交若干个标签,以描述与给定序列相关的危害的大体模式,模型如下:

宿主+环境=后果+关注程度

[0030] 本系统可采用基于标签的方法,以便非先验地施加单个受控词表。由团体注释产生的标签集合可长期形成这种受控词表的基础。

[0031] 在上传每个序列时,可要求用户在四个类别中的每一个中添加标签。标记“宿主”和“关注程度”是强制性的;考虑到所需的额外复杂性和领域知识,为“情景”和“后果”添加标签是可选的。

[0032] 作为实例,编码毒素蓖麻毒蛋白的序列可以被用户标记为:

标签	值
宿主	人
情景	摄取、吸入
后果	发热、咳嗽、呼吸衰竭、死亡
关注程度	极高

[0033] 目标是随着时间的推移使元数据累积超过通用完整性。该系统是集中托管的,并提供整个经审编(curated)序列集(或基于标签查询的子集)以供作为FASTA下载用于筛选。



[0034] 本文提供了用于序列注释的方法,其中数据库接收与生物序列或生物构建体(例如,核苷酸序列或蛋白质序列)相关的特征的列表。示例性特征包括但不限于:核酸序列、蛋白质序列、蛋白质名称、株系来源、指向序列数据库(例如,NCBI)的链接、序列数据库登录号、相同序列(蛋白质或核酸)、相似序列(蛋白质或核酸)、疾病类型(例如,病毒、细菌或真菌)、宿主信息(例如,人、哺乳动物、鸟、昆虫)、有害相互作用的情景或途径(例如,摄取、吸入)和关注程度。本文还提供了一种用户界面,该用户界面呈现每个特征或指向此类特征的附加信息的链接。参见图1。在一些情况下,选择特定毒株的病毒序列。例如,图2图示了用于注释的679种可用的血凝素神经氨酸酶-新城疫病毒毒株的一部分。

[0035] 示例性物种包括动物物种。如本文所用的,“动物”包括但不限于哺乳动物、有袋动物、鸟类、昆虫、节肢动物、两栖动物和爬行动物。示例性哺乳动物包括但不限于绵羊、牛、山羊、猪、兔、野兔、鹿、山羊、小鼠、大鼠、蝙蝠和负鼠等。示例性疾病类型包括来自以下类别的病原体:病毒、细菌、真菌和其他有害病原体。具有有害表达产物的示例性病毒包括但不限于马尔堡病毒、埃博拉病毒、汉坦病毒、禽流感病毒(例如,H5N1毒株)、拉沙病毒、胡宁病毒、克里米亚-刚果热病毒、马秋博病毒、库阿撒鲁尔森林病毒、登革热病毒和屈曲病毒。具有有害表达产物的示例性细菌包括但不限于多重耐药性金黄色葡萄球菌(*Staphylococcus aureus*,MRSA)、大肠杆菌(*E.coli*)、利斯特氏菌(*listeriosis*)、沙门氏菌(*salmonella*)、淋球菌(*gonococcus*)、链球菌(*streptococcus*)和葡萄球菌(*staphylococcus*)。具有有害表达产物的示例性真菌包括但不限于*Amanita arocheae*、双孢鹅膏(*Amanita bisporigera*)、致命鹅膏(*Amanita exitialis*)、*Amanita magnivelaris*、赭鹅膏(*Amanita ocreata*)、春生鹅膏(*Amanita verna*)、白霜杯伞(*Clitocybe dealbata*)、尖顶丝膜菌(*Cortinarius gentilis*)和肉褐环柄菇(*Lepiota brunneoincarnata*)。示例性危害途径包括但不限于摄取、吸入、皮肤接触和性传播。示例性后果包括但不限于发热、头痛、恶心、头晕和腹泻。示例性蛋白质数据库包括美国国家卫生研究院的国家医学图书馆的蛋白质和基因数据库。示例性的疾病关注程度包括低、中、高和极高。

[0036] 本文提供了用于基础审编的方法,如按照生物体名称和/或分类单位(taxon)鉴定与查询相关的序列。一经鉴定,就可任选地更新序列注释,并且任选地针对特定描述性特征重新分类。所鉴定的序列可进一步用于以单个或批量格式下载,任选地采用FASTA格式。

[0037] 数据质量和公众参与都可能是与公共可用数据库相关的问题。为了使即时可用性最大化,公开的系统可进行初始的审编过程,向数据库中添加许多致病蛋白质,以试图包括大多数可能受调控的序列或已知有害的其他序列。该系统可以审编对应于可被认为是无害的基因的NCBI GI标识符的“非受限”列表。该非受限列表也可开放用于审编。

[0038] CAPTCHA的方案可用于防止机器人驱动的审编并且在创建或编辑页面之前需要用户注册。可以定期验证GI标识符(是否存在),并且可以对记录进行标记以供在验证失败时进行人工审查。用户还可以对记录进行标记以请求团体或管理员审查。

[0039] 本公开内容提供了注释和/或筛选至少一种生物序列的系统和方法。在一些情况下,该生物序列是核酸序列。该核酸序列可包含1、10、100、200、300、400、500、600、700、800、900、1000、2000、5000、7000、10000个或更多个核酸残基。在一些情况下,该核酸序列包含100至500个核酸残基。在一些情况下,该核酸序列包含50至1000个核酸残基。在一些情况下,该核酸序列包含20至200个核酸残基。在一些情况下,该核酸序列包含200个残基。在一

些情况下,该生物序列可以是DNA或RNA。在一些情况下,该生物序列是蛋白质序列。该生物序列可包含腺嘌呤(A)、胞嘧啶(C)、鸟嘌呤(G)、胸腺嘧啶(T)或尿嘧啶(U)。在一些情况下,该生物序列是蛋白质序列。该蛋白质可包含1、10、100、200、300、400、500、600、700、800、900、1000、2000个或更多个氨基酸。在一些情况下,该蛋白质序列包含100至300个氨基酸。在一些情况下,该序列包含50至500个氨基酸。在一些情况下,该序列包含10至200个氨基酸。在一些情况下,该序列包含60个氨基酸。在一些情况下,将不超过2、5、10、20、50、100或200个残基的核酸片段经由计算机模拟装配成核酸序列。在一些情况下,核酸片段获自一个或多个来源,或来自相同来源的一种或多种顺序。

#### 筛选工具

[0040] 构建能够确定给定序列是否构成生物安全风险的筛选系统可包括对于所有合成生物学家甚至所有合成生物学公司而言都无法实现的程度的时间投入和专业知识。即使假设可以访问危险序列的数据库,比对器的基本参数化和结果处理(包括剔除相似区域的比对计数以免隐藏与较短区域的同源性)也可能包括领域专业知识。

[0041] 图3A中提供了说明性工作流程。参考图3A,处理器接收包含生物序列信息的查询文件,并且还与具有经鉴定的序列信息的蛋白质数据库通信。生成BLAST报告,其列出与所查询的生物序列部分或全部相关的经鉴定的相同序列和相似序列。然后将BLAST报告针对包含序列注释的数据库(也称为“受限”列表)进行查询,该序列注释鉴定与有害生物序列(蛋白质或核酸)相关的序列。以用户界面的形式生成筛选报告,其概括这些过程的结果。

[0042] 图3B中提供了说明性逻辑工作流程。参考图3B,可以使用搜索一个或多个数据库的算法来评估数据输入源,如实体核酸或蛋白质物质(可对其进行测序)、核酸序列(可将其翻译成蛋白质序列)或蛋白质序列,以确定它是否在受限列表中。示例性算法包括但不限于BLAST、DIAMOND、Smith-Waterman或用于比较序列信息的其他算法。将发现处于受限列表中的序列进一步针对包含已知假阳性的非受限列表进行评估。如果未鉴定出假阳性,则使序列经受专家审查。如果发现序列是无害的,则将其置于非受限列表上以防止进一步将所述序列鉴定为假阳性。如果发现序列是有害的,则生成输出警报。在一些情况下,合成无害序列。在一些情况下,对序列进行修饰以去除有害序列。在一些情况下,重新筛选经修饰的序列。在一些情况下,迭代地重复该过程,直到找到经修饰的无害序列。在一些情况下,合成经修饰的无害序列。

[0043] 参考图4,用户界面显示了可用于筛选过程选择的受限列表。参考图5,说明性用户界面显示了“提交筛选”的提交表单。该表单允许选择针对开放数据库例如公共可用信息的集合进行筛选,或选择针对个人数据库进行筛选,该筛选可以基于非公开可用的选择标准。提交表单还允许选择用于上传的生物序列文件。

[0044] 参考图6,说明性用户界面显示了所进行的生物安全筛选的概要,包括状态信息、筛选的序列、审查状态、关注或不关注状态、序列添加的日期以及指向查看BLAST结果的链接。参考图7,说明性用户界面显示了在筛选期间访问的列表、筛选的序列以及针对序列的有害序列(受限的)指定(assignment)的概要。

[0045] 本文公开的技术可包括筛选系统的基于Python的参考实现。考虑到查询核苷酸序列,该系统可以将序列(例如,通过BLAST)与衍生自通过在前一部分中讨论的界面产生的注释集合的蛋白质序列组进行比较。

[0046] 可以通过同源性程度、E-评分和比对长度来过滤结果。可以通过与这些序列相关的标签以及发现有问题的查询区域的分布来概括传递命中 (passing hit)。可以提供指向原始数据库条目的链接,使得用户可以更详细地跟进。根据预定义的指导,一些实例表明该算法是100%灵敏的,并且可以下载报告以供存档使用。筛选较短 (例如,少于约200个碱基) 序列可导致大量假阳性结果。有效筛选较短的多核苷酸序列可包括算法方法。

[0047] 筛选系统可以位于数据库上,并且包括用于筛选请求提交和结果检索的RESTful应用可编程接口 (API) 以及图形用户界面。该应用可以在膝上型计算机上安装和操作,并且通过API调用可以良好地扩展到高通量使用。

#### 累积生物序列或构建体筛选

[0048] 有可能获得这样的生物序列或构建体片段,该片段在被单独筛选时将不会导致鉴定为有害序列,特别是如果生物序列或构建体是通过多个来源并在多个时间点获得。在一些情况下,该来源可以是客户。例如,任何选择因子调控的细菌或病毒的大部分基因组的累积可以以较小的段获得,然后装配成有害生物序列或构建体。为了解决这个问题,在一些情况下,接收每次请求之后的后台进程,该后台进程在数据库中查询来自该生物序列或构建体请求来源的所有先前订单并收集与任何有害生物序列或构建体具有高度同源性的任何区段的记录。这确保了即使这些区段不足以在单独的订单中触发正式警报或拒绝占有时的评估和警报。在一些情况下,这些高度同源性区段被表示为所关注的选择因子的基因组上的间隔区,然后生成每个生物序列或构建体请求来源和每个基因组的所有间隔区的并集,以确定每个生物序列或构建体请求来源的这些生物体的最大理论构建。在一些情况下,一旦任何生物序列或构建体请求来源寻求设计给定选择因子基因组的20%或更多,就会生成警报以供人工审查并有意地追踪生物序列或构建体请求来源。在一些情况下,一旦任何生物序列或构建体请求来源可生成有害生物序列或构建体的至少5%、10%、20%、30%、40%、50%、60%、70%、80%、90%或超过90%,就会在授权序列构建之前生成警报以供人工审查。在一些情况下,一旦任何生物序列或构建体请求来源可生成有害生物序列或构建体的5%至50%、10%至75%、20%至90%、30%至100%、10%至30%、5%至50%或15%至60%,就会在授权序列构建之前生成警报以供人工审查。

