(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 111095421 B
(45)授权公告日 2024.02.02

(21)申请号 201880054764.5

(22)申请日 2018.08.09

(65)同一申请的已公布的文献号
申请公布号 CN 111095421 A

(43)申请公布日 2020.05.01

(30)优先权数据
15/693,019 2017.08.31 US

(85)PCT国际申请进入国家阶段日
2020.02.24

(86)PCT国际申请的申请数据
PCT/IB2018/056009 2018.08.09

(87)PCT国际申请的公布数据
WO2019/043481 EN 2019.03.07

(73)专利权人 国际商业机器公司
地址 美国纽约

(72)发明人 A·马哈拉纳
M·C·康斯坦丁内斯库

(74)专利代理机构 北京市中咨律师事务所 11247
专利代理师 李永敏 于静

(51)Int.Cl.
G16B 20/00(2019.01)
G06F 16/2458(2019.01)

(56)对比文件
CN 101535945 A,2009.09.16
CN 106687966 A,2017.05.17
CN 1680589 A,2005.10.12
US 2004086861 A1,2004.05.06
US 2008077607 A1,2008.03.27
US 2011119240 A1,2011.05.19
US 2013132353 A1,2013.05.23
杨文涛;张晶;闫冠雄;田苗;袁冬霞;缪炜;曾宏辉;熊杰.四膜虫功能基因组数据库增量更新2016:生活史和减数分裂转录组及磷酸化蛋白组资源建设.基因组学与应用生物学.2016,(第06期),全文.
杨森;夏燕;曹顺良;邓绪斌;朱扬勇.语义异构生物数据源中的数据集成与更新.计算机工程.2008,(第08期),全文.
(续)

审查员 杨哲

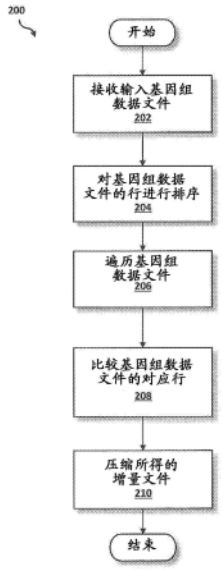
权利要求书2页 说明书19页 附图11页

(54)发明名称

基因文件的上下文感知增量算法

(57)摘要

提供了一种用于压缩多个基因组数据文件的至少一个增量文件的方法、计算机系统和计算机程序产品。本发明可以包括接收多个基因组数据文件作为输入。本发明还可以包括通过遍历所接收的多个基因组数据文件来确定多个行。然后,本发明可以包括比较与所遍历的多个基因组数据文件相关联的多个行。本发明可以进一步包括基于所比较的多行来生成多个所得的增量文件。本发明还可以包括通过利用通用文件压缩器来压缩所生成的多个所得的增量文件。



CN 111095421 B

[转续页]

[接上页]

(56)对比文件

赵秀娟;裴智勇;刘佳;蔡禄.多样性增量结合支持向量机方法预测酵母核小体定位.生物物理学报.2010,(第05期),全文.

陈凤珍;李玲;操利超;严志祥.四种常用的生物序列比对软件比较.生物信息学.2016,(第01期),全文.

1.一种用于压缩基因组数据文件的至少一个增量文件的方法,该方法包括:

接收第一文件和第二文件作为输入,所述第一文件和所述第二文件属于所述基因组数据文件,其中,所述第一文件和所述第二文件是tab分隔的,所述第一文件的每行包括第一文件字段,所述第二文件的每行包括第二文件字段,其中,所述第一文件包括来自与第一样本相关的参考基因组的第一数据,所述第二文件包括来自与第二样本相关的另一基因组的第二数据,所述第一样本和所述第二样本是不同的;

通过遍历所述第一文件和所述第二文件来确定所述第一文件的排序的行和所述第二文件的对应的排序的行,所述排序的行和所述对应的排序的行基于基因组的映射位置按照升序排列;

比较所述排序的行中的所述第一文件字段和所述对应的排序的行中的所述第二文件字段;

基于所述比较生成所得的增量文件;以及

通过利用通用文件压缩器来压缩所生成的所得的增量文件。

2.如权利要求1所述的方法,还包括:

将所压缩的所得的增量文件存储到用户设备上;以及

向用户呈现所压缩的所得的增量文件。

3.如权利要求1所述的方法,其中,

所述第一文件包括至少一个源文件;以及

所述第二文件包括至少一个目标文件。

4.如权利要求3所述的方法,还包括:

确定所述第一文件和所述第二文件未排序;以及

利用排序工具对所述第一文件和所述第二文件中的多个行中的每一个进行排序。

5.如权利要求1所述的方法,还包括:

确定所述第一文件和所述第二文件为兼容格式。

6.如权利要求1所述的方法,其中,所述遍历还包括:

读取与所述第一文件和所述第二文件相关联的多行中的每一行;以及

确定所述第一文件和所述第二文件是同步的。

7.如权利要求1所述的方法,其中,所述比较还包括:

比较与所比较的多个行相关联的多个特定列;

确定所比较的多个特定列与所遍历的所述第一文件和所述第二的所比较的多个行的匹配;

比较多个行中的每一行的多个附加列;

基于所遍历的所述第一文件和所述第二文件的多行中的每一行的所比较的多个附加列,生成所得的增量文件;以及

读取与所遍历的所述第一文件和所述第二文件相关联的多个行中的下一行。

8.如权利要求1所述的方法,其中,所述比较还包括:

比较与所比较的多个行相关联的多个特定列;

确定所比较的多个确定的列和所遍历的所述第一文件和所述第二文件的所比较的多个行的不匹配;以及

读取与所遍历的所述第一文件和所述第二文件相关联的多个行中的下一行。

9.一种用于压缩基因组数据文件的至少一个增量文件的系统,包括:

存储器;以及

与所述存储器通信的至少一个处理器,其中,所述至少一个处理器被配置为执行根据权利要求1-8中任一项所述的方法。

10.一种计算机可读存储介质,包括指令,当在计算机系统上执行所述指令时,所述指令用于执行根据权利要求1-8中任一项所述的方法。

基因文件的上下文感知增量算法

背景技术

[0001] 本发明总体上涉及计算领域,并且更具体地涉及计算生物学。

[0002] 基因组分析流水线(流水线)涉及预处理、变体发现和调用集细化的多个步骤,以便从原始序列读取中提取生物学上有意义的输出。在每个此类步骤中,流水线都会生成输出文件,输出文件的大小在兆字节到TB字节之间,具体取决于输入序列读取的大小。

[0003] 对这些文件的检查显示,并非每一步都保存到输出文件的所有信息都是新生成的。相当数量的数据只是从输入文件中转移到输出中,从而在流水线执行期间甚至之后对存储造成不必要的压力。流水线的每个阶段可能要花费数小时或数天,因此中间文件将保留以备将来调查、更改或在流水线中分支。

发明内容

[0004] 本发明的实施例公开了一种用于压缩多个基因组数据文件的至少一个增量文件的方法、计算机系统和计算机程序产品。本发明可以包括接收多个基因组数据文件作为输入。本发明还可以包括通过遍历所接收的多个基因组数据文件来确定多个行。然后,本发明可以包括比较与所遍历的多个基因组数据文件相关联的多个行。本发明可以进一步包括基于所比较的多行来生成多个所得的增量文件。本发明还可以包括通过利用通用文件压缩器来压缩所生成的多个所得的增量文件。

附图说明

[0005] 通过以下结合附图对示例性实施例的详细描述,本发明的这些和其他目的,特征和优点将变得显而易见。附图的各种特征未按比例绘制,因为图示是为了清楚起见,以帮助本领域技术人员结合详细描述来理解本发明。

[0006] 在附图中:

[0007] 图1示出了根据至少一个实施例的联网计算机环境;

[0008] 图2是示出了根据至少一个实施例的用于压缩基因组数据文件的增量文件的过程的操作流程图;

[0009] 图3A是示出了根据至少一个实施例的用于压缩基因组数据文件的增量文件的示例性过程的操作流程图;

[0010] 图3B是示出了根据至少一个实施例的用于压缩序列比对(Sequence Alignment)/图谱格式(Map format,SAM)的基因组数据文件的增量文件的示例性过程的操作流程图;

[0011] 图3C是示出了根据至少一个实施例的用于压缩以变体调用格式(Variant Call Format,VCF)的基因组数据文件的增量文件的示例性过程的操作流程图;

[0012] 图4是示出根据至少一个实施例的用于比较两个基因组数据文件的示例性过程的框图;

[0013] 图5是示出根据至少一个实施例的用于识别以序列比对/图谱格式(SAM)的基因组数据文件的分层结构的示例性过程的操作流程图;

[0014] 图6是示出了根据至少一个实施例的用于识别以变体调用格式(VCF)的基因组数据文件的分层结构的示例性过程的操作流程图;

[0015] 图7是根据至少一个实施例的图1所示的计算机和服务器的内部和外部组件的框图;

[0016] 图8是根据本公开的实施例的包括图1所示的计算机系统的说明性云计算环境的框图;以及

[0017] 图9是根据本公开的实施例的图8的说明性云计算环境的功能层的框图。

具体实施方式

[0018] 本文公开了要求保护的结构和方法的详细实施例;然而,可以理解,所公开的实施例仅是可以以各种形式体现的所要求保护的结构和方法的说明。然而,本发明可以以许多不同的形式实施,并且不应被解释为限于这里阐述的示例性实施例。而是,提供这些示例性实施例,使得本公开将是彻底和完整的,并且将本发明的范围充分传达给本领域技术人员。在说明书中,可以省略众所周知的特征和技术的细节,以避免不必要地混淆所呈现的实施例。

[0019] 在任何可能的技术细节结合层面,本发明可以是系统、方法和/或计算机程序产品。计算机程序产品可以包括计算机可读存储介质,其上载有用于使处理器实现本发明的各个方面的计算机可读程序指令。

[0020] 计算机可读存储介质可以是可以保持和存储由指令执行设备使用的指令的有形设备。计算机可读存储介质例如可以是——但不限于——电存储设备、磁存储设备、光存储设备、电磁存储设备、半导体存储设备或者上述的任意合适的组合。计算机可读存储介质的更具体的例子(非穷举的列表)包括:便携式计算机盘、硬盘、随机存取存储器(RAM)、只读存储器(ROM)、可擦式可编程只读存储器(EPROM或闪存)、静态随机存取存储器(SRAM)、便携式压缩盘只读存储器(CD-ROM)、数字多功能盘(DVD)、记忆棒、软盘、机械编码设备、例如其上存储有指令的打孔卡或凹槽内凸起结构、以及上述的任意合适的组合。这里所使用的计算机可读存储介质不被解释为瞬时信号本身,诸如无线电波或者其他自由传播的电磁波、通过波导或其他传输媒介传播的电磁波(例如,通过光纤电缆的光脉冲)、或者通过电线传输的电信号。

[0021] 这里所描述的计算机可读程序指令可以从计算机可读存储介质下载到各个计算/处理设备,或者通过网络、例如因特网、局域网、广域网和/或无线网下载到外部计算机或外部存储设备。网络可以包括铜传输电缆、光纤传输、无线传输、路由器、防火墙、交换机、网关计算机和/或边缘服务器。每个计算/处理设备中的网络适配卡或者网络接口从网络接收计算机可读程序指令,并转发该计算机可读程序指令,以供存储在各个计算/处理设备中的计算机可读存储介质中。

[0022] 用于执行本发明操作的计算机程序指令可以是汇编指令、指令集架构(ISA)指令、机器指令、机器相关指令、微代码、固件指令、状态设置数据、集成电路配置数据或者以一种或多种编程语言的任意组合编写的源代码或目标代码,所述编程语言包括面向对象的编程语言—诸如Smalltalk、C++等,以及过程式编程语言—诸如“C”语言或类似的编程语言。计算机可读程序指令可以完全地在用户计算机上执行、部分地在用户计算机上执行、作为一

个独立的软件包执行、部分在用户计算机上部分在远程计算机上执行、或者完全在远程计算机或服务器上执行。在涉及远程计算机的情形中,远程计算机可以通过任意种类的网络—包括局域网(LAN)或广域网(WAN)—连接到用户计算机,或者,可以连接到外部计算机(例如利用因特网服务提供商来通过因特网连接)。在一些实施例中,通过利用计算机可读程序指令的状态信息来个性化定制电子电路,例如可编程逻辑电路、现场可编程门阵列(FPGA)或可编程逻辑阵列(PLA),该电子电路可以执行计算机可读程序指令,从而实现本发明的各个方面。

[0023] 这里参照根据本发明实施例的方法、装置(系统)和计算机程序产品的流程图和/或框图描述了本发明的各个方面。应当理解,流程图和/或框图的每个方框以及流程图和/或框图中各方框的组合,都可以由计算机可读程序指令实现。

[0024] 这些计算机可读程序指令可以提供给通用计算机、专用计算机或其它可编程数据处理装置的处理器,从而生产出一种机器,使得这些指令在通过计算机或其它可编程数据处理装置的处理器执行时,产生了实现流程图和/或框图中的一个或多个方框中规定的功能/动作的装置。也可以把这些计算机可读程序指令存储在计算机可读存储介质中,这些指令使得计算机、可编程数据处理装置和/或其他设备以特定方式工作,从而,存储有指令的计算机可读介质则包括一个制造品,其包括实现流程图和/或框图中的一个或多个方框中规定的功能/动作的各个方面的指令。

[0025] 也可以把计算机可读程序指令加载到计算机、其它可编程数据处理装置、或其它设备上,使得在计算机、其它可编程数据处理装置或其它设备上执行一系列操作步骤,以产生计算机实现的过程,从而使得在计算机、其它可编程数据处理装置、或其它设备上执行的指令实现流程图和/或框图中的一个或多个方框中规定的功能/动作。

[0026] 附图中的流程图和框图显示了根据本发明的多个实施例的系统、方法和计算机程序产品的可能实现的体系架构、功能和操作。在这点上,流程图或框图中的每个方框可以代表一个模块、程序段或指令的一部分,所述模块、程序段或指令的一部分包含一个或多个用于实现规定的逻辑功能的可执行指令。在有些作为替换的实现中,方框中所标注的功能也可以以不同于附图中所标注的顺序发生。例如,两个连续的方框实际上可以基本并行地执行,它们有时也可以按相反的顺序执行,这依所涉及的功能而定。也要注意的是,框图和/或流程图中的每个方框、以及框图和/或流程图中的方框的组合,可以用执行规定的功能或动作的专用的基于硬件的系统来实现,或者可以用专用硬件与计算机指令的组合来实现。

