

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局

(43) 国际公布日
2024 年 12 月 26 日 (26.12.2024)



(10) 国际公布号
WO 2024/259679 A1

- (51) 国际专利分类号:
G16B 30/10 (2019.01) **G06F 16/10** (2019.01)
- (21) 国际申请号: PCT/CN2023/101916
- (22) 国际申请日: 2023 年 6 月 21 日 (21.06.2023)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (71) 申请人: 深圳华大生命科学研究院 (**BGI SHENZHEN**) [CN/CN]; 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。湖北华大基因研究院 (**BGI HUBEI**) [CN/CN]; 中国湖北省武汉市东湖新技术开发区高新大道666号, Hubei 430075 (CN)。
- (72) 发明人: 袁东 (**YUAN, Dong**); 中国湖北省武汉市东湖新技术开发区高新大道666号, Hubei 430075 (CN)。龚淳 (**GONG, Chun**); 中国湖北省武汉市东湖新技术开发区高新大道666号, Hubei 430075 (CN)。李胜康 (**LI, Shengkang**); 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。张勇 (**ZHANG, Yong**); 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。

(CN)。黎宇翔 (**LI, Yuxiang**); 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。陈奥 (**CHEN, Ao**); 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。

(74) 代理人: 上海弼兴律师事务所 (**SHANGHAI BESHINING LAW OFFICE**); 中国上海市徐汇区小木桥路 681 号外经大厦 21 楼, Shanghai 200032 (CN)。

(81) 指定国(除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN, MU, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW。

(84) 指定国(除另有指明, 要求每一种可提供的地区保护): ARIPO (BW, CV, GH, GM, KE, LR, LS, MW, MZ,

(54) **Title:** SUB-SEQUENCE SEARCH METHOD FOR GENE SEQUENCE, COMPARISON METHOD, SYSTEM, AND DEVICE

(54) 发明名称: 基因序列的子序列搜索方法、比对方法、系统及设备

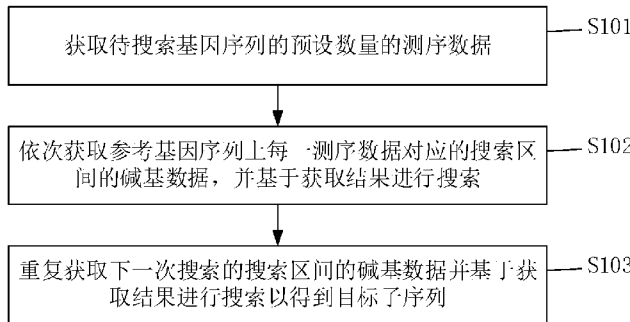


图 1

- S101 Acquire a preset number of pieces of sequencing data of a gene sequence to be searched
- S102 Sequentially acquire, from a reference gene sequence, base data of a search interval corresponding to each piece of sequencing data, and perform a search on the basis of an acquisition result
- S103 Repeatedly acquire base data of a search interval of the next search, and perform a search on the basis of an acquisition result, so as to obtain a target sub-sequence

(57) **Abstract:** Disclosed in the present invention are a sub-sequence search method for a gene sequence, a comparison method, a system, and a device. The sub-sequence search method comprises: acquiring a preset number of pieces of sequencing data of a gene sequence to be searched; sequentially acquiring, from a reference gene sequence, base data of a search interval corresponding to each piece of sequencing data, and performing a search on the basis of an acquisition result, wherein the search interval corresponding to each piece of sequencing data is determined on the basis of a base sequence to be searched of each piece of sequencing data; and repeatedly acquiring base data of a search interval of the next search, and performing a search on the basis of an acquisition result, so as to obtain a target sub-sequence, wherein the search interval of the next search is determined according to the next base to be searched of each piece of sequencing data and a search interval of the previous search. In the search method in the present invention, base data of a search interval is pre-fetched, and by means of batch processing of sequencing data, the calculation of the other sequencing data can be processed when waiting for a pre-fetching result, thereby improving the efficiency of a gene search.