[0049] 为了本文所述的用于核酸设计和/或装配的系统和方法而筛选的生物序列可包含一种或多种核酸或蛋白质序列。对于较短的核酸序列,如包含不超过200个碱基的核酸序列,现有的筛选方法具有非常高的假阳性率。在一些情况下,较短的核酸序列含有不超过2000、1000、500、200、100、75、50、40、30个或不超过20个碱基。在一些情况下,较短的核酸序列含有10至1000个碱基、20至500个碱基、30至300个碱基、40至200个碱基、50至200个碱基、20至200个碱基、10至100个碱基,或100至300个碱基。在一些情况下,核酸序列编码包含不超过300、200、100、75、50、40、30、20、10、5个或不超过5个氨基酸的较短蛋白质。在一些情况下,较短的核酸序列含有10至300个氨基酸、20至200个氨基酸、30至100个氨基酸、10至200个氨基酸、20至100个氨基酸、5至50个氨基酸、10至100个氨基酸或25至75个氨基酸。在一个实例中,采用备选的筛选方法,其在多核苷酸的各组中进行查找以确定生物序列或构建体请求来源何时已经提交了针对足以潜在地装配受调控或有害的生物序列或构建体的多核苷酸的请求。在一些情况下,在订购期间,在一个或多个来源内的后台进程使用装配算法针对选择的有害生物体的基因组跨订单地装配多核苷酸。在一些情况下,装配算法包含下一

代测序装配算法。这些装配体允许将一个或多个订单与一个或多个来源连接的假设生成。例如,将来自来源A和B的订单X、Y和Z组合以装配来自有害生物体的一个或多个基因。在一些情况下,来源的数目为至少2、3、4、5、8、10、15、20、30个或超过30个来源。在一些情况下,来源的数目为2至30个来源、5至50个来源、10至100个来源、5至20个来源、2至10个来源、4至40个来源或15至75来源。在一些情况下,该假设生成了警报以供人工审查,并且任选地触发与生物序列或构建体请求来源的后续讨论或直接向执法部门报告。考虑到与基因长度序列高度同源性的低概率,假阳性率应保持较低。在一些情况下,额外的假阳性减少以评估假设的序列集合的比对结构而确定适当的重叠是否允许装配一个或多个有害生物序列或构建体的形式出现。

[0050] 在一些情况下,实体核酸样品如载体或插入物由用于与待合成的一个或多个核酸序列装配的来源提供。在一些情况下,采用如NGS首先对这些实体核酸材料进行测序,并对一个或多个载体和插入序列的假设装配体进行筛选。在一些情况下,筛选至少两个序列的组合。在一些情况下,筛选至少2、3、4、5、10、15、20、30个或超过30个序列的组合的有害生物序列或构建体。在一些情况下,所筛选的序列的数目为2至30个序列、5至50个序列、10至100个序列、5至20个序列、2至10个序列、4至40个序列或15至75个序列,该筛选是针对有害生物序列或构建体。

#### 数字处理设备

[0051] 在一些实例中,本文所述的平台、系统、介质和方法包括数字处理设备或其使用。在一些实例中,该数字处理设备可包括执行设备功能的一个或多个硬件中央处理单元(CPU)或通用图形处理单元(GPGPU)。在一些实例中,该数字处理设备可进一步包括被配置用于执行可执行指令的操作系统。该数字处理设备可任选地连接到计算机网络。该数字处理设备可任选地连接到因特网,使得其可以访问万维网。该数字处理设备可任选地连接到云计算基础设施。该数字处理设备可任选地连接到内联网。该数字处理设备可任选地连接到数据存储设备。

[0052] 根据本文的描述,作为非限制性实例,合适的数字处理设备可包括服务器计算机、台式计算机、膝上型计算机、笔记本计算机、亚笔记本计算机、上网本计算机、上网平板计算机、机顶盒计算机、流媒体设备、手持计算机、因特网设备、移动智能电话、平板计算机、个人数字助理、视频游戏控制台和载具。许多智能电话可适用于本文所述的系统。具有可选的计算机网络连通性的电视、视频播放器和数字音乐播放器可适用于本文所述的系统。合适的平板计算机可包括具有本领域技术人员已知的小册子、平板和可转换配置的平板计算机。

[0053] 所述数字处理设备可包括被配置用于执行可执行指令的操作系统。例如,该操作系统可以是包括程序和数据的软件,该软件管理设备的硬件并为应用的执行提供服务。作为非限制性实例,合适的服务器操作系统可包括FreeBSD、OpenBSD、NetBSD<sup>®</sup>、Linux、Apple<sup>®</sup> Mac OS XServer<sup>®</sup>、Oracle<sup>®</sup> Solaris<sup>®</sup>、WindowsServer<sup>®</sup>和Novell<sup>®</sup> NetWare<sup>®</sup>。作为非限制性实例,合适的个人计算机操作系统可包括Microsoft<sup>®</sup> Windows<sup>®</sup>、Apple<sup>®</sup> Mac OS X<sup>®</sup>、UNIX<sup>®</sup>和类UNIX操作系统如GNU/Linux<sup>®</sup>。在一些实例中,操作系统可由云计算提供。该设备可包括存储和/或存储器设备。该存储和/或存储器设备可以是用于临时或永久地存储数据或程序的一个或多个物理装置。该设备可以

是易失性存储器,并且可能需要电力来维持所存储的信息。该设备可以是非易失性存储器,并且在数字处理设备未通电时保留所存储的信息。该非易失性存储器可包括闪速存储器、动态随机存取存储器 (DRAM)、铁电随机存取存储器 (FRAM)、相变随机存取存储器 (PRAM)。

[0054] 所述数字处理设备可包括用于向用户发送视觉信息的显示器。该显示器可以是阴极射线管 (CRT)、液晶显示器 (LCD)、薄膜晶体管液晶显示器 (TFT-LCD)、有机发光二极管 (OLED) 显示器、无源矩阵 OLED (PMOLED) 或有源矩阵 OLED (AMOLED) 显示器、等离子体显示器和/或视频投影仪。

[0055] 所述数字处理设备可包括用于从用户接收信息的输入设备。该输入设备可以是键盘。该输入设备可以是定点设备,作为非限制性实例,包括鼠标、追踪球、追踪板、操纵杆、游戏控制器或触笔。该输入设备可以是触摸屏或多点触摸屏。该输入设备可以是用于捕获语音或其他声音输入的麦克风。该输入设备可以是用于捕获运动或视觉输入的摄像机或其他传感器。该输入设备可以是 Kinect、Leap Motion 等。该输入设备可以是诸如本文公开的那些设备的组合。

[0056] 参考图8,在特定实施方案中,示例性数字处理设备801被编程或以其他方式配置用于执行注释或筛选。在该实例中,该数字处理设备801包括中央处理单元 (CPU,本文中也称为“处理器”和“计算机处理器”) 805,其可以是单核或多核处理器,或者用于平行处理的多个处理器。该数字处理设备801还包括存储器或存储位置810 (例如,随机存取存储器、只读存储器、闪速存储器)、电子存储单元815 (例如,硬盘)、用于与一个或多个其他系统通信的通信接口820 (例如,网络适配器) 和外围设备825,如高速缓冲存储器、其他存储器、数据存储和/或电子显示适配器。存储器810、存储单元815、接口820和外围设备825通过诸如主板的通信总线 (实线) 与CPU 805通信。存储单元815可以是用于存储数据的数据存储单元 (或数据存储库)。该数字处理设备801可借助于通信接口820可操作地耦合到计算机网络 (“网络”) 830。网络830可以是因特网、互联网和/或外联网,或与因特网通信的内联网和/或外联网。在一些情况下,网络830是电信和/或数据网络。网络830可包括一个或多个计算机服务器,其可以实现分布式计算,如云计算。在一些情况下借助于设备801,网络830可以实现对等网络,这可以使耦合到设备801的设备能够充当客户端或服务器。

[0057] 继续参考图8,CPU 805可执行一系列机器可读指令,该机器可读指令可体现在程序或软件中。该指令可存储在存储器位置,如存储器810中。该指令可以指向CPU 805,CPU 805可随后编程或以其他方式配置CPU 805以实现本公开内容的方法。由CPU 805执行的操作的实例可包括提取、解码、执行和回写。CPU 805可以是电路如集成电路的一部分。设备801的一个或多个其他组件可包括在电路中。在一些情况下,该电路是专用集成电路 (ASIC) 或现场可编程门阵列 (FPGA)。

[0058] 继续参考图8,存储单元815可以存储文件,如驱动程序、文库和保存的程序。存储单元815可以存储用户数据,例如用户偏好和用户程序。在一些情况下,数字处理设备801可包括一个或多个附加数据存储单元,该附加数据存储单元在外部,如位于通过内联网或因特网进行通信的远程服务器上。

[0059] 继续参考图8,数字处理设备801可通过网络830与一个或多个远程计算机系统通信。例如,设备801可与用户的远程计算机系统通信。远程计算机系统的实例包括个人计算机 (例如,便携式PC)、平板或平板型PC (例如,Apple® iPad、Samsung® Galaxy Tab)、电

话、智能电话(例如, Apple<sup>®</sup>电话、支持Android的设备、Blackberry<sup>®</sup>), 或个人数字助理。

[0060] 如本文所述的方法可通过存储在数字处理设备801的电子存储位置上, 例如, 存储在存储器810或电子存储单元815上的机器(例如, 计算机处理器)可执行代码来实现。可以以软件的形式提供机器可执行代码或机器可读代码。在使用期间, 代码可以由处理器805执行。在一些情况下, 可以从存储单元815检索代码并将其存储在存储器810上以备处理器805获取。在一些情况下, 可以排除电子存储单元815, 并且机器可执行指令存储在存储器810中。

[0061] 其他计算机系统

[0062] 本文所述的任何系统均能够可操作地连接至计算机, 并且可本地或远程地通过计算机进行自动化。在多种情况下, 本公开内容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。相应地, 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。

[0063] 图9中示出的计算机系统900可被理解为能够从介质911和/或网络端口905读取指令的逻辑设备, 其可任选地连接至具有固定介质912的服务器909。诸如图9示出的系统可包括CPU 901、磁盘驱动器903、可选的输入设备如键盘915和/或鼠标916以及可选的监视器907。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内容的数据可经过这样的网络或连接而传输, 以便由图9所示的用户方922接收和/或审阅。

[0064] 图10是示出可与本公开内容的示例实例结合使用的计算机系统1000的第一示例架构的框图。如图10所示, 该示例计算机系统可包括用于处理指令的处理器1002。处理器的非限制性实例包括: Intel Xeon<sup>™</sup>处理器、AMD Opteron<sup>™</sup>处理器、Samsung 32-bit RISC ARM 1176JZ(F)-S v1.0<sup>™</sup>处理器、ARM Cortex-A8 Samsung S5PC100<sup>™</sup>处理器、ARM Cortex-A8 Apple A4<sup>™</sup>处理器、Marvell PXA 930<sup>™</sup>处理器或功能上等效的处理器。多个执行线程可用于平行处理。在一些情况下, 也可以使用多个处理器或具有多核的处理器, 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。