[0027] 以下描述的示例性实施例提供了一种用于压缩用于基因组数据文件的增量文件的系统、方法和程序产品。这样,本实施例具有通过减少由基因组分析流水线产生的中间文件的存储占用空间来改善计算生物学的技术领域的能力。更具体地,可以将至少两个基因组数据文件(例如,至少一个源文件和一个目标文件)作为输入输入到基于基因组的增量压缩程序中,然后,基于基因组的增量压缩程序可以利用已知的开放源代码工具(包括排序算法)对基因组数据文件的行进行排序。基于基因组的增量压缩程序然后可以顺序遍历源文件和目标文件的每一行,并且可以比较源文件和目标文件之间的对应行以生成所得的增量文件。然后,基于基因组的增量压缩程序可以通过利用已知的通用文件压缩器来压缩生成的所得的增量文件。

[0028] 如前所述,基因组分析流水线(流水线)涉及预处理、变体发现和调用集细化的多

个步骤,以便从原始序列读数中提取生物学上有意义的输出。在每个此类步骤中,流水线都会生成输出文件,输出文件的大小在兆字节到TB字节之间,具体取决于输入序列读取的大小。

[0029] 对这些文件的检查显示,并非在每一步保存到输出文件的所有信息都是新生成的。相当数量的数据只是从输入文件中转移到输出中,从而在流水线执行期间甚至之后对存储造成不必要的压力。流水线的每个阶段可能要花费数小时或数天,因此中间文件将保留以备将来调查,更改或在流水线中分支。

[0030] 可能存在增量或差分压缩算法(例如xdelta3或vcdiff);但是,当源文件和目标文件之间存在相对较大的连续公共子字符串时,这些压缩算法可能会更有效。具有某些格式的基因组数据,例如序列比对/图谱格式(SAM)和变异调用格式(VCF),可能包含由字段组成的相对较小的记录,其中某些字段和子字段可能因一个版本而不同。

[0031] 因此,除其他事项外,创建一种数据缩减方法以最小化基因组分析流水线为小记录而生成的中间文件的存储空间或重复信息量可能是有利的。

[0032] 根据至少一个实施例,基于基因组的增量压缩程序可以将一对输入和输出文件之间的增量文件(即增量)减少为一组简单的柱状操作,例如匹配、替代、插入和删除。可以通过具有相应列的行来解析源文件和目标文件。增量文件可以是可以在源文件上执行以将源文件转换为目标文件的一组操作的顺序帐户。可以使用已知的通用文件压缩器(例如.zip)(即,文件格式和用于文件压缩和解压缩的软件应用程序)来压缩所得的增量文件。

[0033] 根据至少一个实施例,基于基因组的增量压缩程序可以使存储和文件系统、产品或云产品能够在医疗保健行业中脱颖而出,以较小的占用空间来支持和运行流水线,并加速通过各种通信的传输网络。基于基因组的增量压缩程序还可以通过从各种中间步骤重新启动来启用流水线中的替代分支,并且可以启用对假设和变异的更快检查。

[0034] 根据至少一个实施例,基因组数据文件可以表示为一组文本行(即记录),其中每个文本行包括固定数目的列或字段(例如,如SAM或VCF格式中所定界的TAB)。通过比较源文件和目标文件,可以计算增量。源文件和目标文件的每一行都可以按照其映射位置的升序进行排序。然后,可以同步地逐行遍历源文件和目标文件,并且可以在源文件和目标文件中比较具有相同位置的相应行,从而得到增量文件。然后可以使用通用压缩程序压缩所得的增量文件。

[0035] 根据至少一个实施例,基于基因组的增量压缩程序可以考虑以下事实:可以对源文件和目标文件进行排序并将其映射到参考基因组文件。这样,基于基因组的增量压缩程序可以对源文件和目标文件之间的两个相似记录使用分层标识系统。找到正确的匹配后,然后可以将两个记录解析为各个字段,并逐字段比较该增量。

[0036] 根据至少一个实施例,基于基因组的增量压缩程序可以通过缓冲映射到相同位置的目标文件中的行,并为源文件中的同一位置的下一行重新扫描目标文件来改善行匹配。基因组文件中的每一行可以对应于映射到参考基因组的“读取”。通常,可能有许多读取被映射到参考基因组中的相同位置。实际上,与基因组相同区域重叠的读取的平均数可能是覆盖率。覆盖范围可能很高,可能会映射到同一位置的数十次读取甚至数百次读取。

[0037] 根据至少一个实施例,可以使用较细的粒度(例如,子列)来查找可能具有预定义的较小粒度的列中的增量。基因组格式的每个记录的最后一列(即行)可以包括一组键值

对。在流水线的步骤中,可以添加一些新的键值对,也可以删除一些。这样,通过使用较小的粒度,可以存在较小的增量。

[0038] 根据至少一个实施例,增量编码(即,压缩方法)可以以文件对之间的差异形式存储数据。即使存在几种已知的增量压缩算法(例如vcdiff和xdelta),它们都在搜索顺序数据之间的差异,但是当应用于基因组数据文件时,这些已知的增量算法可能会生成比实际差异更大的所得的增量文件。这些已知的增量算法可能无法利用基因组数据文件的排序顺序和行的TAB分隔结构的优势。这些文件,包括SAM和VCF格式文件,除了可以在基于基因组的增量压缩程序中为有效的增量压缩而排除的核苷酸序列和质量值之外,还可以包含大量的辅助信息。

[0039] 根据至少一个实施例,算法的设计可以基于流水线和数据文件的显着特征。流水线的第一步可能是序列比对,这可能会导致生成SAM文件,其中每个读取的原始序列都可以映射到参考基因组中的特定位置或未映射。除非在重新对齐期间更改,否则对齐位置在整个流水线的其余部分中可能保持不变。大多数流水线工具只能在对齐步骤之后对文件进行排序。基于基因组的增量压缩程序可以利用排序来同步计算一对文件之间的差异。如果排序顺序在两个文件之间有很大变化,则增量的大小也可能会增加。基于基因组的增量压缩程序可以检测该增量是否可以增加到超过可接受的阈值。在这种情况下,基于基因组的增量压缩程序可能会中止增量压缩并使目标文件保持不变。

[0040] 根据至少一个实施例,基于基因组的增量压缩程序可以快速搜索过源文件和目标文件的头,并到达第一条记录。对于源文件中的每个单独记录,增量文件中可能会有相应的操作。对于每个源记录,基于基因组的增量压缩程序可能首先尝试在目标文件中找到匹配项。,用于查找此匹配项的搜索窗口,对于SAM可能仅限于带有相同值的参考序列名称(REF或RNAME)和基于1的最左侧映射位置(POS);并且对于VCF,来自参考基因组的染色体标识符或者在汇编文件(#CHROM)和位置(POS)中指向重叠群(例如,重叠序列数据或读取)的用尖括号括起来的标识符字符串。基于基因组的增量压缩程序可以顺序读取目标文件,直到找到匹配项或搜索窗口耗尽为止。在搜索一个源记录的匹配过程中可能会遇到的来自目标文件的不匹配记录,其可以被推送到缓冲器(例如,可以临时存储文件的存储器的一小部分),以处理不匹配的记录,当基于基因组的增量压缩程序前进到源文件上的下一个记录时,同样处理。如果推送到缓冲器的记录数超过可接受的阈值,则基于基因组的增量压缩程序可能会放弃搜索。如果找不到匹配项,则源记录可能会在增量文件中标记为“删除记录”。在顺序读取目标文件时,如果基于基因组的增量压缩程序遇到一条记录,该记录的REF、POS或#CHROM、REF可能小于源记录中的记录,则可以标记目标文件中的特定记录作为“插入记录”,并且可以添加到增量文件中。

[0041] 根据至少一个实施例,基于基因组的增量压缩程序可以基于某些权利要求利用分层识别系统来匹配源文件和目标文件。例如,在SAM文件中,基于基因组的增量压缩程序可以利用与查询模板名称(QNAME)、按位标志(FLAG)、REF、POS(即第一至第四列)相关联的列;而在VCF文件中,基于基因组的增量压缩程序可以利用与#CHROM、POS、REF和备用碱基(ALT)(即,第一至第二列以及第四至第五列)相关的列。

[0042] 根据至少一个实施例,在确定源记录与目标记录之间的匹配之后,可以对某些字段(例如,可选字段或非强制字段(OPTIONAL)或SAM文件中第十二列以及其他信息(INFO)或

VCF文件中的第八列)进行逐字段比较。对于比较的字段,当字段不匹配时,该特定字段可以标记为“替换”,并且可以在增量文件中添加新值。排除在比较范围之外的文件(例如,OPTIONAL和INFO)可能包含几个键值对,或者仅包含键和仅包含值,这可以用分号、空格或替代标记来分隔。然后,可以将这些排除在比较之外的字段解析为目标文件和源文件的各自子字段。源文件中存在但目标文件中不存在的键或值可以在增量文件中标记为“删除”。目标中存在但源中不存在的键或值可以标记为“插入”,并可以添加到增量文件中。确定了源记录的操作并将其添加到增量文件后,基于基因组的增量压缩程序可以继续读取源文件中的下一个记录,并且可以再次重复该过程,除非源文件和目标文件文件的末尾(EOF)都遇到了。生成的增量文件可能具有高度可压缩性,并且可以压缩以进行存档。

[0043] 根据至少一个实施例,对于每个源记录,基于基因组的增量压缩程序可以在从目标文件读取新记录之前首先在缓冲器中搜索匹配项。每次搜索窗口更改时,可能会清除缓冲器。所有不匹配的记录都可以标记为“插入”,并添加到增量中。

[0044] 参照图1,示出了根据一个实施例的示例性联网计算机环境100。联网计算机环境100可以包括具有处理器104和数据存储设备106的计算机102,该计算机102能够运行软件程序108和基于基因组的增量压缩程序110a。联网计算机环境100还可以包括服务器112,该服务器112能够运行可以与数据库114和通信网络116交互的基于基因组的增量压缩程序110b。联网计算机环境100可以包括多个计算机102和服务器112,仅示出其中之一。通信网络116可以包括各种类型的通信网络,诸如广域网(WAN)、局域网(LAN)、电信网络、无线网络、公共交换网络和/或卫星网络。应当理解,图1仅提供了一种实现方式的图示,并且不暗示对可以实现不同实施例的环境的任何限制。可以基于设计和实现要求对所描绘的环境进行许多修改。

[0045] 客户端计算机102可以经由通信网络116与服务器计算机112进行通信。通信网络116可以包括诸如有线、无线通信链路或光纤电缆之类的连接。如将参考图7讨论的,服务器计算机112可以分别包括内部组件902a和外部组件904a,并且客户端计算机102可以分别包括内部组件902b和外部组件904b。服务器计算机112还可以在云计算服务模型中运行,例如软件即服务(SaaS)、平台即服务(PaaS)或基础设施即服务(IaaS)。服务器112也可以位于云计算部署模型中,例如私有云、共同体云、公共云或混合云。客户端计算机102可以是例如移动设备、电话、个人数字助理、上网本、膝上型计算机、平板计算机、台式计算机或能够运行程序,访问计算机的任何类型的计算设备。根据本实施例的各种实现,基于基因组的增量压缩程序110a、110b可以与数据库114交互,该数据库114可以嵌入各种存储设备中,例如但不限于存储设备。计算机/移动设备102、联网服务器112或云存储服务。

[0046] 根据本实施例,使用客户端计算机102或服务器计算机112的用户可以(分别)使用基于基因组的增量压缩程序110a、110b以压缩用于基因组数据文件的增量文件。基于基因组的增量压缩方法将在下面参照图2-6更详细地说明。

[0047] 现在参考图2,描绘了根据至少一个实施例的用于基于基因组的增量压缩程序110a和110b使用的基因组数据文件的示例性压缩过程200的操作流程图。

[0048] 在202处,接收基因组数据文件作为到基于基因组的增量压缩程序110a、110b中的输入。使用用户设备(例如,用户计算机102)上的软件程序108,基因组数据文件可以作为输入经由通信网络116传输到基于基因组的增量压缩程序110a、110b中。基因组数据可以包括

与生物体相关联基因组数据,并且基因组数据上的文件可能包括映射到参考基因组以及针对参考基因组检测到的变异的脱氧核糖核酸(DNA)序列读取(即,DNA分子中核苷酸的精确顺序)。基因组数据文件还可包括一组行(即记录),每个行包括用于至少两个文件,即源文件(即参考文件)和目标文件,的固定数目的列(即字段)。源文件是可以直接源自生物体参考基因组的数据。可以根据文件是目标文件还是源文件来标记每个文件(例如,“target_name.format type”或“source_name.format type”)。基于基因组的增量压缩程序110a、110b可以利用源文件来重构目标文件。

[0049] 例如,基于基因组的增量压缩程序110a、110b接收从基因组分析工具包(GATK)传输的源文件(即source_liverHS1567.sam)和目标文件(即target_liverHS1567.sam)。源文件包括来自与健康人类肝细胞样本相关的参考基因组的数据,目标文件包括与来自患有肝癌的人类的肝细胞样本相关的数据。下面将参照图3A更详细地描述用于基因组数据文件的示例性压缩过程300,包括基于源基因组的增量压缩程序110a、110b在302处接收源文件和目标文件作为输入。

[0050] 在本实施例中,基因组数据文件可以作为诸如GATK之类的基因组分析流水线的执行的输出而产生。流水线脚本可以富含用于对目标文件和源文件执行的基于基因组的增量压缩程序110a、110b的特定命令。另外,在启动基于基因组的增量压缩程序110a、110b之前,可以使用外部引擎对已压缩的一些输出文件(例如,SAM文件的的二进制版本(BAM)和VCF文件的二进制版本(BCF))进行解压缩。

[0051] 在本实施例中,可以从从文件夹或目录(即数据库114)中运行流水线时生成的文件中检索源文件和目标文件,可以在文件夹或目录(即数据库114)中生成和存储输出文件。