NA, RW, SC, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, ME, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

根据细则4.17的声明:

- 关于申请人有权申请并被授予专利(细则4.17(ii))
- 发明人资格(细则4.17(iv))

本国际公布:

- 包括国际检索报告(条约第21条(3))。

(57) 摘要: 本发明公开了一种基因序列的子序列搜索方法、比对方法、系统及设备, 所述子序列搜索方法包括: 获取待搜索基因序列的预设数量的测序数据; 依次获取参考基因序列上每一所述测序数据对应的搜索区间的碱基数据, 并基于获取结果进行搜索; 每一所述测序数据对应的搜索区间基于每一所述测序数据的待搜索碱基序列确定; 重复获取下一次搜索的搜索区间的碱基数据并基于获取结果进行搜索以得到目标子序列; 所述下一次搜索的搜索区间根据每一所述测序数据下一个待搜索碱和上一次搜索的搜索区间确定。本发明的搜索方法, 预先获取搜索区间的碱基数据, 通过批量处理测序数据, 在等待预取结果时可以处理其他测序数据的计算, 提高了基因搜索的效率。

基因序列的子序列搜索方法、比对方法、系统及设备

技术领域

本发明涉及核酸信息领域，特别涉及一种基因序列的子序列搜索方法、比对方法、系统及设备。

背景技术

基因序列是包含 ACGT 四种碱基的长字符串，基因序列比对的一种方式是将短基因序列与参考基因序列对比，找出短基因序列在参考基因序列中的位置及两者间差异。现有的比对软件在处理基因序列时存在延迟高，搜索效率低，性能不够好的问题，不能满足基因序列的处理需求。

发明内容

本发明要解决的技术问题是为了克服现有技术中基因搜索算法的搜索效率低的缺陷，提供一种基因序列的子序列搜索方法、比对方法、系统及设备。

本发明是通过下述技术方案来解决上述技术问题：

本发明提供一种基因序列的子序列搜索方法，所述子序列搜索方法包括：
获取待搜索基因序列的预设数量的测序数据；

依次获取参考基因序列上每一所述测序数据对应的搜索区间的碱基数据，并基于获取结果进行搜索；每一所述测序数据对应的搜索区间基于每一所述测序数据的待搜索碱基序列确定；

重复获取下一次搜索的搜索区间的碱基数据并基于获取结果进行搜索以得到目标子序列；所述下一次搜索的搜索区间根据每一所述测序数据下一个待搜索碱基和上一次搜索的搜索区间确定。

较佳地，所述待搜索基因序列的测序数据的数量基于所述搜索区间的碱

基数据的预取时长和计算时长确定。

较佳地，在依次获取参考基因序列上每一所述测序数据对应的搜索区间的碱基数据的步骤之前，所述子序列搜索方法还包括：

根据每一所述测序数据对应的搜索方向和起始位置，为每一所述测序数据配置多个搜索阶段；

重复获取下一次搜索的搜索区间的碱基数据的步骤包括：

当下一次搜索不存在对应的搜索区间时，以上一次搜索的结束位置为起始位置在所述搜索阶段内重新开始搜索区间的碱基数据的预取。

较佳地，在以上一次搜索的结束位置为起始位置在所述搜索阶段内重新开始搜索区间的碱基数据的预取的步骤之后，所述重复获取下一次搜索的搜索区间的碱基数据的步骤还包括：

当当前搜索阶段的待搜索碱基序列对应的搜索区间的碱基数据均完成预取，在下一搜索阶段内重新开始搜索区间的碱基数据的预取；和/或，

当在当前搜索阶段中当前测序数据的正向待搜索碱基序列对应的搜索区间的碱基数据均完成预取，获取待搜索基因序列的下一个测序数据；和/或，

当所有搜索阶段的搜索区间的碱基数据均完成预取，获取待搜索基因序列的下一个测序数据。

较佳地，获取待搜索基因序列的预设数量的测序数据的步骤包括：

利用缓存池的空队列从IO（输入输出）线程中接收所述测序数据；所述空队列由未存储数据的缓存区排列而成，缓存区存储数据后进入缓存池的满队列；

利用所述满队列将所述测序数据传输到比对线程以进行子序列搜索；所述比对线程完成子序列搜索后释放缓存区存储的数据。

本发明还提供一种基因序列的比对方法，所述比对方法包括：

获取待比对基因序列；

利用如上所述的基因序列的子序列搜索方法得到所述待比对基因序列

的目标子序列；

扩展所述目标子序列以得到所述待比对基因序列的比对结果。

本发明还提供一种基因序列的子序列搜索系统，所述子序列搜索系统包括：

测序数据获取模块，用于获取待搜索基因序列的预设数量的测序数据；

搜索数据获取模块，用于依次获取参考基因序列上每一所述测序数据对应的搜索区间的碱基数据，并基于获取结果进行搜索；每一所述测序数据对应的搜索区间基于每一所述测序数据的待搜索碱基序列确定；