[0065] 如图10所示, 高速缓冲存储器1004可连接至或并入处理器1002, 以提供由处理器1002新近或频繁使用的指令或数据的高速存储器。处理器1002通过处理器总线1008连接至北桥1006。北桥1006通过存储器总线1012连接至随机存取存储器(RAM) 1010, 并管理处理器1002对RAM 1010的访问。北桥1006还通过芯片集总线1016连接至南桥1014。南桥1014又连接至外围总线1018。外围总线可以是例如PCI、PCI-X、PCI Express或其他外围总线。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1018上的外围组件之间的数据传送。在一些供选择的架构中, 北桥的功能性可以并入处理器, 而不是使用单独的北桥芯片。在一些情况下, 系统1000可包括附接至外围总线1018的加速器卡1022。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他硬件。例如, 加速器可用于适应性数据重建或用来评估在扩展集处理中使用的代数表达式。

[0066] 软件和数据存储在外部存储器1024中并可加载至RAM 1010和/或高速缓冲存储器1004中,以供处理器使用。系统1000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、Windows<sup>TM</sup>、MACOS<sup>TM</sup>、BlackBerry OS<sup>TM</sup>、iOS<sup>TM</sup>和其他功能上等价的操作系统,以及在操作系统顶部运行的、用于根据本公开内容的示例实施方案管理数据存储和优化的应用软件。在该实例中,系统1000还包括与外围总线连接的网络接口卡(NIC) 1020和1021,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他计算机系统的网络接口。

[0067] 图11是显示了具有多个计算机系统1102a和1102b、多个蜂窝电话和个人数据助理1102c以及网络附加存储(NAS) 1104a和1104b的网络1100的示图。在示例实例中,系统1102a、1102b和1102c可管理数据存储并优化对存储在网络附加存储(NAS) 1104a和1104b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1102a和1102b和蜂窝电话以及个人数据助理系统1102c的分布式并行处理进行评估。计算机系统1102a和1102b和蜂窝电话以及个人数据助理系统1102c也可提供对存储在网络附加存储(NAS) 1104a和1104b中的数据的适应性数据重建的并行处理。图11仅示出了一个实例,而多种多样的其他计算机架构和系统可与本公开内容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、背板或其他连接器传输数据以便由其他处理器并行处理。在其他实例中,部分或全部的处理器可使用共享的虚拟地址存储空间。

[0068] 图12是根据示例情况使用共享虚拟地址存储空间的多处理器计算机系统1200的框图。该系统包括可访问共享的存储器子系统1204的多个处理器1202a-f。该系统中并入存储器子系统1204中的多个可编程硬件存储算法处理器(MAP) 1206a-f。MAP 1206a-f中的每一个可包括存储器1208a-f和一个或多个现场可编程门阵列(FPGA) 1210a-f。MAP提供了可配置的功能单元,并且可向FPGA 1210a-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可用来评估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。在该实例中,每一个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,每一个MAP可使用直接存储器访问(DMA)以访问相关联的存储器1208a-f,使其独立于且异步于各自的微处理器1202a-f而执行任务。在这一配置中,MAP可将结果直接提供给另一MAP以用于流水处理和并行执行算法。

[0069] 以上计算机架构和系统仅为实例,并且多种多样的其他计算机、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,其包括使用普通处理器、协处理器、FPGA和其他可编程逻辑设备、芯片上系统(SOC)、专用集成电路(ASIC)和其他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,其包括随机存取存储器、硬盘驱动器、闪速存储器、磁带驱动器、磁盘阵列、网络附加存储(NAS)和其他的本地或分布式数据存储设备和系统。

[0070] 在示例实例中,计算机系统可使用在任何上述或其他计算机架构和系统上执行的软件模块来实现。在其他实例中,系统的功能可部分或完全地在固件、可编程逻辑设备如图12所示的现场可编程门阵列(FPGA)、芯片上系统(SOC)、专用集成电路(ASIC)或其他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(例



如图10所示的加速器卡1022)用硬件加速方式实现。

#### 非暂时性计算机可读存储介质

[0071] 本文公开的平台、系统、介质和方法可包括利用程序进行编码的一个或多个非暂时性计算机可读存储介质,该程序包括可由任选地联网的数字处理设备的操作系统执行的指令。计算机可读存储介质可以是数字处理设备的有形组件。计算机可读存储介质任选地可从数字处理设备移除。作为非限制性实例,计算机可读存储介质包括CD-ROM、DVD、闪存存储器设备、固态存储器、磁盘驱动器、磁带驱动器、光盘驱动器、云计算系统和服务,等等。在一些情况下,该程序和指令在介质上永久地、基本上永久地、半永久地或非暂时地编码。

#### 计算机程序

[0072] 在一些实施方案中,本文公开的平台、系统、介质和方法可包括至少一个计算机程序或其使用。计算机程序包括可在数字处理设备的CPU中执行的一系列指令,该指令被编写用于执行指定的任务。计算机可读指令可被实现为执行特定任务或实现特定抽象数据类型的程序模块,如函数、对象、应用编程接口(API)、数据结构等。根据本文提供的公开内容,计算机程序可以用各种语言的各种版本来编写。

#### 网络应用

[0073] 计算机程序可包括网络应用。根据本文提供的公开内容,网络应用可利用一个或多个软件框架以及一个或多个数据库系统。可在诸如**Microsoft® .NET**或**Ruby on Rails (RoR)**等软件框架上创建网络应用。网络应用可利用一个或多个数据库系统,作为非限制性实例,该数据库系统包括关系型数据库系统、非关系型数据库系统、面向对象数据库系统、关联数据库系统和XML数据库系统。在进一步的实施方案中,作为非限制性实例,合适的关系型数据库系统包括**Microsoft® SQL**服务器、**MySQL™**和**Oracle®**。本领域技术人员还将认识到,在各个实施方案中,网络应用以一种或多种语言的一个或多个版本编写。网络应用可以用一种或多种标记语言、表示定义语言、客户端脚本语言、服务器端编码语言、数据库查询语言或其组合来编写。在一些实施例中,网络应用在某种程度上以诸如超文本标记语言(HTML)、可扩展超文本标记语言(XHTML)或可扩展标记语言(XML)等标记语言编写。网络应用可在某种程度上以诸如级联样式表(CSS)等表示定义语言编写。网络应用可在某种程度上以诸如异步Javascript和XML(AJAX)、**Flash®** Actionscript、Javascript或**Silverlight®**等客户端脚本语言编写。网络应用可在某种程度上以诸如活动服务器网页(ASP)、**ColdFusion®**、Perl、Java™、JavaServer Pages(JSP)、超文本预处理器(PHP)、Python™、Ruby、Tcl、Smalltalk、**WebDNA®**或Groovy等服务器端编码语言编写。网络应用可在某种程度上以诸如结构化查询语言(SQL)等数据库查询语言编写。

#### 移动应用

[0074] 计算机程序可包括被提供至移动数字处理设备的移动应用。该移动应用可在其制造时被提供至移动数字处理设备。可以经由本文所述的计算机网络将移动应用提供至移动数字处理设备。

[0075] 例如,可以使用硬件、语言和开发环境来创建移动应用。移动应用可以以多种编程语言编写。作为非限制性实例,合适的编程语言包括C、C++、C#、Objective-C、Java™、



Javascript、Pascal、Object Pascal、Python<sup>TM</sup>、Ruby、VB.NET、WML以及具有或不具有CSS的 XHTML/HTML或其组合。

[0076] 合适的移动应用开发环境可从数个来源获得。作为非限制性实例，市售的开发环境包括AirplaySDK、alcheMo、Appcelerator<sup>®</sup>、Celsius、Bedrock、Flash Lite、.NET Compact Framework、Rhomobile和WorkLight移动平台。其他开发环境可免费获得，作为非限制性实例，包括Lazarus、MobiFlex、MoSync和Phonegap。此外，移动设备制造商分发软件开发工具包，作为非限制性实例，包括iPhone和iPad (iOS) SDK、Android<sup>TM</sup> SDK、BlackBerry<sup>®</sup> SDK、BREW SDK、Palm<sup>®</sup> OS SDK、Symbian SDK、webOS SDK和Windows<sup>®</sup> Mobile SDK。

#### 独立应用

[0077] 计算机程序可包括独立应用，该独立应用是作为独立的计算机过程运行的程序，而非现有过程的附加项，例如，不是插件。可编译独立应用。编译器是将以编程语言编写的源代码转换为二进制目标代码如汇编语言或机器代码的计算机程序。作为非限制性实例，合适的编译编程语言包括C、C++、Objective-C、COBOL、Delphi、Eiffel、Java<sup>TM</sup>、Lisp、Python<sup>TM</sup>、Visual Basic以及VB.NET或其组合。通常至少部分地执行编译以创建可执行程序。

#### 网络浏览器插件

[0078] 所述计算机程序可包括网络浏览器插件。在计算中，插件可以是向较大的软件应用添加特定功能的一个或多个软件组件。软件应用的制造者支持插件，以使第三方开发人员能够创建扩展应用的能力，能够支持轻易地添加新特征，并且能够减小应用的大小。插件当受支持时可以能够对软件应用的功能进行定制。例如，插件通常用于网络浏览器中，以播放视频、产生互动、扫描病毒以及显示特定文件类型。网络浏览器插件包括但不限于Adobe<sup>®</sup> Flash<sup>®</sup> Player、Microsoft<sup>®</sup> Silverlight<sup>®</sup>和Apple<sup>®</sup> QuickTime<sup>®</sup>。工具栏可包括一个或多个网络浏览器扩展项、加载项或附加项。在一些实施方案中，工具栏包含一个或多个浏览器栏、工具栏或桌面栏。

[0079] 若干插件框架是可用的，其可以能够以多种编程语言开发插件，包括，作为非限制性实例，这些编程语言包括C++、Delphi、Java<sup>TM</sup>、PHP、Python<sup>TM</sup>和VB.NET或其组合。

[0080] 网络浏览器（也称为因特网浏览器）是可被配置用于与联网的数字处理设备一起使用的软件应用，用于检索、呈现和遍历万维网上的信息资源。作为非限制性实例，合适的网络浏览器包括Microsoft<sup>®</sup> Internet Explorer<sup>®</sup>、Mozilla<sup>®</sup> Firefox<sup>®</sup>、Google<sup>®</sup> Chrome、Apple<sup>®</sup> Safari<sup>®</sup>、Opera Software<sup>®</sup> Opera<sup>®</sup>和KDE Konqueror。在一些实施方案中，网络浏览器是移动网络浏览器。移动网络浏览器（也称为微浏览器、迷你浏览器和无线浏览器）可被配置用于在移动数字处理设备上使用，作为非限制性实例，该移动数字处理设备包括手持计算机、平板计算机、上网本计算机、亚笔记本计算机、智能电话、音乐播放器、个人数字助理（PDA）和手持视频游戏系统。作为非限制性实例，合适的移动网络浏览器包括：Google<sup>®</sup> Android<sup>®</sup> 浏览器、RIM BlackBerry<sup>®</sup> 浏览器、Apple<sup>®</sup> Safari<sup>®</sup>、Palm<sup>®</sup>