[0052] 在本实施例中,如果仅一个基因组数据文件作为输入被输入到基于基因组的增量压缩程序110a、110b,则基于基因组的增量压缩程序110a、110b可能不计算所得的增量文件。而是,基于基因组的增量压缩程序110a、110b可以将错误消息返回给用户。用户可以重新输入相同的输入文件;但是,该文件应与基于基因组的增量压缩程序110a、110b的相应源文件或目标文件一起输入,以计算所得的增量文件。

[0053] 接下来在204,通过基于基因组的增量压缩程序110a、110b对基因组数据文件的行进行排序。源文件和目标文件的行可以基于它们在参考基因组中的映射位置(POS)按照升序排列,这可以缩小搜索窗口的范围,以匹配源文件和目标文件之间的行,从而将缓冲器占用保持在最小化并且优化基于基因组的增量压缩程序110a、110b的运行时间。可以利用包括已知排序算法的已知开源排序工具来对基因组数据文件的行进行排序。

[0054] 根据目标文件和源文件的格式类型,文件可能已经在流水线中排序。如果源文件和目标文件在流水线中被排序,则在202处当输入到基于基因组的增量压缩程序110a、110b作为输入时,目标文件和源文件可以在其各自的行中保留排序的顺序。

[0055] 继续先前的示例,基于基因组的增量压缩程序110a、110b确定从GATK获得的源文件和目标文件先前已被排序,因此,基于基因组的增量压缩程序110a、110b不必基于映射位置(POS)对行排序。

[0056] 出于该示例的目的,基于基因组的增量压缩程序110a、110b仅与SAM或VCF格式的文件兼容。如果文件是任何其他格式,则将向用户显示错误消息,并且基于基因组的增量压缩程序110a、110b停止运行,直到将兼容的源文件和目标文件作为输入来输入为止。在该示

例中,基于基因组的增量压缩程序110a、110b确定源文件和目标文件为SAM格式,其与基于基因组的增量压缩程序110a、110b兼容。这样,基于基因组的增量压缩程序110a、110b启动缓冲器(即,如果目标文件与源文件不匹配,则可以存储目标文件的存储器的一小部分)。下面将相对于图3A更详细地描述用于基因组数据文件的压缩增量文件的示例性过程,包括基于基因组的增量压缩程序110a、110b检查文件以识别格式并启动(INIT)缓冲器。

[0057] 但是,如果在基于基因组的增量压缩程序110a、110b接收文件作为输入之前未对源文件和目标文件进行排序,则基于基因组的增量压缩程序110a、110b通过使用开源排序工具可以对源文件和目标文件的每个文件的行基于其POS(即每个文件行的第四列)以升序进行排序。

[0058] 在本实施例中,基于基因组的增量压缩程序110a、110b可以检查文件格式类型(例如,SAM、BAM、VCF、BCF)的兼容性。如果文件格式类型与基于基因组的增量压缩程序110a、110b兼容,则可以继续进行基于基因组的增量压缩程序110a、110b。然而,如果文件格式类型与基于基因组的增量压缩程序110a、110b不兼容,则基于基因组的增量压缩程序110a、110b可以停止运行,并且可以向用户呈现错误消息。

[0059] 在本实施例中,可能必须根据基因组数据文件的复杂性来修改可用于对基因组数据文件的行进行排序的排序算法。

[0060] 然后在206处,遍历源文件和目标文件。基于基因组的增量压缩程序110a、110b可以顺序地遍历(即,读取)源文件和目标文件,其中基因组数据文件中的每一行可以被解释为一行。输入文件可以由基于基因组的增量压缩程序110a、110b读取,直到遇到换行符(即,表示行的结尾和新行的开始的特殊字符或字符序列)。此外,可以将行划分为列(即字段)。为了确认两个文件是同步的,可以比较行的值。可以同步遍历两个文件,以节省时间和内存的方式计算文件内容之间的差异。

[0061] 继续前面的示例,基于基因组的增量压缩程序110a、110b横穿源文件的每一行。源文件和目标文件包括四行,每行有三段数据。下面将参照图4更详细地描述比较两个基因组数据文件的框图。在遍历源文件的每一行之后,基于基因组的增量压缩程序110a、110b确定缓冲器是否为空(即,如果缓冲器为空,则可以从文件而不是缓冲器中读取目标文件)。对于此示例,缓冲器为空。因此,基于基因组的增量压缩程序110a、110b遍历文件中的每个目标行。在基于基因组的增量压缩程序110a、110b遍历源文件和目标文件的每一行之后,基于基因组的增量压缩程序110a、110b确认源文件和目标文件是同步的。下面将参照图3A更详细地描述用于压缩基因组数据文件的增量文件的示例性过程,包括基于基因组的增量压缩程序110a、110b读取源文件和目标文件的行。

[0062] 然后在208,比较目标文件和源文件的行。基于基因组的增量压缩程序110a、110b可以比较源文件和目标文件上的相应行(即,具有相同映射位置的行)以确定源文件和目标文件之间的任何差异,并生成所得的增量文件(即,生成的D文件)。

[0063] 基于基因组的增量压缩程序110a、110b可以对部分匹配的行(例如,比较每个文件的某些列)进行分层的按字段比较,以及对匹配的行(例如,比较两个文件中的其他列)进行进行全行比较。取决于所比较的行是否匹配,基于基因组的增量压缩程序110a、110b可以用于确定使用哪种比较。来自目标文件和源文件的每一行的比较可以通过对某些列进行逐字段比较来启动。对于在分层的逐字段比较期间的每个比较,可以记录中间匹配或不匹配。在

分层按字段进行比较的最后,行之间可能存在匹配或不匹配。匹配或不匹配可能是确定比较行是否同步以及确定是否存在不匹配的快速方法。如果两行之间存在匹配,则可以执行全行比较(即比较所有列)。比较的结果可以被捕获在所得增量文件中,该增量文件是基于基因组的增量压缩程序110a、110b的输出。

[0064] 继续先前的示例,基于基因组的增量压缩程序110a、110b首先对目标文件和源文件内的每一行执行分层的按字段进行的比较。由于源文件和目标文件均为SAM格式,因此有11个固定列。但是,分层的按字段比较仅比较四列(即POS、REF、QNAME和FLAG)。基于基因组的增量压缩程序110a、110b首先比较目标文件和源文件中第一行的POS和REF列,以确定POS和REF列是否匹配。由于目标文件和源文件中第一行的POS和REF列匹配,因此基于基因组的增量压缩程序110a、110b确定存在中间匹配,然后比较相同行的QNAME列。基于基因组的增量压缩程序110a、110b确定同一行的QNAME列匹配,并且存在中间匹配。然后,基于基因组的增量压缩程序110a、110b比较相同行的FLAG(按位)列。由于这些行的FLAG(按位)列不匹配,因此基于基因组的增量压缩程序110a、110b确定这些行的FLAG(按位)列之间的差是否等于或小于0x400(即,优选重复)。基于基因组的增量压缩程序110a、110b确定该差等于0x400。因此,目标文件和源文件的第一行匹配。通过比较源文件和目标文件创建的D文件的第一行中包含一个“m”。然后,基于基因组的增量压缩程序110a、110b可以在源文件和目标文件中的下一行上进行分层的按字段的比较。下面将参照图5更详细地描述用于识别SAM格式的基因组数据文件的分层结构的示例性过程。

[0065] 当执行行的部分匹配的逐字段比较方式时,随后的结果被生成并用于创建所得的D文件:

[0066] 当field_i(S)与field_i(T)不同时:D←(field_i(T),i)

[0067] 如果在分层逐字段比较中比较的行匹配,则基于基因组的增量压缩程序110a、110b将执行完整行比较,其中包括对目标文件和源文件行中的可选列的比较(即,其余可选列或非必选的7列或第12列),可选列包括RNAME(即第三列,表示序列名称)、MAPQ(即第五列,映射为质量)、CIGAR(即第六列,表示与是否序列匹配相关的字符串操作)、RNEXT(即第七列,表示下一个读取的参考名称)、PNEXT(即第八列,表示下一个读数的位置)、TLEN(即第九列,表示观察到的模板长度)、SEQ(即第十列,表示片段序列)以及QUAL(即,第十一列,表示具有基本错误概率的基本质量),并通过可选列的子字段进行解析,以计算源文件和目标文件中各行之间的差异。然后将列差异添加到所得的文件中,并且基于基因组的增量压缩程序110a、110b从源文件中读取下一行。下面将参照图3B更详细地描述用于基因组数据文件的示例性压缩过程300,包括用于SAM格式文件的用于层次识别的过程。

[0068] 由于行不匹配,因此可能无法执行全行比较。但是,如果执行了全行比较,则可能会产生以下结果:

[0069] 匹配(D←“m”),删除(D←“d”),插入(D←T(i)

[0070] 然而,如果源文件和目标文件为VCF格式,则每行中有八个固定和强制列。分层字段比较仅比较四列(即#CHROM、POS、REF和ALT)。基于基因组的增量压缩程序110a、110b首先比较目标文件和源文件中第一行的#CHROM和POS列,以确定#CHROM和POS列是否匹配。取决于基于基因组的增量压缩程序110a、110b是否确定#CHROM和POS列是中间匹配还是不匹配,然后将比较相同行的REF和ALT列。下面将参照图6更详细地描述用于识别VCF格式的基因组

数据文件的分层结构的示例性过程。

[0071] 如果在分层逐字段比较中比较的行匹配,则基于基因组的增量压缩程序110a、110b可执行全行比较,其包括在目标文件和源文件的行中的INFO列(即,代表其他信息的8列),并通过INFO列的子字段进行解析,以计算源文件和目标文件中的行之间的差异,以创建所得的文件。然后,基于基因组的增量压缩程序110a、110b从源文件读取下一行。下面将参照图3C更详细地描述用于基因组数据文件的示例性压缩过程300,包括用于VCF格式文件的用于层次识别的过程。

[0072] 然后在210,使用已知的通用文件压缩器压缩所得的增量文件。可以利用用于临时改变所得的D文件的代码来自动地将所得的增量文件(即,所得的D文件)传输到通用文件压缩器。在完成对所得D文件的压缩之后,可以将基于基因组的增量压缩程序110a、110b保存在用户设备(例如,用户计算机102)的存储器上。基于基因组的增量压缩程序110a、110b可以提示用户(例如,通过对话框)压缩的所得的D文件已完成。例如,该对话框可能包含一条消息,提示压缩后的所得的D文件已完成,并且该对话框下方有一个“查看详细信息”按钮。一旦用户单击“查看详细信息”按钮,该对话框就会消失,并且可能会向用户显示另一个带有压缩的D文件的对话框。生成的D文件可与源文件一起使用以重新生成目标文件。

[0073] 继续前面的示例,通过.zip压缩由源文件和目标文件的比较生成的所得的D文件(即,源文件被用于将目标文件重构为所得的D文件)。生成的增量文件减少了原始大小的52%,压缩后的生成的D文件被保存到用户计算机(即用户设备102)的存储器中。对话框提示用户压缩的所得的D文件已完成,并且用户单击消息下方的“查看详细信息”按钮。一旦用户单击“查看详细信息”按钮,则该对话框将消失,然后出现另一个对话框,该对话框将压缩后的D文件呈现给用户。

[0074] 现在参考图3A,该操作流程图示出了根据至少一个实施例的基于基因组的增量压缩程序110a和110b使用的基因组数据文件的示例性压缩过程300。如图所示,基于基因组的增量压缩程序110a、110b利用源文件和目标文件来创建D文件,并利用通用文件压缩器来压缩所得的D文件。

[0075] 在302,作为输入将源文件和目标文件输入到基于基因组的增量压缩程序110a、110b中。使用用户设备(例如,用户计算机102)上的软件程序108,可以将源文件和目标文件作为输入经由通信网络116传输到基于基因组的增量压缩程序110a、110b中。每个文件可以被标记以指示文件是源文件还是目标文件。这样,基于基因组的增量压缩程序110a、110b可能能够识别所传送的文件的类型(即,该文件是源文件还是目标文件)。

[0076] 例如,基于基因组的增量压缩程序110a、110b接收从基因组分析工具包(GATK)发送的源文件和目标文件。源文件包括来自与健康亚洲虎蚊相关的参考基因组的数据,目标文件包括与感染了西尼罗河病毒(即黄病毒)的亚洲虎蚊的基因组相关的数据。

[0077] 接下来,在304,基于基因组的增量压缩程序110a、110b检查文件以识别格式和启动缓冲器(INIT缓冲器)。基于基因组的增量压缩程序110a、110b可以检查源文件和目标文件是否具有兼容格式(例如,SAM或VCF)。然后,基于基因组的增量压缩程序110a、110b可以启动缓冲器,该缓冲器可以是可以存储来自目标文件的行的存储器的一小部分,直到基于基因组的增量压缩程序110a、110b找到来自源文件的匹配行为止。继续前面的示例,基于基因组的增量压缩程序110a、110b确认接收到的源文件和目标文件为VCF格式。另外,基于基

因组的增量压缩程序110a、110b启动缓冲器。

[0078] 然后,在306,从源文件读取一行。根据格式类型以及是否从流水线中检索文件,可能已经对各个文件的行进行了排序,因此,基于基因组的增量压缩程序110a、110b可以继续遍历(或读取)源文件。继续前面的示例,基于基因组的增量压缩程序110a、110b确定在传输到基于基因组的增量压缩程序110a、110b之前,源文件和目标文件已经在流水线中排序。这样,在基于基因组的增量压缩程序110a、110b读取接收到的源文件行之前,不必对接收到的源文件和目标文件进行排序。这样,基于基因组的增量压缩程序110a、110b可以继续从源文件读取第一行。

[0079] 如果在308处基于基因组的增量压缩程序110a、110b确定缓冲器为空(例如,缓冲器中没有目标文件),则基于基因组的增量压缩程序110a、110b可以在312处从文件中读取目标行。基于基因组的增量压缩程序110a、110b搜索缓冲器以确定缓冲器中是否有任何目标行。如果缓冲器中没有目标行,则基于基因组的增量压缩程序110a、110b确定缓冲器为空。

[0080] 然而,如果在308缓冲器不为空(例如,缓冲器包括至少一个目标行),则在310从缓冲器读取目标行。如果缓冲器中存在目标行,则基于基因组的增量压缩程序110a、110b确定缓冲器不为空。