子序列确定模块，用于重复调用所述搜索数据获取模块以获取下一次搜索的搜索区间的碱基数据并基于获取结果进行搜索以得到目标子序列；所述下一次搜索的搜索区间根据每一所述测序数据下一个待搜索碱基和上一次搜索的搜索区间确定。

较佳地，所述待搜索基因序列的测序数据的数量基于所述搜索区间的碱基数据的预取时长和计算时长确定。

较佳地，所述子序列搜索系统还包括：

搜索阶段配置模块，用于根据每一所述测序数据对应的搜索方向和起始位置，为每一所述测序数据配置多个搜索阶段；

所述子序列确定模块具体用于当下一次搜索不存在对应的搜索区间时，调用所述搜索数据获取模块以上一次搜索的结束位置为起始位置在所述搜索阶段内重新开始搜索区间的碱基数据的预取。

较佳地，所述子序列确定模块具体用于当当前搜索阶段的待搜索碱基序列对应的搜索区间的碱基数据均完成预取，调用所述搜索数据获取模块在下一搜索阶段内重新开始搜索区间的碱基数据的预取；和/或，

所述子序列确定模块具体用于当在当前搜索阶段中当前测序数据的正向待搜索碱基序列对应的搜索区间的碱基数据均完成预取，调用所述测序数据获取模块获取待搜索基因序列的下一个测序数据；和/或，

所述子序列确定模块具体用于当所有搜索阶段的搜索区间的碱基数据均完成预取,调用所述测序数据获取模块获取待搜索基因序列的下一个测序数据。

较佳地,所述测序数据获取模块具体用于利用缓存池的空队列从 IO 线程中接收所述测序数据;所述空队列由未存储数据的缓存区排列而成,缓存区存储数据后进入缓存池的满队列;

所述测序数据获取模块具体用于利用所述满队列将所述测序数据传输到比对线程以进行子序列搜索;所述比对线程完成子序列搜索后释放缓存区存储的数据。

本发明还提供一种基因序列的比对系统,所述比对系统包括:

基因序列获取模块,用于获取待比对的基因序列;

目标序列确定模块,用于利用如上所述的基因序列的子序列搜索系统得到所述待比对的基因序列的目标子序列;

目标序列扩展模块,用于扩展所述目标子序列以得到所述待比对基因序列的比对结果。

本发明还提供一种电子设备,包括存储器、处理器及存储在存储器上并可在处理器上运行的计算机程序,所述处理器执行所述计算机程序时实现如上所述的基因序列的子序列搜索方法或如上所述的基因序列的比对方法。

本发明还提供一种计算机可读存储介质,其上存储有计算机程序,所述计算机程序被处理器执行时实现如上所述的基因序列的子序列搜索方法或如上所述的基因序列的比对方法。

本发明的积极进步效果在于:

本发明提供的基因序列的子序列搜索方法,通过批量获取待搜索基因序列的测序数据,并预先提取参考基因序列上每一测序数据对应的搜索区间的碱基数据,通过批量处理测序数据,使得 CPU 在等待预取结果时可以处理其他测序数据的计算,即通过批量处理使得 CPU 的计算时间填补了等待预

取的时间，提高了基因搜索的效率。

附图说明

为了更清楚地说明本说明书实施例的技术方案，下面将对实施例描述中所需要使用的附图作简单的介绍。显而易见地，下面描述中的附图仅仅是本说明书的一些示例或实施例，对于本领域的普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图将本说明书应用于其它类似情景。

图 1 为本发明实施例 1 中的基因序列的子序列搜索方法的流程图。

图 2 为本发明实施例 1 中的 FASTQ 格式的测序数据的示例图。

图 3 为本发明实施例 1 中的确定待搜索基因序列的测序数据的数量的示例图。

图 4 为本发明实施例 1 中的解析测序数据的示例图。

图 5 为本发明实施例 1 中的设置搜索阶段的示意图。

图 6 为本发明实施例 2 中的基因序列的比对方法的流程图。

图 7 为本发明实施例 3 中的基因序列的子序列搜索系统的结构示意图。

图 8 为本发明实施例 4 中的基因序列的比对系统的结构示意图。

图 9 为本发明实施例 5 中的电子设备的结构示意图。

具体实施方式

下面通过实施例的方式进一步说明本发明，但并不因此将本发明限制在所述的实施例范围之中。

在本文提及“实施例”意味着，结合实施例描述的特定特征、结构或特性可以包含在本申请的至少一个实施例中。在文中的各个位置出现该短语并不一定均是指相同的实施例，也不是与其它实施例互斥的独立的或备选的实施例。本领域技术人员显式地和隐式地理解的是，本文所描述的实施例可以与其它实施例相结合。