Blazer、Palm<sup>®</sup> WebOS<sup>®</sup> 浏览器、适用于移动设备的Mozilla<sup>®</sup> Firefox<sup>®</sup>、Microsoft<sup>®</sup> Internet Explorer<sup>®</sup> Mobile、Amazon<sup>®</sup> Kindle<sup>®</sup> Basic Web、Nokia<sup>®</sup> 浏览器、Opera Software<sup>®</sup> Opera<sup>®</sup> Mobile和Sony<sup>®</sup> PSP™浏览器。

#### 软件模块

[0081] 本文所述的系统、介质、网络和方法可包括软件、服务器和/或数据库模块或其使用。可以使用各种机器、软件和编程语言来创建软件模块。本文公开的软件模块以多种方式实现。软件模块可包含文件、代码段、编程对象、编程结构或其组合。软件模块可包含多个文件、多个代码段、多个编程对象、多个编程结构或其组合。作为非限制性实例，一个或多个软件模块包括网络应用、移动应用和独立应用。在一些实施方案中，软件模块位于一个计算机程序或应用中。软件模块可位于超过一个计算机程序或应用中。软件模块可被托管在一台机器上。软件模块可被托管在超过一台机器上。软件模块可被托管在云计算平台上。软件模块可被托管在一个位置处的一台或多台机器上。软件模块可被托管在超过一个位置处的一台或多台机器上。

#### 数据库

[0082] 本文公开的平台、系统、介质和方法可包括一个或多个数据库或其使用。根据本文提供的公开内容，许多数据库适用于存储和检索生理数据。在各个实施方案中，作为非限制性实例，合适的数据库包括关系型数据库、非关系型数据库、面向对象的数据库、对象数据库、实体关系模型数据库、关联数据库和XML数据库。进一步的非限制性实例包括SQL、PostgreSQL、MySQL、Oracle、DB2和Sybase。在一些实施方案中，数据库是基于互联网的。数据库是基于网络的。数据库可以是基于云计算的。数据库可以基于一个或多个本地计算机存储设备。

[0083] 提出以下实施例是为了更清楚地说明本文公开的实施方案的原理和实践，但不应解释为限制任何要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。

#### 算法

[0084] 本文公开的平台、系统、介质和方法可包括一个或多个算法或其使用。根据本文提供的公开内容，许多算法适用于搜索和比较序列数据。在各个实施方案中，作为非限制性实例，合适的算法包括BLAST、DIAMOND、BLAT、BWT、PLAST、Smith-Waterman或用于序列搜索和比对的其他算法。算法可包括现有算法的加速或扩展版本，或者使用这些算法的软件工具。在一些情况下，作为非限制性实例，合适的加速或扩展算法和软件工具包括CS-BLAST、Tera-BLAST、GPU-Blast、G-BLASTN、MPIBLAST、Paracel BLAST、CaBLAST或任何其他加速BLAST算法的算法或软件工具。

[0085] 本文提供了用于设计和合成具有增强的生物安全性和生物安全的生物序列或构建体的系统和方法。在一些情况下，生物安全性是指增强个体的安全性，例如，该增强是通过旨在防止与在制造期间或由制造产生的有害生物制剂接触的预防措施。在一些情况下，生物安全是指保护人群的安全，例如，通过旨在防止有害生物制剂的使用或传播的预防措施。在一些情况下，接收包含一个或多个生物序列的一个或多个生物构建体，使用数据库针对生物安全风险进行筛选，并且如果一个或多个生物序列或构建体被确定为有害表达构建

体或有害产物,则生成警报。在一些情况下,生物序列或构建体是指合成序列。在一些情况下,生物序列或构建体是指天然存在的序列。在一些情况下,生物序列或构建体包括核酸或氨基酸。在一些情况下,生物序列是指合成序列。在一些情况下,生物序列是指天然存在的序列。在一些情况下,生物序列包含核酸或氨基酸。在一些情况下,使用用户注释来提供关于数据库中的生物序列或构建体的性质的额外信息。在一些情况下,所述方法和系统适合于自动化,以便无缝地适应高通量设计/构建/检验工作流程。在一些情况下,筛选生物构建体包括将在多个时间点从单个或多个来源获得的较小生物序列的组合进行比较。在一些情况下,人类专家进一步评估被确定为有害的生物序列或构建体以减少未来的假阳性。在一些情况下,这些系统和方法包括与用户和数据库接口的计算机、软件应用和网络。

[0086] 本文提供了这样的系统,其包含:处理器和存储器;用于评估生物构建体的生物安全的机器指令,该机器指令包括:与生物构建体相关的多个标签的数据库;注释工具;以及可选的筛选工具。本文进一步提供了这样的系统,其中所述生物序列或构建体包含一个或多个生物序列。本文进一步提供了这样的系统,其中所述生物序列是核酸序列。本文进一步提供了这样的系统,其中所述生物序列是蛋白质序列。本文进一步提供了这样的系统,其中所述注释工具被配置为允许用户提供生物构建体的序列的一个或多个带注释的标签。本文进一步提供了这样的系统,其中所述一个或多个带注释的标签至少包括宿主和关注程度。本文进一步提供了这样的系统,其中所述一个或多个带注释的标签包括后果。本文进一步提供了这样的系统,其中所述后果包括疾病。本文进一步提供了这样的系统,其中所述一个或多个带注释的标签包括情景。本文进一步提供了这样的系统,其中所述一个或多个带注释的标签包括致病性。本文进一步提供了这样的系统,其中所述一个或多个带注释的标签包括危害。本文进一步提供了这样的系统,其中所述一个或多个带注释的标签基于一个或多个术语。本文进一步提供了这样的系统,其中所述一个或多个带注释的标签基于一个或多个语句描述。本文进一步提供了这样的系统,其中所述注释工具被进一步配置用于生成一个或多个带注释的标签的受控词表。本文进一步提供了这样的系统,其中所述注释工具包括审编过程。本文进一步提供了这样的系统,其中所述审编过程包括将来自外部数据库的关于生物序列或构建体的信息整合到数据库中。本文进一步提供了这样的系统,其中所述审编过程包括确定生物构建体的无害特征。本文进一步提供了这样的系统,其中所述注释工具包括将序列与数据库中的生物序列或构建体的序列进行比对。本文进一步提供了这样的系统,其中所述筛选工具被配置为允许用户搜索生物构建体的给定序列的生物安全风险。本文进一步提供了这样的系统,其中所述给定序列包含核苷酸序列。本文进一步提供了这样的系统,其中所述给定序列包含蛋白质序列。本文进一步提供了这样的系统,其中所述筛选工具包括将给定序列与数据库中生物序列或构建体的序列进行比对的序列比对器。本文进一步提供了这样的系统,其中所述搜索生物安全风险包括通过一定程度的同源性进行过滤。本文进一步提供了这样的系统,其中所述搜索生物安全风险包括评估序列比对长度。本文进一步提供了这样的系统,其中所述搜索生物安全风险包括生成评估分数。本文进一步提供了这样的系统,其中所述筛选工具进一步包括应用可编程接口。本文还提供了这样的系统,其中所述机器指令进一步包括用于注释和筛选的图形用户界面。

[0087] 本文提供了用于评估生物安全风险的计算机实现的方法,该方法包括:通过处理器使用数据库来存储与生物构建体相关的多个标签;通过处理器使用注释工具来注释生物

构建体的特征;以及任选地,通过处理器使用筛选工具来搜索生物构建体的特征。本文进一步提供了这样的方法,其中所述生物构建体包含生物序列。本文进一步提供了这样的方法,其中所述生物序列是核酸序列。本文进一步提供了这样的方法,其中所述生物序列是蛋白质序列。本文进一步提供了这样的方法,其中所述注释工具被配置为允许用户提供生物构建体的序列的一个或多个带注释的标签。本文进一步提供了这样的方法,其中所述一个或多个带注释的标签至少包括宿主和关注程度。本文进一步提供了这样的方法,其中所述一个或多个带注释的标签包括后果。本文进一步提供了这样的方法,其中所述后果包括疾病。本文进一步提供了这样的方法,其中所述一个或多个带注释的标签包括情景。本文进一步提供了这样的方法,其中所述一个或多个带注释的标签包括致病性。本文进一步提供了这样的方法,其中所述一个或多个带注释的标签包括危害。本文进一步提供了这样的方法,其中所述一个或多个带注释的标签基于一个或多个术语。本文进一步提供了这样的方法,其中所述一个或多个带注释的标签基于一个或多个语句描述。本文进一步提供了这样的方法,其中所述注释工具被进一步配置用于生成一个或多个带注释的标签的受控词表。本文进一步提供了这样的方法,其中所述注释工具包括审编过程。本文进一步提供了这样的方法,其中所述审编过程包括将来自外部数据库的关于生物序列或构建体的信息整合到数据库中。本文进一步提供了这样的方法,其中所述审编过程包括确定生物构建体的无害特征。本文进一步提供了这样的方法,其中所述注释工具包括将序列与数据库中的生物构建体的序列进行比对。本文进一步提供了这样的方法,其中所述筛选工具被配置为允许用户搜索生物构建体的给定序列的生物安全风险。本文进一步提供了这样的方法,其中所述给定序列包含核苷酸序列。本文进一步提供了这样的方法,其中所述给定序列包含蛋白质序列。本文进一步提供了这样的方法,其中所述筛选工具包括将给定序列与数据库中生物构建体的序列进行比对的序列比对器。本文进一步提供了这样的方法,其中所述搜索生物安全风险包括通过一定程度的同源性进行过滤。本文进一步提供了这样的方法,其中所述搜索生物安全风险包括评估序列比对长度。本文进一步提供了这样的方法,其中所述搜索生物安全风险包括生成评估分数。本文进一步提供了这样的方法,其中所述筛选工具进一步包括应用可编程接口。本文进一步提供了这样的方法,其中所述机器指令进一步包括用于注释和筛选的图形用户界面。

[0088] 本文提供了用于评估生物安全风险的计算机实现的方法,该方法包括:通过处理器访问数据库以存储与生物构建体相关的多个标签;通过处理器评估筛选工具以搜索生物构建体的特征;以及通过处理器传输报告工具以发送筛选工具的搜索结果。本文进一步提供了这样的方法,其中所述生物构建体包含生物序列。本文进一步提供了这样的方法,其中所述生物序列是核酸序列。本文进一步提供了这样的方法,其中所述生物序列是蛋白质序列。本文进一步提供了这样的方法,其进一步包括被配置为允许用户提供生物构建体的序列的一个或多个带注释的标签的注释工具。本文进一步提供了这样的方法,其中所述一个或多个带注释的标签至少包括宿主和关注程度。本文进一步提供了这样的方法,其中所述一个或多个带注释的标签包括后果。本文进一步提供了这样的方法,其中所述后果包括疾病。本文进一步提供了这样的方法,其中所述一个或多个带注释的标签包括情景。本文进一步提供了这样的方法,其中所述一个或多个带注释的标签包括致病性。本文进一步提供了这样的方法,其中所述一个或多个带注释的标签包括危害程度。本文进一步提供了这样的