[0081] 继续先前的示例,基于基因组的增量压缩程序110a、110b确定缓冲器不为空并且包括三个目标行。这样,从缓冲器读取目标行之一。

[0082] 如果在310中基于基因组的增量压缩程序110a、110b从缓冲器读取目标行,或者在312中从文件读取目标行,则在314基于基因组的增量压缩程序110a、110b确定用于源文件和目标文件的格式是否是SAM或VCF。在基于基因组的增量压缩程序110a、110b读取源文件和目标文件之后,基于基因组的增量压缩程序110a、110b进行创建和压缩所得的D文件,创建和压缩操作不同,具体取决于格式类型。如果在314,基于基因组的增量压缩程序110a、110b确定文件为SAM格式,则基于基因组的增量压缩程序110a、110b进行到图3B。然而,如果在314处基于基因组的增量压缩程序110a、110b确定文件为VCF格式,则基于基因组的增量压缩程序110a、110b进行至图3C。

[0083] 继续先前的示例,基于基因组的增量压缩程序110a、110b确认源文件和目标文件均处于VCF格式,然后基于基因组的增量压缩程序110a、110b进行入图3C。

[0084] 现在参考图3B,示出了根据至少一个实施例的基于基因组的增量压缩程序110a和110b使用的SAM格式的基因组数据文件的示例性压缩过程300的操作流程图。显示了SAM格式的源文件和目标文件行中字段的比较。在图3B中,源文件和目标文件都为SAM格式。

[0085] 如果在316处基于基因组的增量压缩程序110a、110b确定POS和REF不相同,则在318处,基于基因组的增量压缩程序110a、110b确定源文件的POS和REF是否小于目标文件的。基于基因组的增量压缩程序110a、110b可以评估SAM格式的源文件和目标文件的每一行,以确定POS和REF是否相同。

[0086] 继续先前的示例,如果源文件和目标文件是SAM格式,则基于基因组的增量压缩程序110a、110b可以评估源文件和目标文件的第一行,以确定每个文件的POS和REF文件是相同的。如果源文件和目标文件相同,则基于基因组的增量压缩程序110a、110b可以评估源文件和目标文件的QNAME是否相同。但是,如果源文件和目标文件的POS和REF不相同,则基于

基因组的增量压缩程序110a、110b可以确定源文件的POS和REF小于目标文件的。

[0087] 如果在318处基于基因组的增量压缩程序110a、110b确定源文件的POS和REF大于目标文件的,则在320处,基于基因组的增量压缩程序110a、110b在增量文件中插入目标记录,并在306返回从源读取的行。因此,如果源文件中行的POS和REF大于目标文件中行的,则目标记录可能会包含在所得的增量文件。继续前面的示例,如果源文件大于目标文件,则可以将目标记录插入增量中。因此,目标记录可以插入到所得的增量文件的该行中。

[0088] 然而,如果在318处基于基因组的增量压缩程序110a、110b确定源文件行的POS和REF小于目标文件中对应行的,则在322处,在增量文件中将对应行删除。增量和基于基因组的增量压缩程序110a、110b返回以在306从源文件读取下一行。继续前面的示例,如果源文件小于目标文件,则可以从产生的增量中删除源记录。这样,增量可能在所得增量文件的该行中而不是源记录中具有“d”。

[0089] 然而,如果在316,基于基因组的增量压缩程序110a、110b确定源文件中的行的POS和REF与目标文件中的相应行的相同,则在324,进行评估行以确定每行QNAME是否相同。继续前面的示例,如果接收到的源文件和目标文件的行具有相同的POS和REF,则基于基因组的增量压缩程序110a、110b可以确定接收到的源文件和目标文件的对应行的QNAME是否相同。

[0090] 如果在324处基于基因组的增量压缩程序110a、110b确定源文件行和目标文件行的QNAME不相同,则在326处,来自目标文件的行被保存到缓冲器中,并且基于基因组的增量压缩程序110a、110b返回以在312从文件读取下一个目标行。继续前面的示例,如果目标文件和源文件的相应行的QNAME不相同(即不同),目标行将保存在缓冲器中,直到确定与目标行的更好匹配为止。

[0091] 然而,如果在324处基于基因组的增量压缩程序110a、110b确定每一行的QNAME相同,则在328处,评估目标文件和源文件的行以确定FLAG是否是相同的。继续前面的示例,如果源文件和目标文件的相应行的QNAME相同,则基于基因组的增量压缩程序110a、110b将比较接收到的源文件和目标文件的相应行的FLAG(按位)。

[0092] 如果在328处基于基因组的增量压缩程序110a、110b确定每行的FLAG不相同,则在330处,基于基因组的增量压缩程序110a、110b确定差是否等于或小于阈值(例如0x400)。如果在330差大于阈值(例如0x400),则在326,将来自目标文件的行保存到缓冲器,并且基于基因组的增量压缩程序110a、110b返回312以读取文件中的目标行。继续前面的示例,如果接收到的源文件和目标文件的每个对应行的FLAG(按位)不同,则基于基因组的增量压缩程序110a、110b可以确定差异是否等于或小于阈值值。阈值可以由用户预先确定,并且在该示例中,阈值0x400(即,光学副本)由用户预先选择。这样,基于基因组的增量压缩程序110a、110b可以确定FLAG(逐位)差是否等于或小于0x400。

[0093] 然而,如果基于基因组的增量压缩程序110a、110b在330确定差等于或小于阈值(例如0x400),或者如果基于基因组的增量压缩程序110a、110b在328确定目标文件和源文件的行的FLAG相同,则在332处基于基因组的增量压缩程序110a、110b计算除了的OPTIONAL字段的列差。基于基因组的增量压缩程序110a,110b可以搜索列差异,然后指示哪些列包括这样的差异。根据不同列的位置和列差异的程度,所得的增量文件中的记录可能会受到影响。继续前面的示例,如果基于基因组的增量压缩程序110a、110b确定接收到的源文件和目

标文件的相应行之间的差异等于或小于0x400,或者确定接收到的源文件和目标文件的应的行的FLAG(按位)相同,则基于基因组的增量压缩程序110a、110b可以继续计算列差,不包括OPTIONAL字段。

[0094] 然后,在334,基于基因组的增量压缩程序110a、110b将OPTIONAL字段解析为子字段。OPTIONAL子字段可以包括在SAM格式文件的对齐记录中,该格式可以包括二十多个预定义的标准标签,这些标签可以由基于基因组的增量压缩程序110a、110b进行解析,以确定源文件和目标文件的相应行的OPTIONAL子字段之间的相似性或差异。继续前面的示例,基于基因组的增量压缩程序110a、110b可以在OPTIONAL字段和OPTIONAL字段内的子字段中的预定义标准标签中搜索,以确定接收到的源文件和目标文件相应行之间是否存在任何相似性或差异性。

[0095] 然后,在336,计算目标文件和源文件的每一行的子字段的差。基于基因组的增量压缩程序110a,110b然后可以针对目标和源文件的每一行搜索OPTIONAL子字段中的差异,并且可以指示针对目标和源文件的每一行哪些子字段是不同的。根据差异的程度,所得的增量文件中的记录可能会受到影响。继续前面的示例,基于基因组的增量压缩程序110a,110b可以为接收到的源文件和目标文件的相应行计算OPTIONAL子字段之间的差异。

[0096] 然后,在338,基于基因组的增量压缩程序110a,110b将列差异添加到增量文件中,并且在306,基于基因组的增量压缩程序110a,110b返回以从源文件读取下一行。基于基因组的增量压缩程序110a,110b继续循环进行用于基因组数据文件的示例性压缩过程300,直到最短的源文件和目标文件结束。继续前面的示例,基于基因组的增量压缩程序110a,110b将列差异添加到所得的增量文件中,并返回到306以读取接收到的源文件的下一个对应行。

[0097] 现在参考图3C,描绘了根据至少一个实施例的基于基因组的增量压缩程序110a和110b使用的VCF格式的基因组数据文件的示例性压缩过程300的操作流程图,示出了以VCF格式的源文件和目标文件行中的字段的比较。在图在图3C中,源文件和目标文件为VCF格式。

[0098] 如果在340处基于基因组的增量压缩程序110a、110b确定#CHROM和POS不相同,则在342处,基于基因组的增量压缩程序110a、110b确定源文件行的#CHROM和POS是否小于目标文件行的。如果在342处基于基因组的增量压缩程序110a、110b确定源文件行的#CHROM和POS大于目标文件行的,则在320处,基于基因组的增量压缩程序110a、110b在增量文件中插入目标记录,并在306返回从源文件读取行。

[0099] 然而,如果在342处基于基因组的增量压缩程序110a、110b确定源文件行的#CHROM和POS小于目标文件中对应行的,则在322处,在增量文件中删除对应行,并且在306,基于基因组的增量压缩程序110a、110b返回以在306从源文件读取下一行。

[0100] 然而,如果在340,基于基因组的增量压缩程序110a、110b确定源文件和目标文件的每一行的#CHROM和POS相同,则在344,基于基因组的增量压缩程序110a、110b确定源文件和目标文件的每一行的REF和ALT是否相同。继续前面的示例,先前确定接收到的源文件和目标文件为VCF格式。这样,基于基因组的增量压缩程序110a、110b前进到图3C,其中对接收到的源文件和目标文件的第一行进行评估以确定#CHROM和POS是否相同。接收到的源文件和目标文件对应行的#CHROM为21,每行POS为3。因此,源文件和目标文件对应行的#CHROM和POS相同,并且基于基因组的增量的压缩程序110a、110b进行到344以确定源文件行中的

REF、ALT是否与目标文件中对应行的相同。

[0101] 如果在344处基于基因组的增量压缩程序110a、110b确定源文件行和目标文件行的REF和ALT不同,则在326处,来自目标文件的行被保存到缓冲器中,然后,在312处,基于基因组的增量压缩程序110a、110b返回以从文件中读取目标行。

[0102] 但是,如果在344中基于基因组的增量压缩程序110a、110b确定源文件行和目标文件行的REF和ALT相同,则在346中,基于基因组的增量压缩程序110a、110b计算除INFO字段以外的列差。基于基因组的增量压缩程序110a、110b可以搜索列差异,然后指示哪些列包括这样的差异。根据不同列的位置和列差异的程度,所得的增量文件中的记录可能会受到影响。继续前面的示例,两行的REF为G,两行的ALT为C。这样,基于基因组的增量压缩程序110a、110b确定接收到的源文件和目标文件的两行的ALT和REF是相同的。由于源文件和目标文件行的REF和ALT相同,因此基于基因组的增量压缩程序110a、110b可以搜索VCF格式的源文件和目标文件中包括的所有八个强制固定列,以计算任何列差异。基于基因组的增量压缩程序110a、110b在ID和QUAL两个列中生成若干差异。

[0103] 然后,在348处,基于基因组的增量压缩程序110a、110b将INFO字段解析为子字段。INFO字段(即,VCF文件的第八列)可以包括关于源文件或目标文件的附加信息,可以对其进行搜索以确定源文件和目标文件的每一行之间的相似性和差异。继续前面的示例,基于基因组的增量压缩程序110a、110b搜索VCF格式的源文件和目标文件行的INFO字段或第八列。源文件行中每个INFO子字段的格式在元信息中指定为DP=154;MQ=52;H2。但是,目标文件行中每个INFO子字段的格式为DP=159;MQ=79;H2。

[0104] 然后,在350,基于基因组的增量压缩程序110a、110b比较目标文件和源文件的每一行的子字段,并在306返回以从源文件读取下一行。基于基因组的增量压缩程序110a、110b整个300继续循环,直到最短的源文件和目标文件结束。继续前面的示例,基于基因组的增量压缩程序110a、110b比较子字段以计算差异,并将其插入到生成的增量文件中。然后,基于基因组的增量压缩程序110a、110b返回到306以读取所接收的源文件的下一行。

[0105] 现在参考图4,示出了根据至少一个实施例的由基于基因组的增量压缩程序110a和110b使用的比较基因组数据文件400的示例性过程的框图。如图所示,用于基因组数据的文件402和404可以被组织为一组行和列。

[0106] 在402处的文件1是源文件,在404处的文件2是目标文件。当基于基因组的增量压缩程序110a、110b比较两个文件时,利用了基于402的映射位置与404的映射位置比较的逐行指示符。当将源文件406的一个记录中包含的数据与目标文件408中的另一个记录进行比较时,基于基因组的增量压缩程序110a、110b可以确定两个文件之间存在差异。尽管图4用颜色表示差异,该差异可能包括记录内的不同数据。这样,基于基因组的增量压缩程序110a、110b可以为所得的D文件写入以下结果:

[0107] 增量:<匹配行>

[0108] <第二列-替代><新值>

[0109] <匹配行>

[0110] <匹配行>

[0111] 因此,除了在408中的新值之外,源文件402和目标文件404是相同的,并且在比较行的四分之三中彼此匹配。

[0112] 现在参照图5,描绘了根据至少一个实施例的用于识别由基于基因组的增量压缩程序110a和110b使用的SAM格式500的基因组数据文件的分层结构的示例性过程的操作流程图。如图所示,当比较作为SAM记录的目标文件和源文件的行时,基于基因组的增量压缩程序110a、110b利用标识的分层系统。

[0113] 如果在502,基于基因组的增量压缩程序110a、110b确定该行的POS和REF不匹配(即,不匹配),则在504,确定目标文件和源文件不匹配,并且目标文件和源文件的比较移至源文件中的下一行。基于基因组的增量压缩程序110a、110b比较目标文件和源文件中每一行的POS和REF,以确定POS、REF是否相同。

[0114] 但是,如果基于基因组的增量压缩程序110a、110b在502确定POS和REF匹配,则在506基于基因组的增量压缩程序110a、110b确定目标和源文件行是中间匹配,在508基于基因组的增量压缩程序110a、110b向下移动该行以比较源文件和目标文件的下一行的QNAME。

[0115] 如果在508处基于基因组的增量压缩程序110a、110b确定行的QNAME不匹配,则在510处确定目标文件和源文件不匹配并且目标和源文件的行的比较被确定结束。但是,如果在508,基于基因组的增量压缩程序110a、110b确定QNAME匹配,则在512,基于基因组的增量压缩程序110a、110b确定目标文件和源文件行是中间匹配,在514处,基于基因组的增量压缩程序110a、110b向下移动该行以比较源文件和目标文件的FLAG(按位)。

[0116] 如果在514处基于基因组的增量压缩程序110a、110b确定行的FLAG(按位)匹配,则在516处基于基因组的增量压缩程序110a、110b确定目标和源文件的行匹配,进行到图3B中的更复杂的操作流程图。

[0117] 然而,如果在514处基于基因组的增量压缩程序110a、110b确定FLAG(逐位)不匹配,则在518处基于基因组的增量压缩程序110a、110b确定源文件行和目标文件行出现中间不匹配,并且在520处基于基因组的增量压缩程序110a、110b可以比较行以确定该差是否等于或小于阈值(例如0x400)。如果在520处差值等于或者小于阈值,则在522处目标文件行和源文件行被认为匹配,并且将用于匹配的“m”写入所得的增量文件中。然而,如果在520处该差大于阈值,则在524确定源文件和目标文件行不匹配,并且源文件和目标文件行的层次标识结束。

[0118] 现在参照图6,描绘了根据至少一个实施例的由基于基因组的增量压缩程序110a和110b使用的用于识别VCF格式600的基因组数据文件的层次结构的简单示例过程的操作流程图。如图所示,当比较VCF记录的目标文件和源文件的行时,基于基因组的增量压缩程序110a、110b利用识别的分层系统。

[0119] 如果在602处基于基因组的增量压缩程序110a、110b确定该行的#CHROM和POS不匹配(即,不匹配),则在604处将目标文件和源文件确定为不匹配,并且目标文件和源文件的比较结束。基于基因组的增量压缩程序110a、110b比较目标文件和源文件中每行的#CHROM和POS。

[0120] 然而,如果在602处基于基因组的增量压缩程序110a、110b确定#CHROM和POS匹配,则在606处目标和源文件行被确定为中间匹配,并且基于基因组的增量压缩程序110a、110b在608处向下移动该行以比较源文件和目标文件的下一行的REF和ALT。

[0121] 如果在608处基于基因组的增量压缩程序110a、110b确定行的REF和ALT匹配,则在612处基于基因组的增量压缩程序110a、110b确定目标文件行和源文件行匹配并且目标文

件行和源文件行的比较结束。匹配的“m”会写入所得的增量文件中。

[0122] 然而,如果在608处基于基因组的增量压缩程序110a、110b确定REF和ALT不匹配,则在610处基于基因组的增量压缩程序110a、110b确定目标文件行和源文件行不匹配,并且源文件和目标文件的层次标识结束。

[0123] 可以认识到,图2-6仅提供了一个实施例的图示,并不暗示关于可以如何实现不同实施例的任何限制。可以基于设计和实施要求对所描绘的实施例进行许多修改。

[0124] 图7是根据本发明的说明性实施例的图1中描绘的计算机的内部和外部组件的框图900。应该理解的是,图7仅提供了一种实现方式的图示,并不暗示对可以实现不同实施例的环境的任何限制。可以基于设计和实现要求对所描绘的环境进行许多修改。

[0125] 数据处理系统902、904代表能够执行机器可读程序指令的任何电子设备。数据处理系统902、904可以代表智能电话、计算机系统、PDA或其他电子设备。可以由数据处理系统902、904表示的计算系统、环境和/或配置的示例包括但不限于个人计算机系统、服务器计算机系统、瘦客户端、胖客户端、手持或膝上型设备、多处理器系统、基于微处理器的系统、网络PC、小型计算机系统以及包括上述任何系统或设备的分布式云计算环境。

[0126] 用户客户端计算机102和网络服务器112可以包括图7中所示的内部组件902a、b和外部组件904a、b的相应集合。内部组件组902a、b中的每一组都包括一个或多个总线912上的一个或多个处理器906、一个或多个计算机可读RAM 908和一个或多个计算机可读ROM 910、以及一个或多个操作系统914和一个或多个计算机可读有形存储设备916。客户端计算机102中的一个或多个操作系统914、软件程序108和基于基因组的增量压缩程序110a,以及在网络服务器112中的基于基因组的增量压缩程序110b可以存储在一个或多个计算机可读有形存储设备916上,以由一个或多个处理器906经由一个或多个RAM 908(通常包括高速缓冲存储器)来执行。在图7所示的实施例中,每个计算机可读有形存储设备916是内部硬盘驱动器的磁盘存储设备。备选地,每个计算机可读有形存储设备916是半导体存储设备,例如ROM 910、EPROM、闪存或可以存储计算机程序和数字信息的任何其他计算机可读有形存储设备。

[0127] 每组内部组件902a、b还包括R/W驱动器或接口918,以从一个或多个便携式计算机可读有形存储设备920(例如CD-ROM、DVD、存储棒、磁带、磁盘、光盘或半导体存储设备)中读取和写入。可以将诸如软件程序108和基于基因组的增量压缩程序110a和110b之类的软件程序存储在相应的便携式计算机可读有形存储设备920中的一个或多个上,通过相应的R/W驱动器或接口918读取,并加载到相应的硬盘驱动器916中。

[0128] 每组内部组件902a、b还可以包括网络适配器(或交换机端口卡)或例如TCP/IP适配器卡、无线Wi-Fi接口卡或3G或4G无线接口卡的接口922,或其他有线或无线通信链接。客户端计算机102中的软件程序108和基于基因组的增量压缩程序110a以及网络服务器计算机112中的基于基因组的增量压缩程序110b可以经由网络(例如,互联网、局域网或其他广域网)和相应的网络适配器或接口922从外部计算机(例如,服务器)下载。从网络适配器(或交换机端口适配器)或接口922,客户计算机102中的软件程序108和基于基因组的增量压缩程序110a和网络服务器计算机112中的基于基因组的增量压缩程序110b被加载到相应的硬盘驱动器916中。网络可以包括铜线、光纤、无线传输、路由器、防火墙、交换机、网关计算机和/或边缘服务器。

[0129] 每组外部组件904a、b可以包括计算机显示监视器924、键盘926和计算机鼠标928。外部组件904a、b还可以包括触摸屏、虚拟键盘、触摸板、指点设备和其他人机界面设备。内部组件组902a、b中的每组还包括设备驱动程序930以与计算机显示监视器924、键盘926和计算机鼠标928进行接口。设备驱动程序930、R/W驱动器或接口918,以及网络适配器或接口922包括硬件和软件(存储在存储设备916和/或ROM 910中)。

[0130] 预先理解的是,尽管本公开包括关于云计算的详细描述,但是本文叙述的教导的实施方式不限于云计算环境。相反,本发明的实施例能够与现在已知或以后开发的任何其他类型的计算环境结合实现。

[0131] 云计算是一种服务交付模式,用于对共享的可配置计算资源池进行方便、按需的网络访问。可配置计算资源是能够以最小的管理成本或与服务提供者进行最少的交互就能快速部署和释放的资源,例如可以是网络、网络带宽、服务器、处理、内存、存储、应用、虚拟机和服务。这种云模式可以包括至少五个特征、至少三个服务模型和至少四个部署模型。

[0132] 特征包括:

[0133] 按需自助式服务:云的消费者在无需与服务提供者进行人为交互的情况下能够单方面自动地按需部署诸如服务器时间和网络存储等的计算能力。

[0134] 广泛的网络接入:计算能力可以通过标准机制在网络上获取,这种标准机制促进了通过不同种类的瘦客户机平台或厚客户机平台(例如移动电话、膝上型电脑、个人数字助理PDA)对云的使用。

[0135] 资源池:提供者的计算资源被归入资源池并通过多租户(multi-tenant)模式服务于多重消费者,其中按需将不同的实体资源和虚拟资源动态地分配和再分配。一般情况下,消费者不能控制或甚至并不知晓所提供的资源的确切位置,但可以在较高抽象程度上指定位置(例如国家、州或数据中心),因此具有位置无关性。

[0136] 迅速弹性:能够迅速、有弹性地(有时是自动地)部署计算能力,以实现快速扩展,并且能迅速释放来快速缩小。在消费者看来,用于部署的可用计算能力往往显得是无限的,并能在任意时候都能获取任意数量的计算能力。

[0137] 可测量的服务:云系统通过利用适于服务类型(例如存储、处理、带宽和活跃用户帐号)的某种抽象程度的计量能力,自动地控制和优化资源效用。可以监测、控制和报告资源使用情况,为服务提供者和消费者双方提供透明度。

[0138] 服务模型如下:

[0139] 软件即服务(SaaS):向消费者提供的能力是使用提供者在云基础架构上运行的应用。可以通过诸如网络浏览器的瘦客户机接口(例如基于网络的电子邮件)从各种客户机设备访问应用。除了有限的特定于用户的应用配置设置外,消费者既不管理也不控制包括网络、服务器、操作系统、存储、乃至单个应用能力等的底层云基础架构。

[0140] 平台即服务(PaaS):向消费者提供的能力是在云基础架构上部署消费者创建或获得的应用,这些应用利用提供者支持的程序设计语言和工具创建。消费者既不管理也不控制包括网络、服务器、操作系统或存储的底层云基础架构,但对其部署的应用具有控制权,对应用托管环境配置可能也具有控制权。

[0141] 基础架构即服务(IaaS):向消费者提供的能力是消费者能够在其中部署并运行包括操作系统和应用的任意软件的处理、存储、网络和其他基础计算资源。消费者既不管理也

不控制底层的云基础架构,但是对操作系统、存储和其部署的应用具有控制权,对选择的网络组件(例如主机防火墙)可能具有有限的控制权。

[0142] 部署模型如下:

[0143] 私有云:云基础架构单独为某个组织运行。云基础架构可以由该组织或第三方管理并且可以存在于该组织内部或外部。

[0144] 共同体云:云基础架构被若干组织共享并支持有共同利害关系(例如任务使命、安全要求、政策和合规考虑)的特定共同体。共同体云可以由共同体内的多个组织或第三方管理并且可以存在于该共同体内部或外部。

[0145] 公共云:云基础架构向公众或大型产业群提供并由出售云服务的组织拥有。

[0146] 混合云:云基础架构由两个或更多部署模型的云(私有云、共同体云或公共云)组成,这些云依然是独特的实体,但是通过使数据和应用能够移植的标准化技术或私有技术(例如用于云之间的负载平衡的云突发流量分担技术)绑定在一起。

[0147] 云计算环境是面向服务的,特点集中在无状态性、低耦合性、模块性和语意的互操作性。云计算的核心是包含互连节点网络的基础架构。

[0148] 现在参考图8,描绘了说明性的云计算环境1000。如图所示,云计算环境1000包括一个或多个云计算节点100,由云消费者使用的本地计算设备,例如个人数字助理(PDA)或蜂窝电话1000A、台式计算机1000B、膝上型计算机1000C、和/或汽车计算机系统1000N可以通信。节点100可以彼此通信。可以在一个或多个网络(如上文所述的私有云、共同体云、公共云或混合云)或其组合中对它们进行物理或虚拟分组(未显示)。这允许云计算环境1000提供基础设施、平台和/或软件作为服务,云消费者不需要为其维护本地计算设备上的资源。应当理解,图8中所示的计算设备1000A-N的类型仅旨在说明,并且计算节点100和云计算环境1000可以通过任何类型的网络和/或网络可寻址连接(例如,使用网络浏览器)与任何类型的计算机化设备通信。

[0149] 现在参考图9,其中显示了云计算环境1000提供的一组功能抽象层1100。首先应当理解,图9所示的组件、层以及功能都仅仅是示意性的,本发明的实施例不限于此。如图所示,提供下列层和对应功能:

[0150] 硬件和软件层1102包括硬件和软件组件。硬件组件的例子包括:主机1104;基于RISC(精简指令集计算机)体系结构的服务器1106;服务器1108;刀片服务器1110;存储设备1112;网络和网络组件1114。软件组件的例子包括:网络应用服务器软件1116以及数据库软件1118。

[0151] 虚拟层1120提供一个抽象层,该层可以提供下列虚拟实体的例子:虚拟服务器1122、虚拟存储1124、虚拟网络1126(包括虚拟私有网络)、虚拟应用和操作系统1128,以及虚拟客户端1130。

[0152] 在一个示例中,管理层1132可以提供下述功能:资源供应功能1134:提供用于在云计算环境中执行任务的计算资源和其它资源的动态获取;计量和定价功能1136:在云计算环境内对资源的使用进行成本跟踪,并为此提供帐单和发票。在一个例子中,该资源可以包括应用软件许可。安全功能:为云的消费者和任务提供身份认证,为数据和其它资源提供保护。用户门户功能1138:为消费者和系统管理员提供对云计算环境的访问。服务水平管理功能1140:提供云计算资源的分配和管理,以满足必需的服务水平。服务水平协议(SLA)计划

和履行功能1142:为根据SLA预测的对云计算资源未来需求提供预先安排和供应。

[0153] 工作负载层1144提供云计算环境可能实现的功能的示例。在该层中,可提供的工作负载或功能的示例包括:地图绘制与导航1146;软件开发及生命周期管理1148;虚拟教室的教学提供1150;数据分析处理1152;交易处理1154以及基于基因的增量压缩1156。基于基因组的增量压缩程序110a、110b提供了一种为基因组数据文件压缩增量文件的方法。

[0154] 以上已经描述了本发明的各实施例,上述说明是示例性的,并非穷尽性的,并且也不限于所披露的各实施例。在不偏离所说明的各实施例的范围的情况下,对于本技术领域的普通技术人员来说许多修改和变更都是显而易见的。本文中所用术语的选择,旨在最好地解释各实施例的原理、实际应用或对市场中的技术的技术改进,或者使本技术领域的其它普通技术人员能理解本文披露的各实施例。