方法,其中所述一个或多个带注释的标签基于一个或多个术语。本文进一步提供了这样的方法,其中所述一个或多个带注释的标签基于一个或多个语句描述。本文进一步提供了这样的方法,其中所述注释工具进一步被配置用于生成一个或多个带注释的标签的受控词表。本文进一步提供了这样的方法,其中所述注释工具包括审编过程。本文进一步提供了这样的方法,其中所述审编过程包括将来自外部数据库的关于生物序列或构建体的信息整合到数据库中。本文进一步提供了这样的方法,其中所述审编过程包括确定生物构建体的无害特征。本文进一步提供了这样的方法,其中所述注释工具包括将序列与数据库中的生物构建体的序列进行比对。本文进一步提供了这样的方法,其中所述筛选工具被配置为允许用户搜索生物构建体的给定序列的生物安全风险。本文进一步提供了这样的方法,其中所述给定序列包含核苷酸序列。本文进一步提供了这样的方法,其中所述给定序列包含蛋白质序列。本文进一步提供了这样的方法,其中所述筛选工具包括使给定序列与数据库中生物构建体的序列进行比对的序列比对。本文进一步提供了这样的方法,其中所述搜索生物安全风险包括通过一定程度的同源性进行过滤。本文进一步提供了这样的方法,其中所述搜索生物安全风险包括评估序列比对长度。本文进一步提供了这样的方法,其中所述搜索生物安全风险包括生成评估分数。本文进一步提供了这样的方法,其中所述筛选工具进一步包含应用可编程接口。本文进一步提供了这样的方法,其进一步包括传输针对用于注释的图形用户界面的机器指令。本文进一步提供了这样的方法,其中进一步包括传输针对用于筛选的图形用户界面的机器指令。本文进一步提供了这样的方法,其进一步包括传输针对用于报告的图形用户界面的机器指令。本文进一步提供了这样的方法,其中所述生物构建体包含与有害表达产物(例如,由翻译产生的蛋白质)或有害产物(例如,由转录产生的RNA)相关的生物序列。本文进一步提供了这样的方法,其中所述生物序列是病毒、细菌或真菌。本文进一步提供了这样的方法,其进一步包括用于访问数据库以存储与生物构建体相关的多个标签的接收到的机器指令。本文进一步提供了这样的方法,其中所述机器指令包括与生物构建体相关的信息。本文进一步提供了这样的方法,其中所述与生物序列或构建体相关的信息包含核酸序列或蛋白质序列。本文进一步提供了这样的方法,其中所述与生物序列或构建体相关的信息包含数据库登录号。

[0089] 应当理解,可以单独地、共同地或彼此组合地理解本公开内容的不同方面。本文描述的本公开内容的各个方面可应用于下面阐述的任何特定应用。通过查看说明书、权利要求和附图,本公开内容的其他目的和特征将变得显而易见。

#### 实施例

##### [0090] 实施例1:序列注释

[0091] 由处理器单元接收生物序列。在该实施例中,生物序列是蛋白质序列。处理器单元访问蛋白质数据库并鉴定与所接收的蛋白质序列相匹配的蛋白质序列。处理器单元接收与蛋白质序列的各种特征相关的信息。特征包括:与蛋白质序列相关的核酸序列、蛋白质序列、蛋白质名称、株系来源信息、指向序列数据库(例如,NCBI)的链接、序列数据库登录号、相同序列(蛋白质或核酸)、相似序列(蛋白质或核酸)、疾病来源(例如,病毒、细菌)、生物体的分类学描述(例如,界、门、纲、目、科、属、种)、宿主信息(例如,人、哺乳动物、鸟类、昆虫)、有害相互作用的情景或途径(例如,摄取、吸入)、症状和关注程度。在该实施例中,获取的蛋白质是新城疫病毒-3。提供了用于注释的特征的示例性用户界面在图1中提供。当由处理器

接收具有与生物序列相关的特征的信息的机器指令时,更新与生物序列相关的标签信息。例如,参考图1,新城疫病毒-3具有蛋白质序列、相同蛋白质(AHL4519.1.1和AHL45193.1)、宿主类型(鸟)、有害相互作用途径(吸入)和症状(呼吸衰竭)的标签信息。

[0092] 当处理器单元接收到对“血凝素神经氨酸酶-新城疫病毒”家族的选择时,访问病毒毒株信息的列表,并且任选地,与使用户界面显示该毒株的机器指令一起传输。参见例如图2,其中提供了679种可用的血凝素神经氨酸酶-新城疫病毒毒株的部分列表以供注释。

[0093] 在一些情况下,还使用与本说明书一致的附加标签信息,包括但不限于FSAP控制或出口控制。

[0094] 实施例2:序列筛选

[0095] 参考图3A,处理器接收呈包含生物序列信息的查询文件形式的机器指令,在本情况下该生物序列信息是核酸信息。处理器还与核酸和蛋白质数据库通信。处理器访问核酸和蛋白质数据库。生成BLAST处理的报告,其列出了经鉴定与所查询的生物序列部分或全部相关的相同和相似序列。然后将来自BLAST处理的报告的序列针对包含序列注释的数据库进行查询,该序列注释鉴定出与也被称为“受限”列表的有害生物序列(蛋白质或核酸)相关的序列。以概括这些过程的结果的用户界面的形式生成筛选报告。筛选报告以针对用户界面的机器指令的形式传输。处理器接收针对数据库的特定指令以访问受限列表信息。参见图4。受限列表可以通过互联网打开,或是关闭的并且只能经过授权访问。还生成了包括生物序列筛选的概况的筛选报告。进行了5次筛选。参见图6。还生成了包括“受限指定”的列表的筛选报告,鉴定出有害生物序列。参见图7。筛选报告鉴定了Gcra细胞周期调控家族-猪布鲁氏菌2型蛋白质。

[0096] 实施例3:针对特定基因组的预筛选

[0097] 对重型天花(*Variola major*)或轻型天花(*Variola Minor*)的基因组中超过500个核苷酸的访问受到世界卫生组织(WHO)政策的限制。需要更长序列的人必须在合成之前进行申请并获得WHO的许可。由于天花的独特性质,仅对重型天花和轻型天花以及痘苗和其他密切相关的正痘病毒的基因组进行预筛选。使用实施例2的一般生物安全筛选程序和正痘病毒的基因组评估核酸序列。该筛选在短于1秒内进行(通过商品硬件上的blastx)。包括痘苗和其他正痘参考序列以确保在发出警报之前所请求序列与天花的同源性最大(类似于2010HHS指南‘最佳匹配’标准)。这可以任选地在订单报价生成过程期间进行,其中如果检测到有害序列,则在开始制造之前生成警报以供人工审查。

[0098] 实施例4:文库模板筛选

[0099] 选择编码约200个氨基酸的基因的约600个核苷酸的基因长度核酸序列以供产生变体文库。获得序列并将其提交至实施例2的一般生物安全筛选程序,以确保变体文库不含有害序列。该程序旨在当检测到有害序列时生成警报以供人工审查。

[0100] 实施例5:定制核酸筛选

[0101] 获得含有实体核酸的物质,如载体,并通过下一代测序(NGS)进行测序。将从NGS获得的共有序列数据提交至实施例2的一般生物安全筛选程序。这确保了核酸材料不构成生物安全或生物安全性问题,如在远离预期使用的插入位点的载体骨架中编码毒素的表达,使得转化到大肠杆菌中将导致有害物剂如毒素的表达。该程序旨在当检测到有害序列时生成警报以供人工审查。

[0102] 实施例6:在相同的查询内,相对于选择因子基因组的订单间装配

[0103] 为了管理请求者(生物序列或构建体请求来源,如客户)可能随着时间的推移和跨越各个订单积累任何选择因子调控的细菌或病毒的大部分基因组的风险,在每个请求之后,后台进程在数据库中查询来自该请求者的所有先前订单,并使用实施例2的一般方法收集与任何选择因子细菌或病毒具有高度同源性的任何区段的记录。这确保了即使这些区域不足以在单个订单中触发正式警报或拒绝占有时也可作出评估和发出警报。这些高度同源性区段表示为所关注的选择因子的基因组上的间隔区,然后生成每个请求者和每个基因组的所有间隔区的并集,以确定每个请求者的这些生物体的最大理论构建。一旦任何请求者可以生成给定选择因子基因组的20%或更多,就会生成警报以供人工审查并有意地追踪该请求者。

[0104] 实施例7:用于假设生成的相对于选择因子基因组的多核苷酸池装配

[0105] 对于较短的多核苷酸序列,如含有不超过200个碱基的多核苷酸序列,现有的筛选方法具有非常高的假阳性率。采用备选的筛选方法,其在各组多核苷酸中进行查找以确定请求者(生物序列或构建体请求来源,即客户)何时已经订购了足以潜在地装配受调控或有害的序列的多核苷酸。在订购期间,在一个或多个请求来源内的后台进程使用来自NGS的装配算法针对选择因子细菌和病毒的基因组跨订单地装配多核苷酸。这些装配体允许假设生成,例如,“如果将来自请求者A和B的订单X、Y和Z组合,则可完全装配来自天花的三个基因”。这些假设生成警报以供人工审查,并且任选地触发与请求者的后续讨论或直接向执法部门报告。考虑到与基因长度序列高度同源性的低概率,假阳性率应保持较低;额外的假阳性减少以评估假设的多核苷酸集合的比对结构而确定是否存在允许容易装配的适当重叠(即,它看起来是有意设计的)的形式出现。

[0106] 实施例8:机器学习引导的风险注释

[0107] 筛选平台和人工审查建立较大的非受限列表和一组真阳性警报案例,其中生物序列或构建体请求来源被确认为正订购所关注的受限序列。机器学习算法在序列本身(例如,隐马尔可夫模型(HMM)类型的情景感知状态模型)和/或GenBank记录注释(例如,自然语言处理(NLP)类型的模型,用于基于共享语言和含义以及先前列出非受限序列的记录来估计未来非受限序列指定的概率)上进行训练。

[0108] 虽然本文已经示出并描述了本公开内容的优选实施方案,但对于本领域技术人员显而易见的是,这些实施方案仅以示例的方式提供。本领域技术人员在不脱离本公开内容的情况下现将会想到多种变化、改变和替换。应当理解,本文中所述的本公开内容实施方案的各种替代方案可用于实施本公开内容。