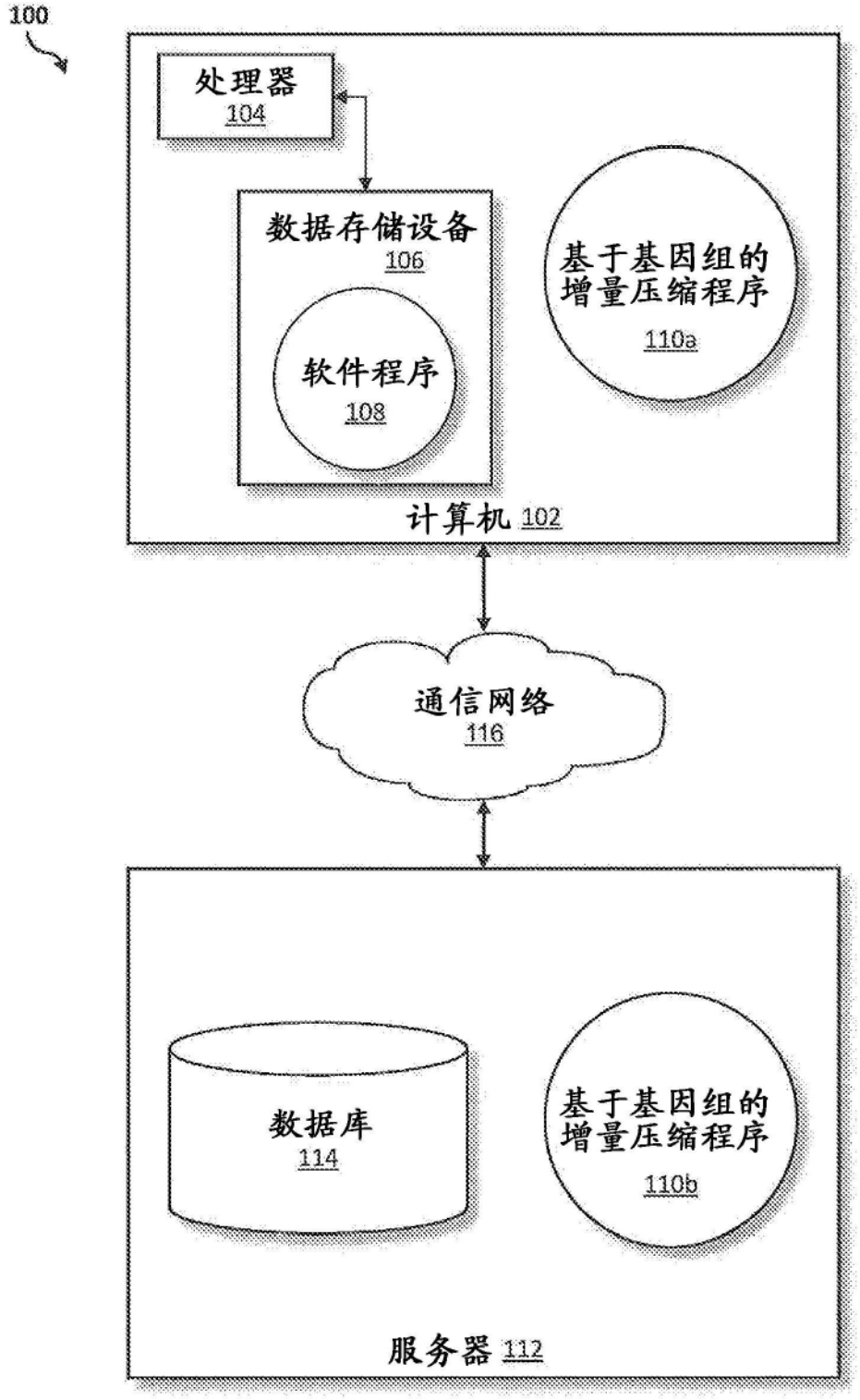

100
处理器
104
数据存储设备
106
软件程序
108
基于基因组的
增量压缩程序
110a
计算机 102
通信网络
116
数据库
114
基于基因组的
增量压缩程序
110b
服务器 112

图1

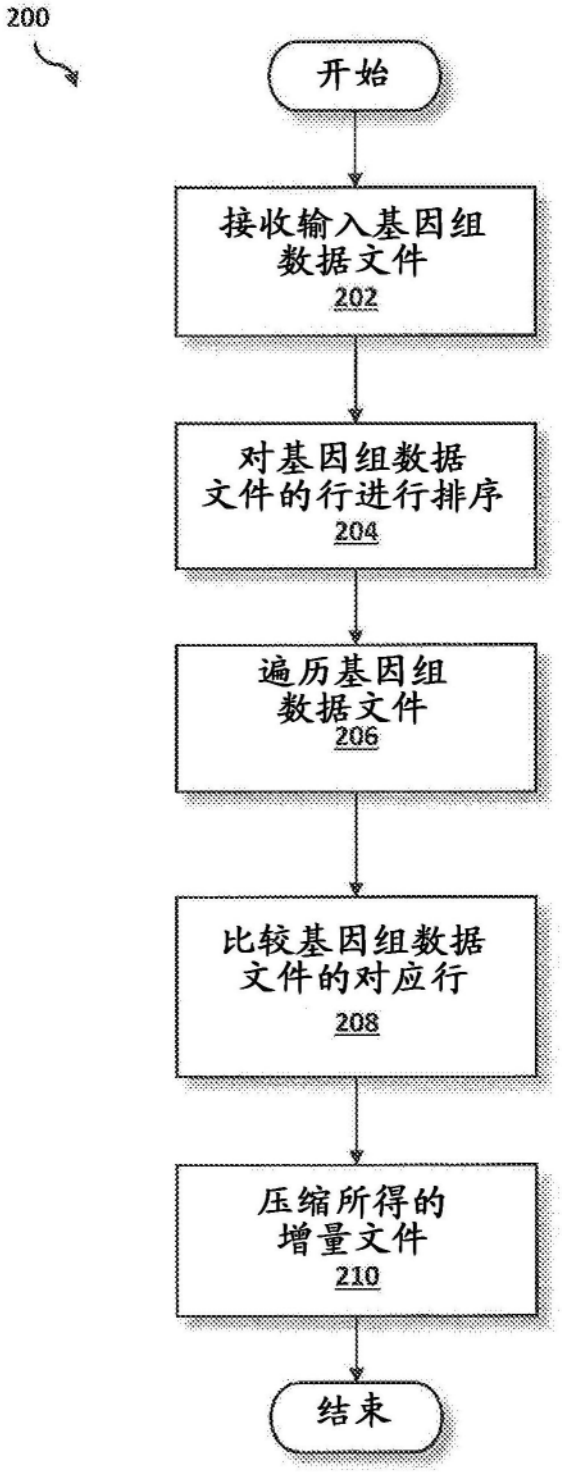

200
开始
接收输入基因组数据文件
202
对基因组数据文件的行进行排序
204
遍历基因组数据文件
206
比较基因组数据文件的对应行
208
压缩所得的增量文件
210
结束

图2

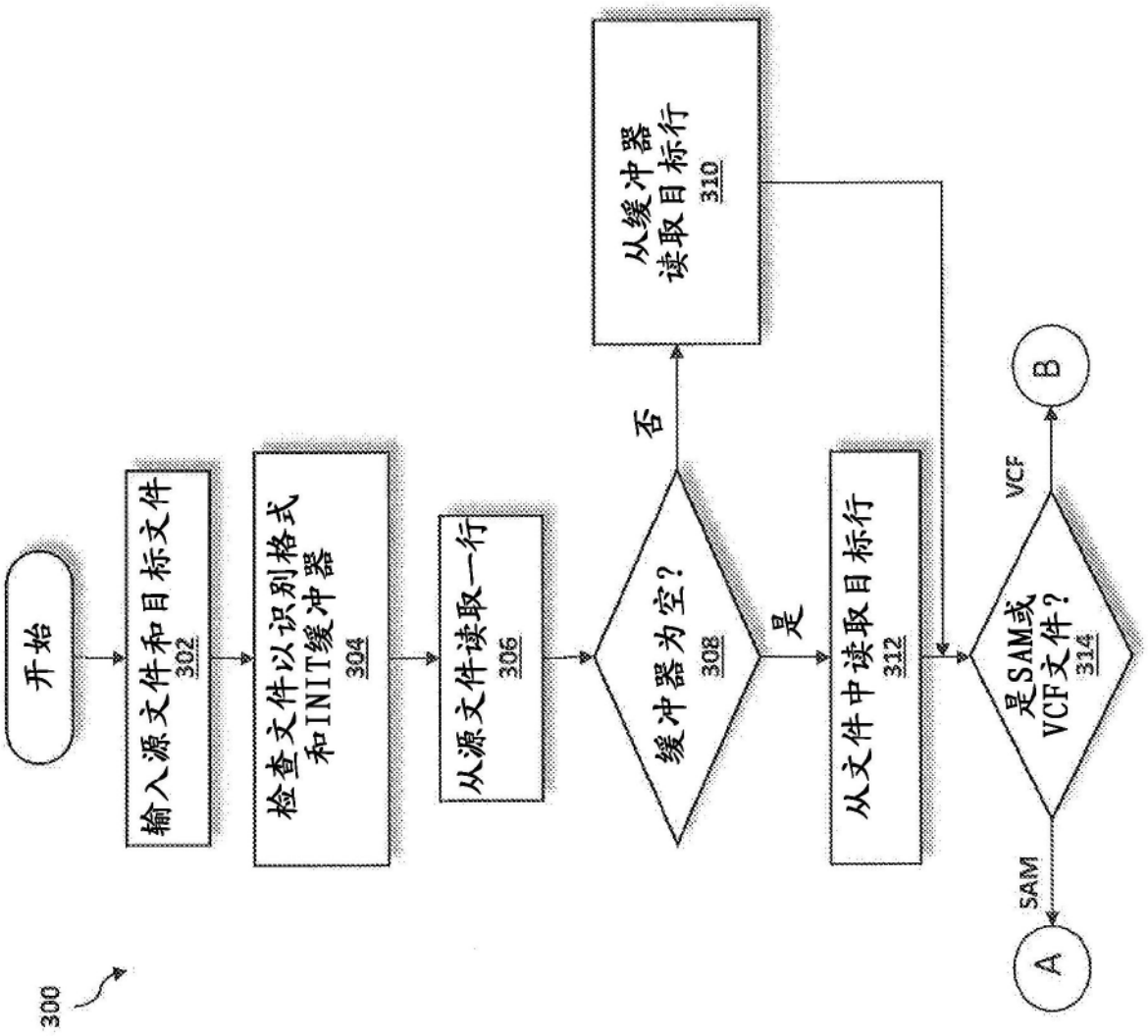

300
开始
输入源文件和目标文件
302
检查文件以识别格式
和INIT缓冲器
304
从源文件读取一行
306
缓冲器为空?
308
否
从缓冲器
读取目标行
310
是
从文件中读取目标行
312
是SAM或
VCF文件?
314
VCF
B
SAM
A

图3A

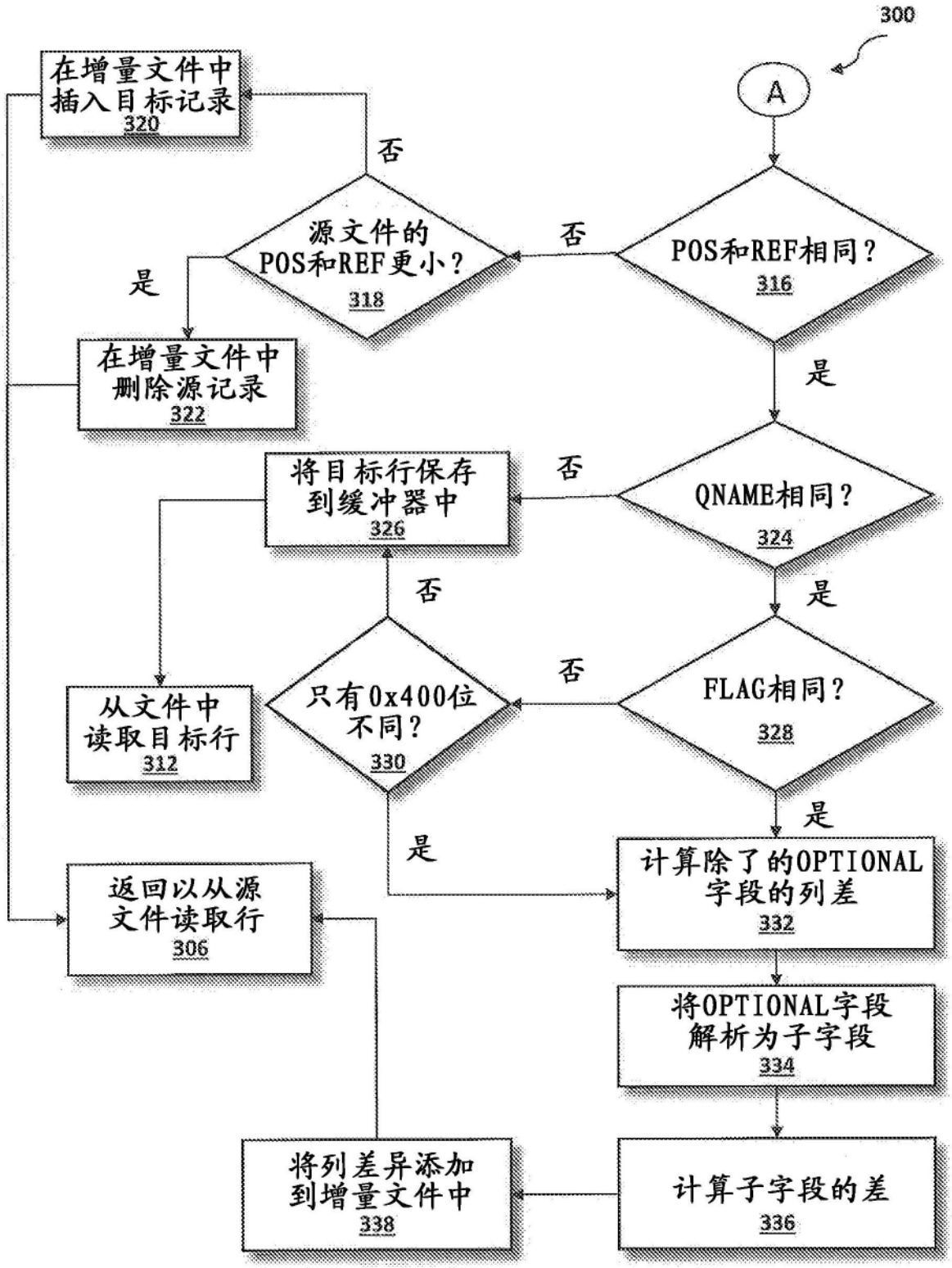

300
A
POS和REF相同？
316
否
是
源文件的POS和REF更小？
318
否
是
在增量文件中插入目标记录
320
在增量文件中删除源记录
322
QNAME相同？
324
否
是
将目标行保存到缓冲器中
326
否
是
FLAG相同？
328
否
是
只有0x400位不同？
330
否
是
从文件中读取目标行
312
返回以从源文件读取行
306
计算除了的OPTIONAL字段的列差
332
将OPTIONAL字段解析为子字段
334
计算子字段的差
336
将列差异添加到增量文件中
338

图3B

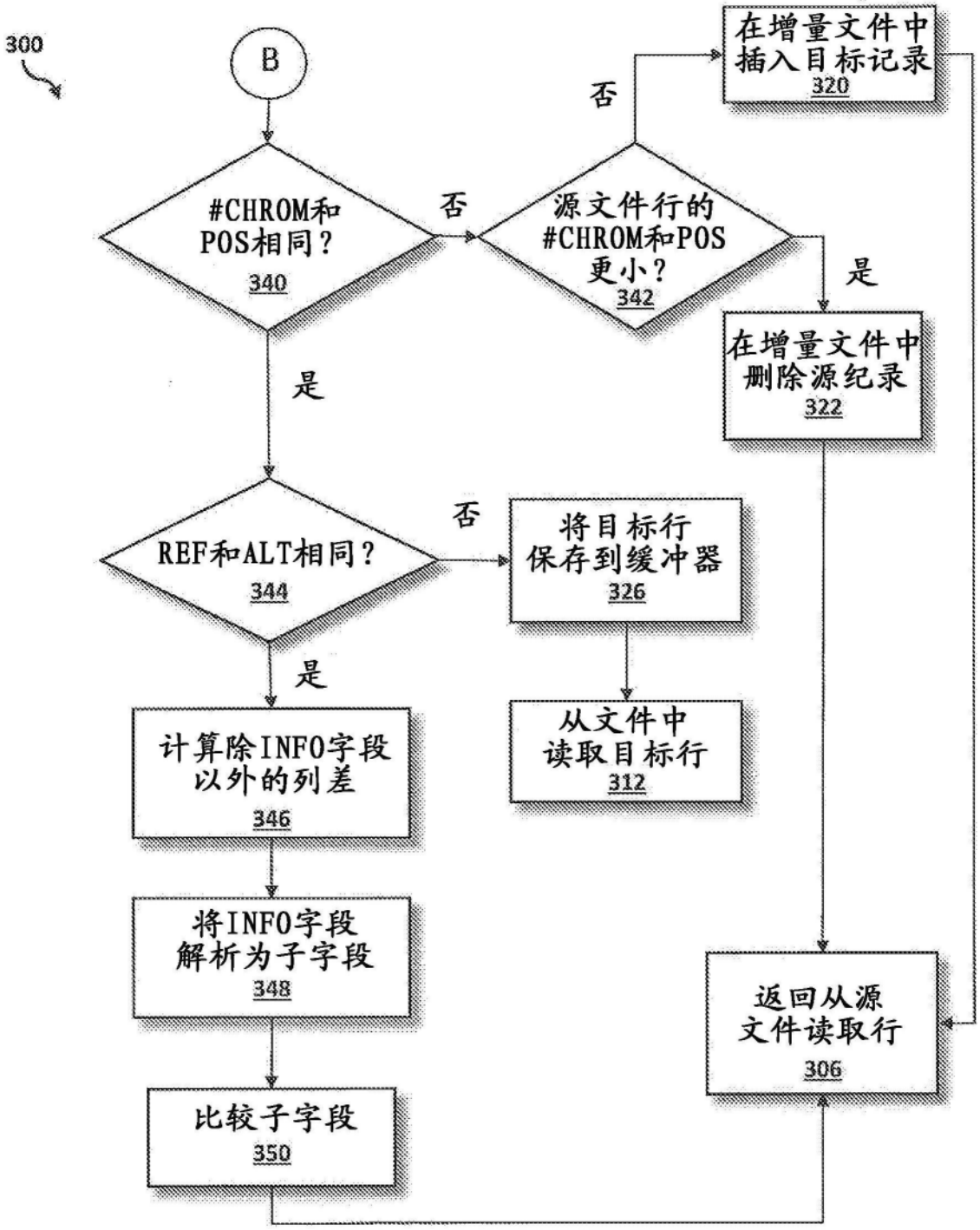

300
B
#CHROM和POS相同？
340
否
源文件行的#CHROM和POS更小？
342
否
在增量文件中插入目标记录
320
是
在增量文件中删除源纪录
322
是
REF和ALT相同？
344
否
将目标行保存到缓冲器
326
是
计算除INFO字段以外的列差
346
从文件中读取目标行
312
将INFO字段解析为子字段
348
比较子字段
350
返回从源文件读取行
306

图3C

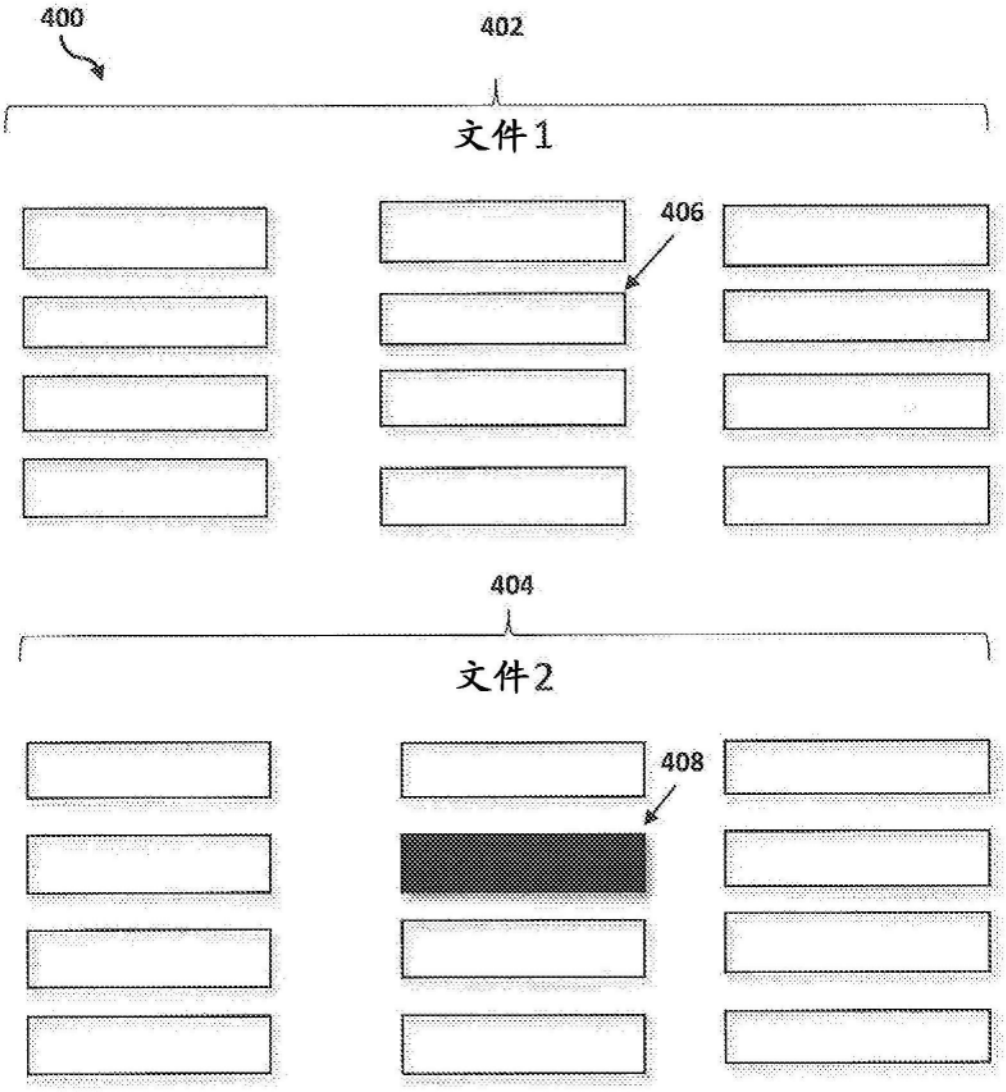


图4

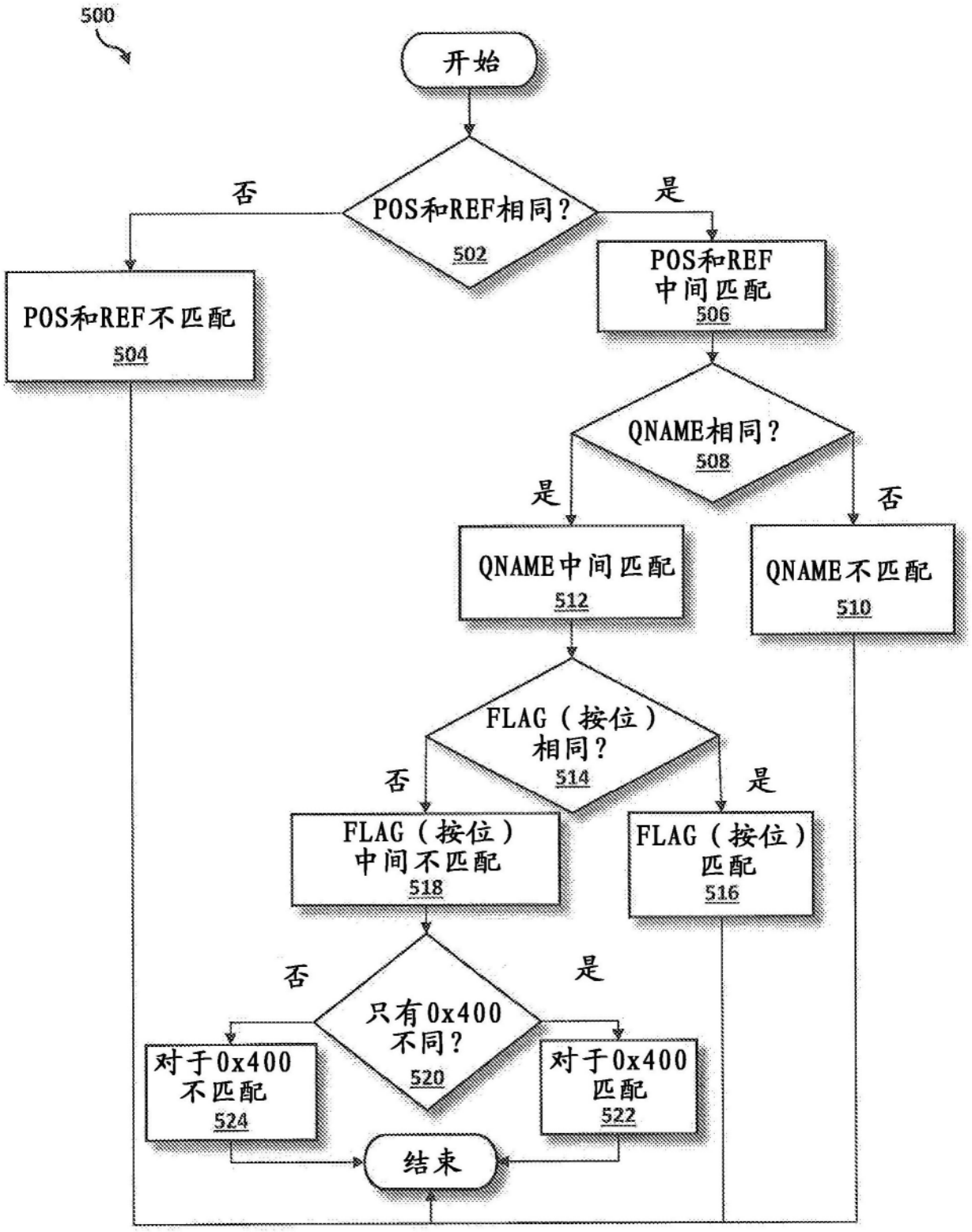

500
开始
POS和REF相同？
502
否
是
POS和REF不匹配
504
POS和REF
中间匹配
506
QNAME相同？
508
是
否
QNAME中间匹配
512
QNAME不匹配
510
FLAG（按位）
相同？
514
否
是
FLAG（按位）
中间不匹配
518
FLAG（按位）
匹配
516
只有0x400
不同？
520
否
是
对于0x400
不匹配
524
对于0x400
匹配
522
结束