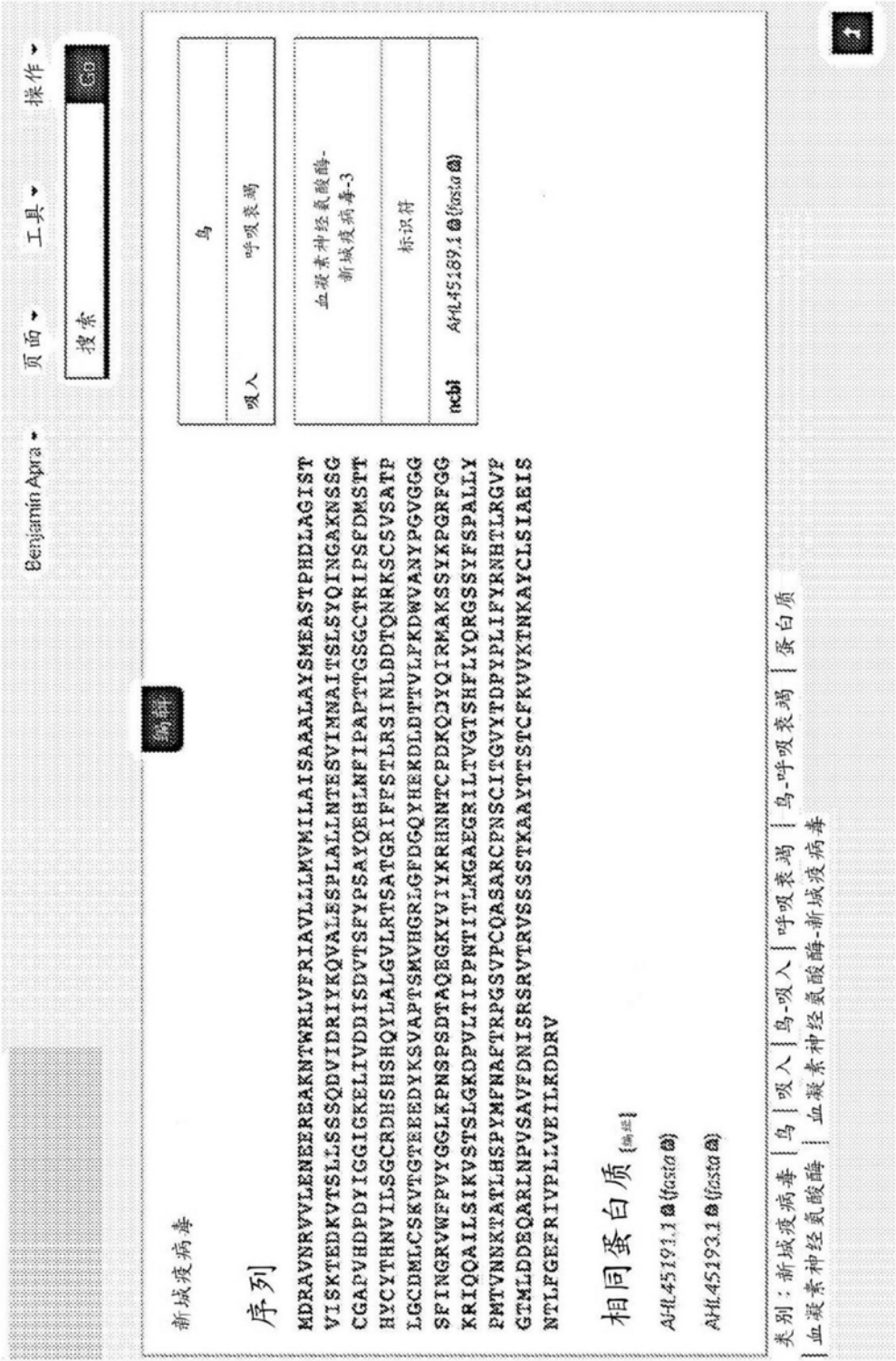


图1



## 类别“血凝素神经氨酸酶-新城疫病毒”中的页面

以下200页属于此类别，共679页。

(上一页) (下一页)

## H

- 血凝素神经氨酸酶-新城疫病毒-1
- 血凝素神经氨酸酶-新城疫病毒-10
- 血凝素神经氨酸酶-新城疫病毒-100
- 血凝素神经氨酸酶-新城疫病毒-101
- 血凝素神经氨酸酶-新城疫病毒-102
- 血凝素神经氨酸酶-新城疫病毒-103
- 血凝素神经氨酸酶-新城疫病毒-104
- 血凝素神经氨酸酶-新城疫病毒-105
- 血凝素神经氨酸酶-新城疫病毒-106
- 血凝素神经氨酸酶-新城疫病毒-107
- 血凝素神经氨酸酶-新城疫病毒-108
- 血凝素神经氨酸酶-新城疫病毒-109
- 血凝素神经氨酸酶-新城疫病毒-11
- 血凝素神经氨酸酶-新城疫病毒

- 血凝素神经氨酸酶-新城疫病毒-159
- 血凝素神经氨酸酶-新城疫病毒-16
- 血凝素神经氨酸酶-新城疫病毒-160
- 血凝素神经氨酸酶-新城疫病毒-161
- 血凝素神经氨酸酶-新城疫病毒-162
- 血凝素神经氨酸酶-新城疫病毒-163
- 血凝素神经氨酸酶-新城疫病毒-164
- 血凝素神经氨酸酶-新城疫病毒-165
- 血凝素神经氨酸酶-新城疫病毒-166
- 血凝素神经氨酸酶-新城疫病毒-167
- 血凝素神经氨酸酶-新城疫病毒-168
- 血凝素神经氨酸酶-新城疫病毒-169
- 血凝素神经氨酸酶-新城疫病毒-17
- 血凝素神经氨酸酶-新城疫病毒-170
- 血凝素神经氨酸酶-新城疫病毒-171
- 血凝素神经氨酸酶-新城疫病毒-219
- 血凝素神经氨酸酶-新城疫病毒-22
- 血凝素神经氨酸酶-新城疫病毒-220
- 血凝素神经氨酸酶-新城疫病毒-221
- 血凝素神经氨酸酶-新城疫病毒-222
- 血凝素神经氨酸酶-新城疫病毒-223
- 血凝素神经氨酸酶-新城疫病毒-224
- 血凝素神经氨酸酶-新城疫病毒-225
- 血凝素神经氨酸酶-新城疫病毒-226
- 血凝素神经氨酸酶-新城疫病毒-227
- 血凝素神经氨酸酶-新城疫病毒-228
- 血凝素神经氨酸酶-新城疫病毒-229
- 血凝素神经氨酸酶-新城疫病毒-23
- 血凝素神经氨酸酶-新城疫病毒-230
- 血凝素神经氨酸酶-新城疫病毒-231

图2

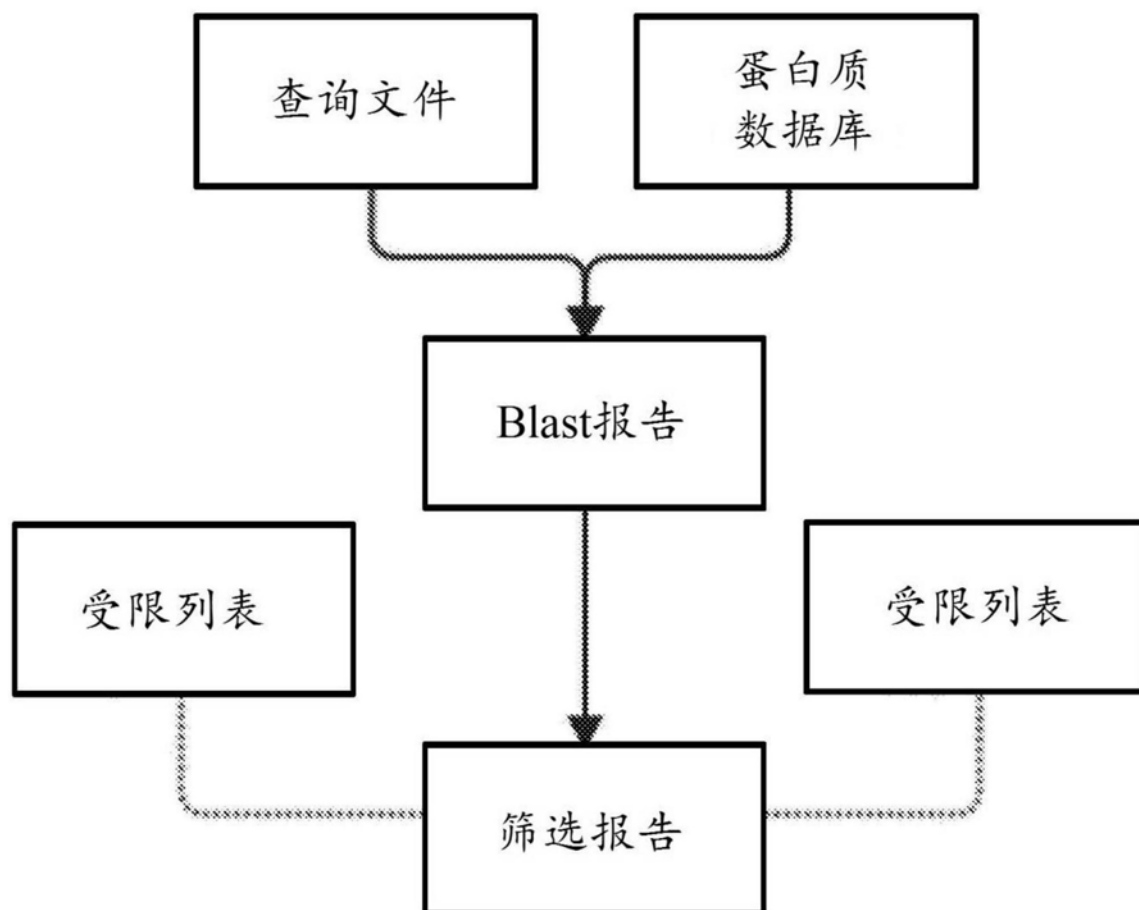


图3A

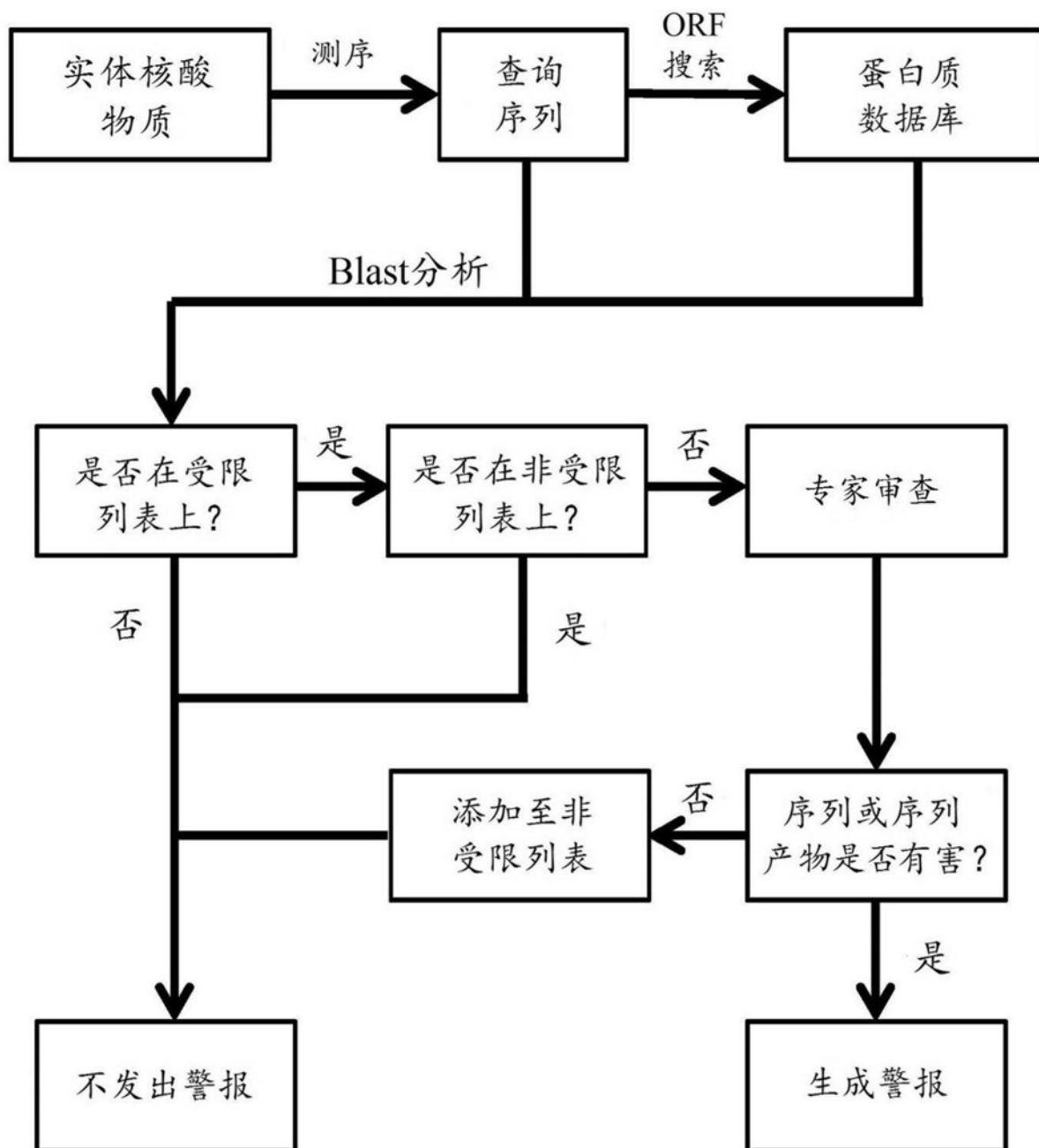


图3B

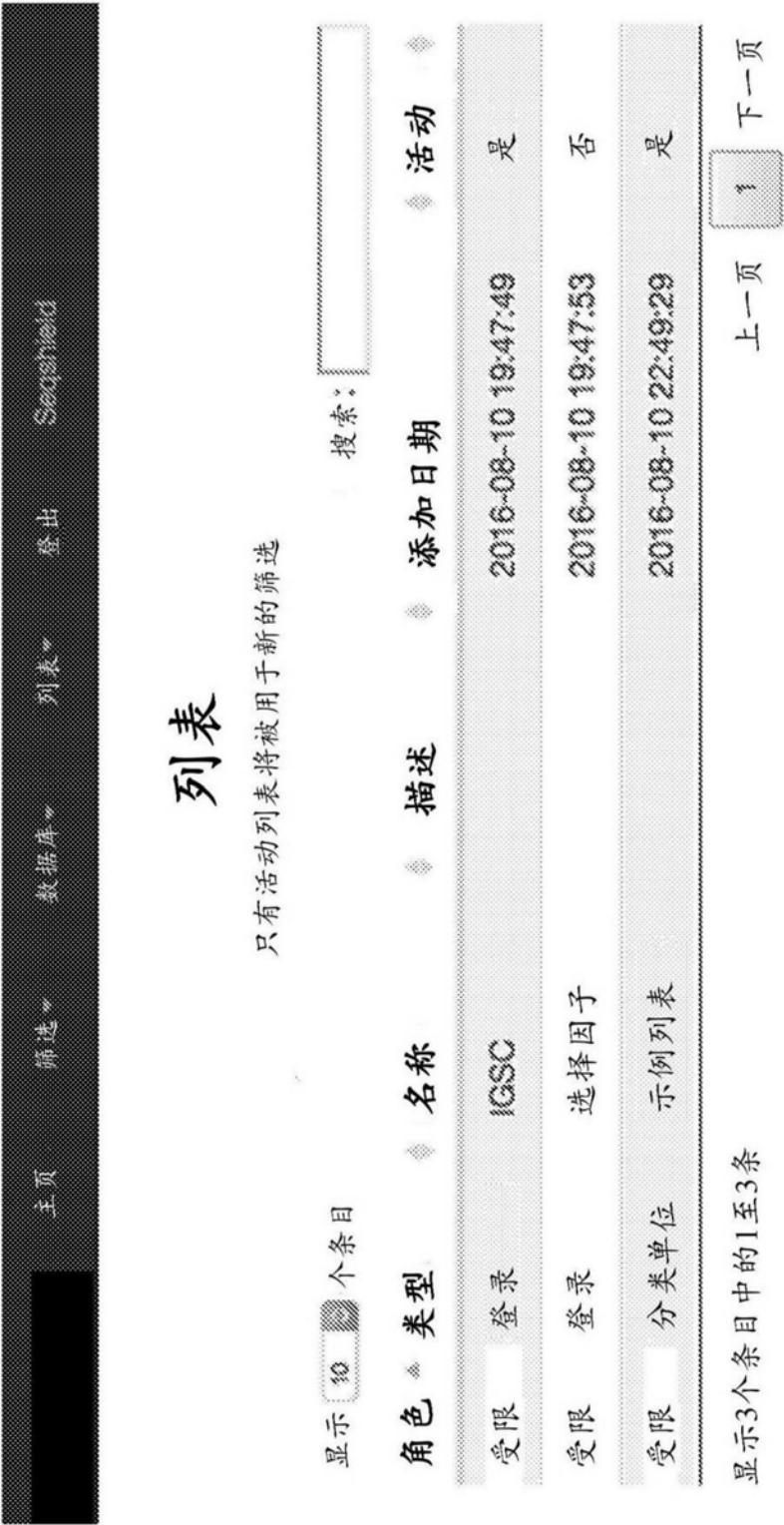


图4

主页

筛选

数据库\*

列表\*

登出

Seqshard

## 提交筛选

名称

数据库

描述

fasta

选择文件 未选择文件

提交

图5

主页

筛选

数据库

列表

登出

Seqshield

生物安全筛选

显示10个条目

搜索

名称	用户	状态	序列	未经审查	未关注	关注	添加日期	Blast
筛选1	Ben	完成	16	15	0	0	2016-08-11 05:56:30	查看
筛选2	Ben	完成	16	11	0	0	2016-08-11 16:44:07	查看
筛选3	Ben	完成	16	16	0	0	2016-08-11 16:44:22	查看
筛选4	Ben	完成	6990	6990	0	0	2016-08-11 16:45:09	查看
筛选5	Ben	Blast 运行中	0	0	0	0	2016-08-11 16:50:29	查看

显示5个条目中的1-5条

上一页1下一页

图6

用于筛选的列表

显示 10 个条目

搜索:

角色	类型	名称	描述	添加日期	活动
受限	无	IGSC		2015-08-11 16:40:37	是
受限	无	示例列表	分类单位列表	2015-08-11 16:41:38	是

显示2个条目中的1-2条

上一页1下一页

筛选出的序列

序列 ID	描述	关注	原因	命中	未经审查	安全	有害
end_Gera	end_Gera	未经审查		1	1	0	0
end_Gera	end_Gera						

受限的比对

登录	分类单元	描述	关注	原因	起始	结束	长度	链	类型
672735645	无	Gera细胞周期 调控家族-猪布 鲁氏菌2型	✓ 未经审查		32	64	33	1	登录
			未关注						