图5

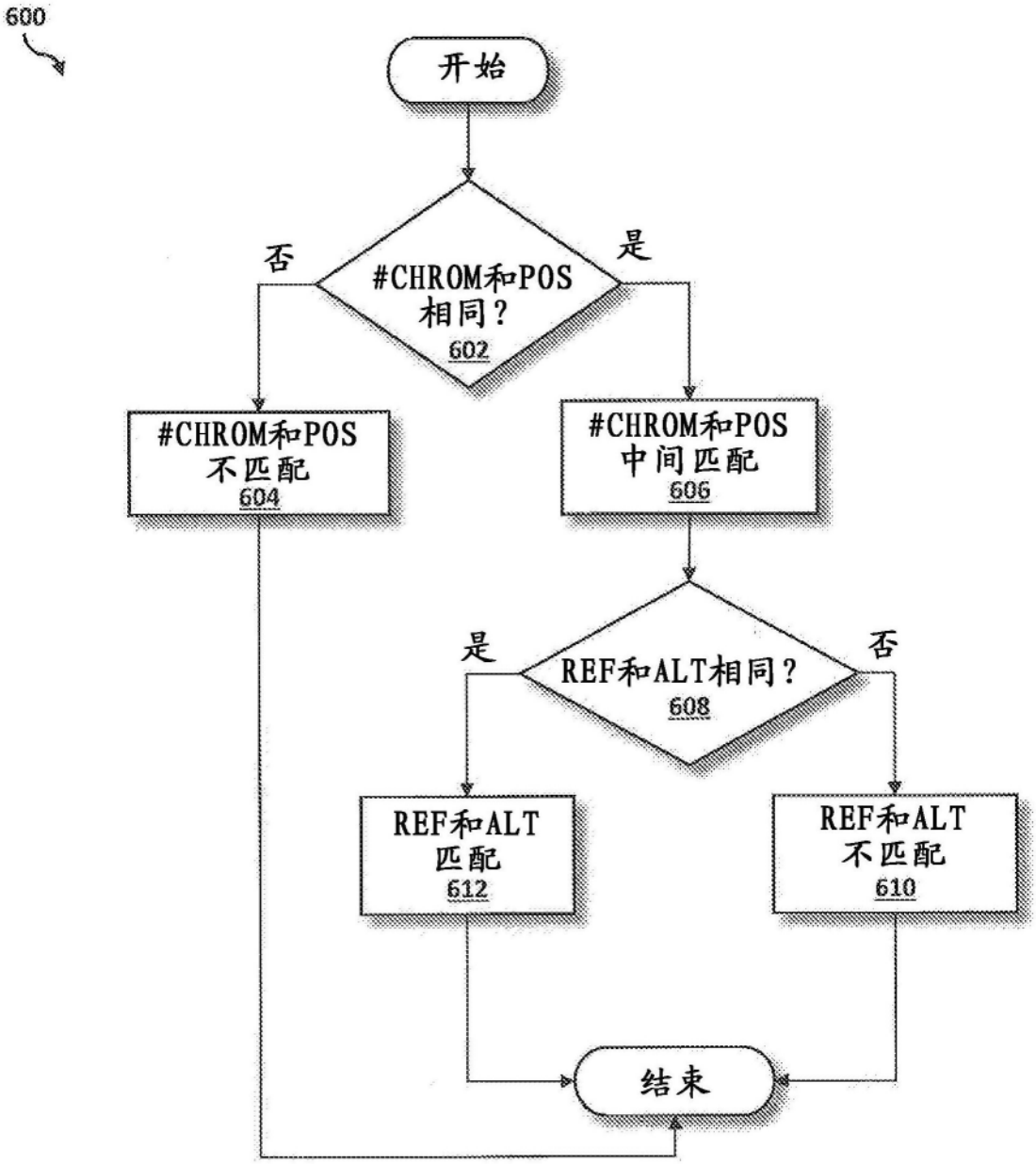

600
开始
#CHROM和POS
相同?
602
否
是
#CHROM和POS
不匹配
604
#CHROM和POS
中间匹配
606
是
REF和ALT相同?
608
否
REF和ALT
匹配
612
REF和ALT
不匹配
610
结束

图6

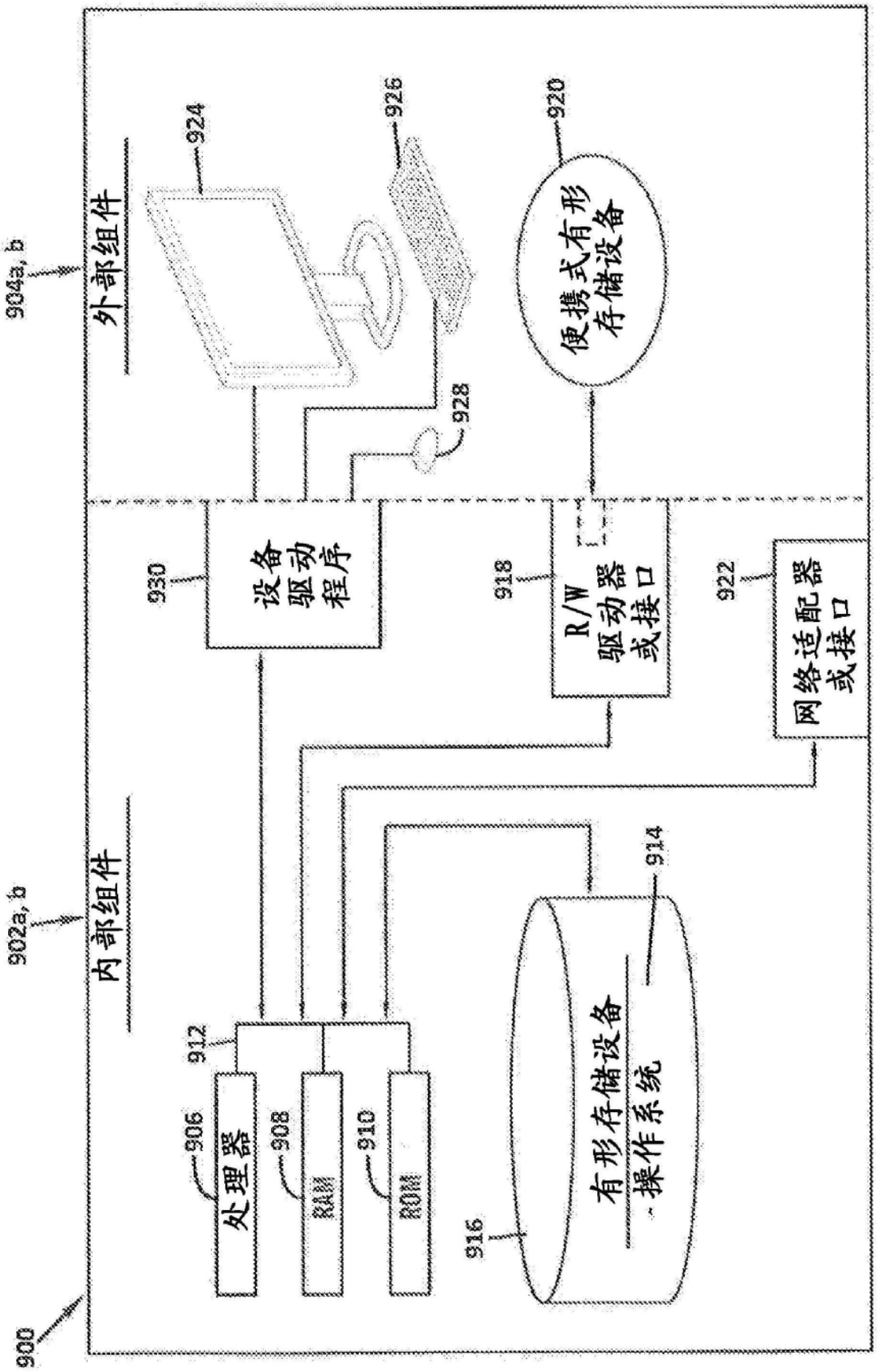

900
902a, b
内部组件
904a, b
外部组件
906
处理器
908
RAM
910
ROM
912
914
有形存储设备
操作系统
916
918
R/W
驱动器
或接口
920
便携式有形
存储设备
922
网络适配器
或接口
924
925
928
930
设备
驱动
程序

图7

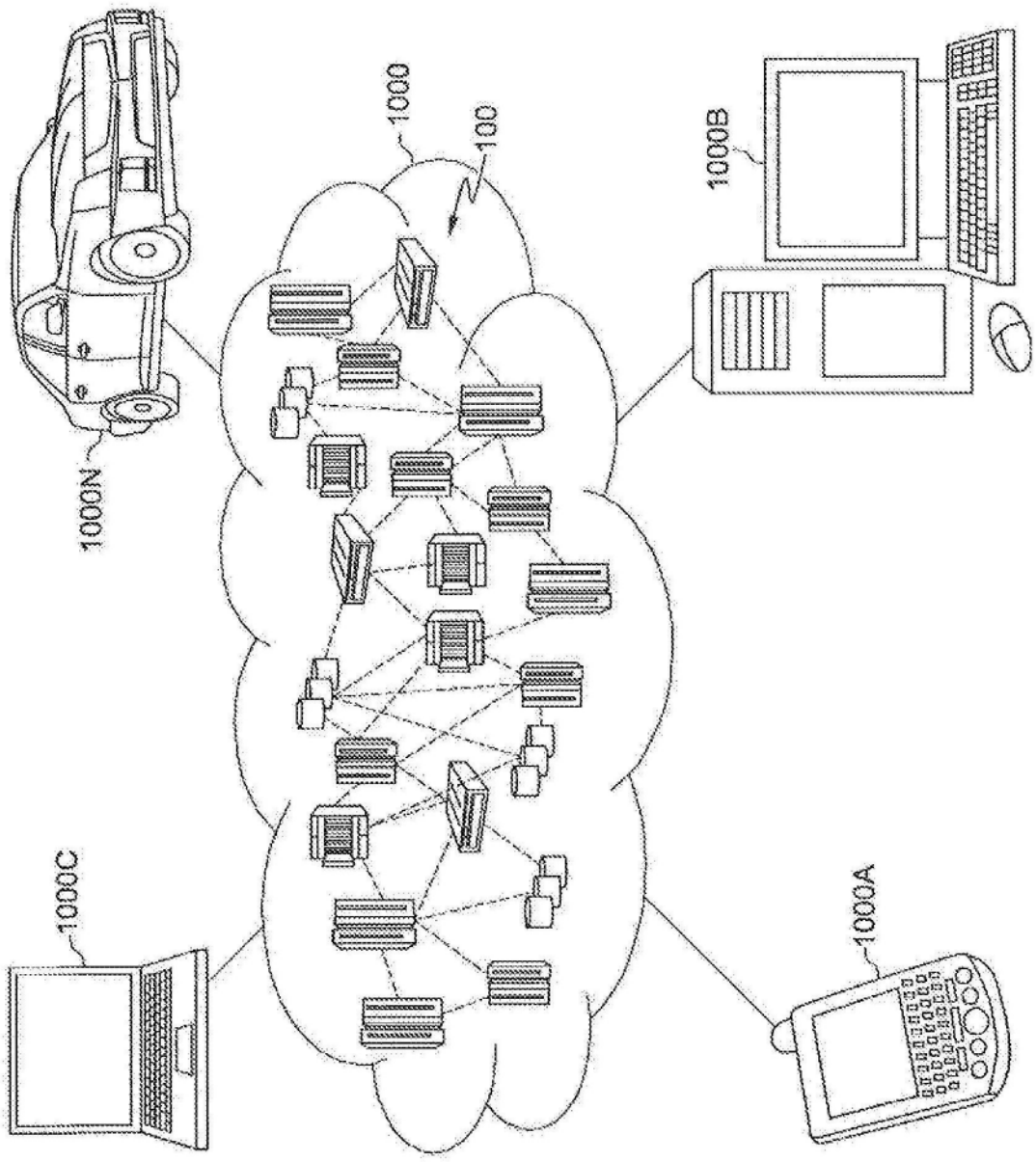


图8

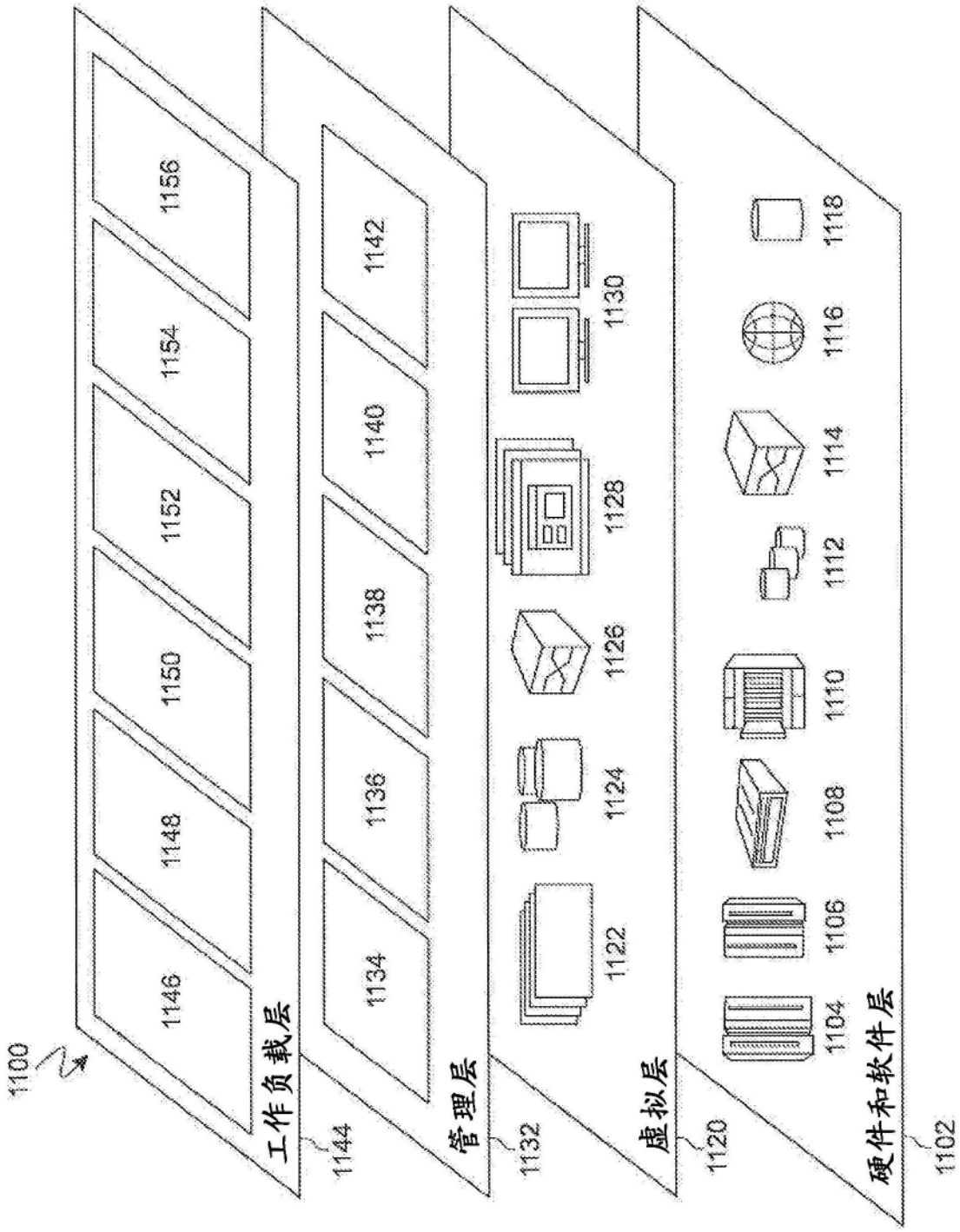

1100
工作负载层
1144
1146
1148
1150
1152
1154
1156
1134
1136
1138
1140
1142
管理层
1132
1122
1124
1126
1128
1130
虚拟层
1120
1104
1106
1108
1110
1112
1114
1116
1118
硬件和软件层
1102

图9