图7

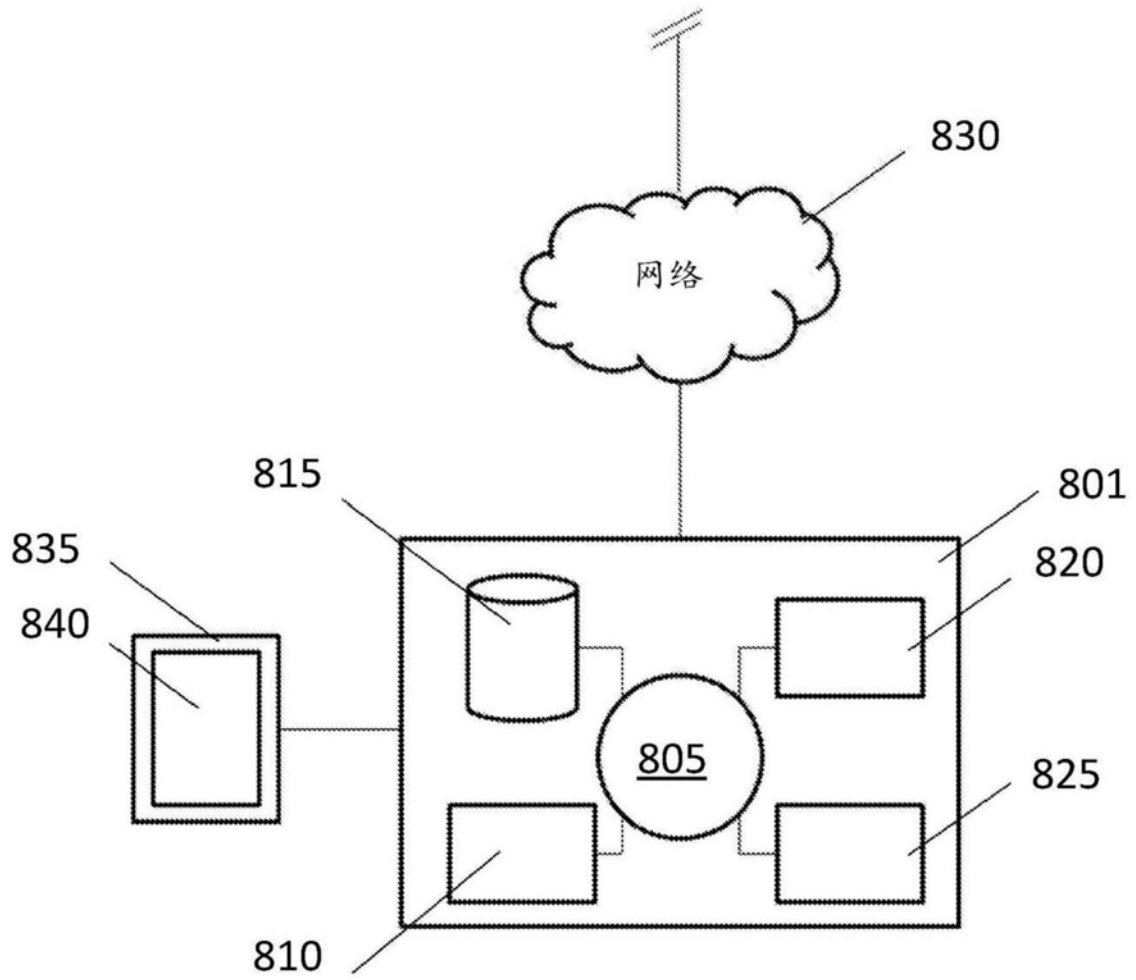


图8



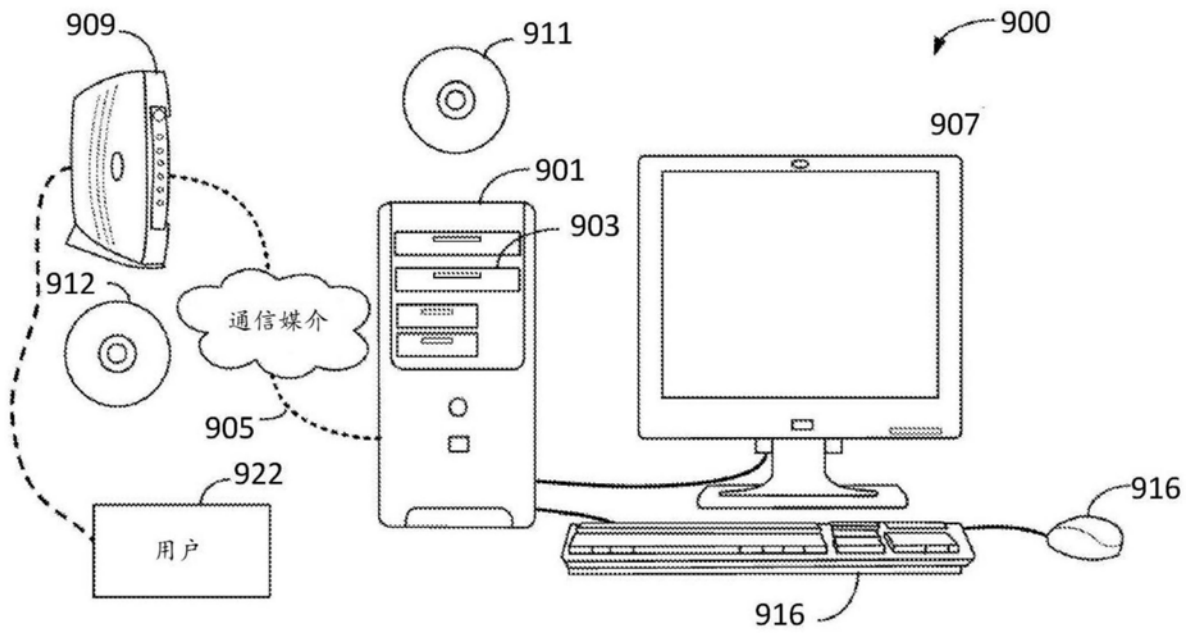


图9

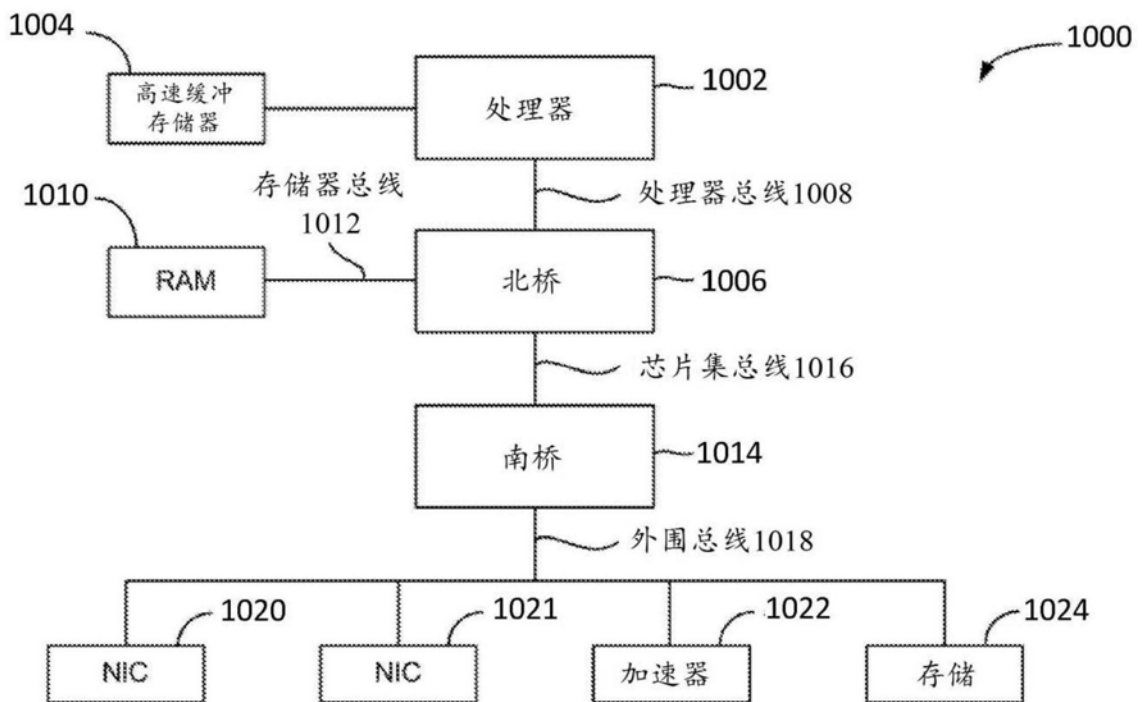


图10

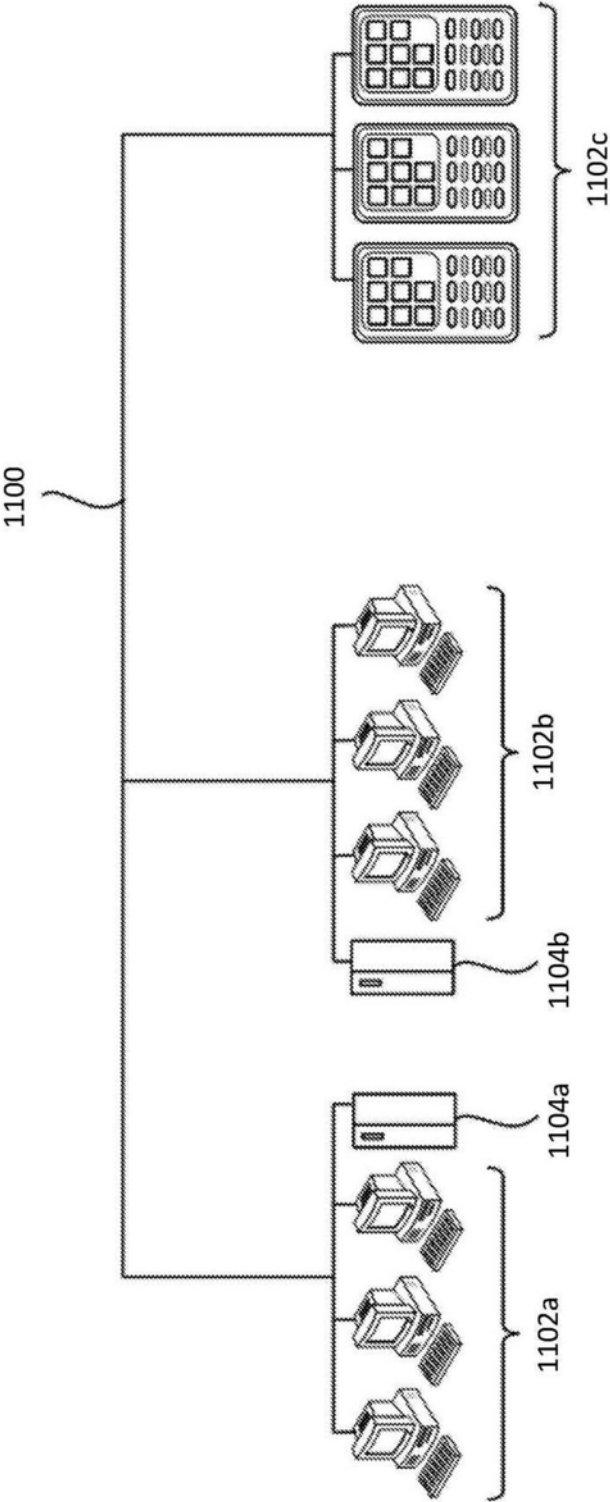


图11

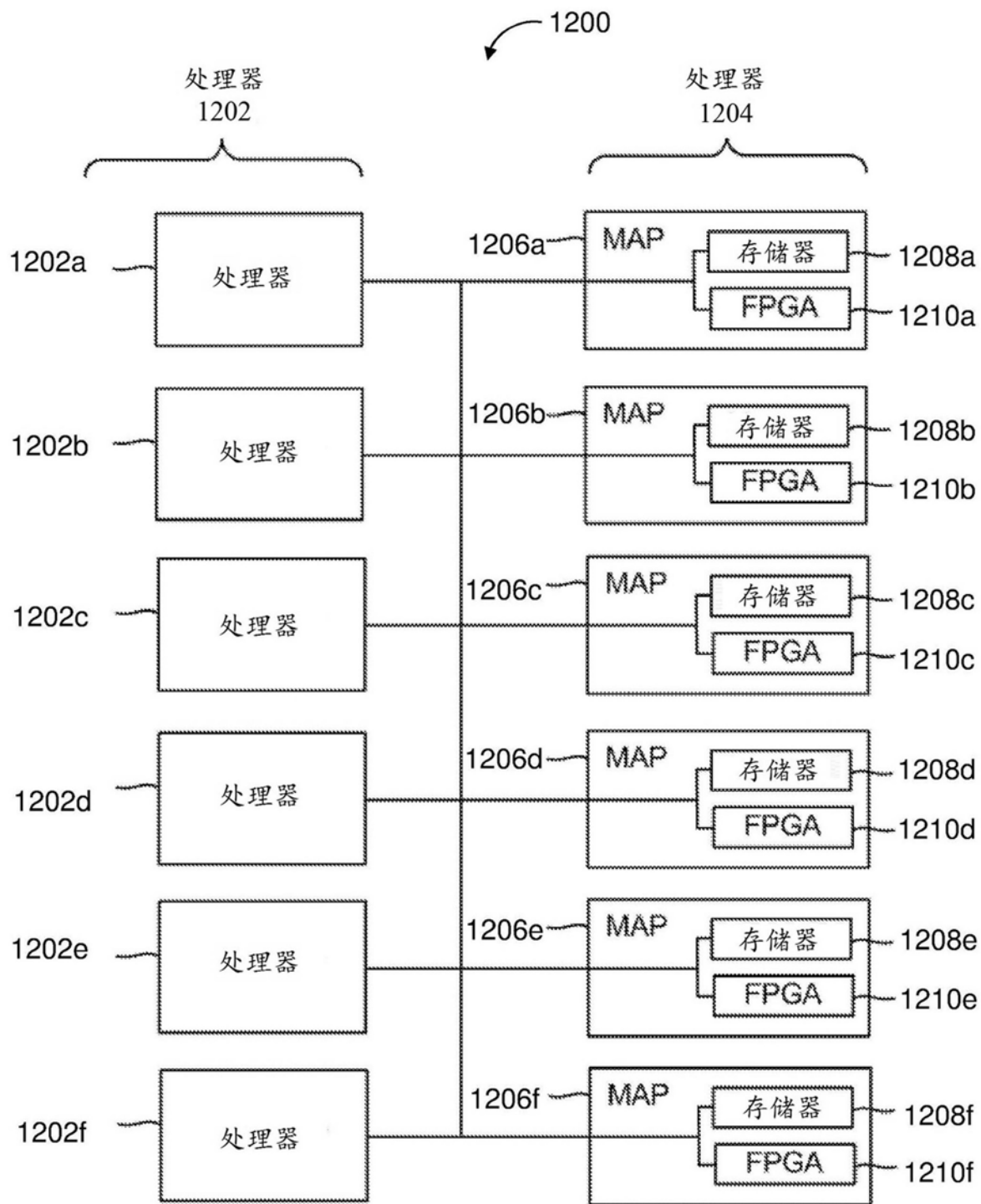


图12