应当理解，本文使用的“系统”、“装置”、“单元”和/或“模组”是用于区分不同级别的不同组件、元件、部件、部分或装配的一种方法。然而，如果其他词语可实现相同的目的，则可通过其他表达来替换所述词语。

如本文中所示，除非上下文明确提示例外情形，“一”、“一个”、“一种”和/或“该”等词并非特指单数，也可包括复数。一般说来，术语“包括”与“包含”仅提示包括已明确标识的步骤和元素，而这些步骤和元素不构成一个排它性的罗列，方法或者设备也可能包含其它的步骤或元素。

本文中的包括的定义，如这里所使用的术语“具有”、“可以具有”、“包括”或“可以包括”指示本文的相应功能、操作、元件等的存在，并且不限制其它的一个或多个功能、操作、元件等的存在。此外应当理解到，如这里所使用的术语“包括”或“具有”是指示在说明书中所描述的特点、数字、步骤、操作、元件、部件或其组合的存在，而不排除一个或多个其它特点、数字、步骤、操作、元件、部件或其组合的存在或增加。

本文中和/或的定义，如这里所使用的术语“A或B”、“A和/或B的至少之一”或“A和/或B的一个或多个”包括与其一起列举的单词的任意和所有组合。例如，“A或B”、“A和B的至少之一”或“A或B的至少之一”意味着（1）包括至少一个A，（2）包括至少一个B，或（3）包括至少一个A和至少一个B两者。

本文中使用了流程图用来说明根据本文的实施例的系统所执行的操作。应当理解的是，前面或后面操作不一定按照顺序来精确地执行。相反，可以按照倒序或同时处理各个步骤。同时，也可以将其他操作添加到这些过程中，或从这些过程移除某一步或数步操作。

实施例 1

比对软件的核心算法主要分为两个部分：**Seed** 搜索和基于 **Seed** 进行扩展；其中 **Seed** 指的是一条测序碱基序列的子序列。**Seed** 的搜索效率对比对效率影响很大，所以一般的比对软件在 **Seed** 阶段采用后缀数组（SA, suffix

array)、哈希表 (HashTable) 等索引结构提升 Seed 搜索效率。

以基于 SA 的二分搜索算法为例,每次搜索需要访问参考序列的不同位置,内存访问空间局部性差,内存延迟高,且搜索 MMP (Maximal Mappable Prefix, 最大可映射前缀) 结束后需要再次在 SA 数组中确认匹配区间,导致搜索效率低,性能不够好。针对比对软件的核心算法存访问空间局部性差,内存延迟高的缺陷,本实施例提出一种基因序列的子序列搜索方法。

请参考图 1, 其为本实施例中的基因序列的子序列搜索方法的流程图。具体的, 如图 1 所示, 所述子序列搜索方法包括:

S101、获取待搜索基因序列的预设数量的测序数据; 具体地, 如图 2 所示, 以 FASTQ (一种测序数据的文件格式) 为例, 该文件每 4 行为一个记录 (名称、序列、注释、质量值), 该记录称为 reads。

S102、依次获取参考基因序列上每一测序数据对应的搜索区间的碱基数据, 并基于获取结果进行搜索; 每一测序数据对应的搜索区间基于每一测序数据的待搜索碱基序列确定; 具体地, 在一种可选的实施方式中, 可以通过预取技术获取搜索区间的碱基数据; 预取是 CPU (Central Processing Unit, 中央处理器) 使用的一种技术, 在指令或数据实际需要之前, 将从其在较慢内存中提取到较快的本地 Cache (高速缓冲存储器) 中, 从而提高执行性能; 如果没有预取机制, 由于数据之间存在依赖性, 大部分时间都浪费在等待内存的相关数据, 计算单元大部分时间处于空闲状态, 采用预取机制使得计算单元和相关存取单元可以保持高效地运作, 最大化地利用相关的硬件资源; 本实施例中, 通过批量处理测序数据, 使得 CPU 在等待预取结果时可以处理其他 reads 的计算, 即通过批量处理使得 CPU 的计算时间填补了等待预取的时间, 相当于隐藏了 CPU 等待预取的时间, 提高了基因搜索的效率。优选地, 采用软件预取的方式。

在一种可选的实施方式中, 待搜索基因序列的测序数据的数量基于搜索区间的碱基数据的预取时长和计算时长确定。如图 3 所示, 当测序数据 R4

的计算结束时，CPU 正好接收测序数据 R1 的预取结果，可以继续基于测序数据 R1 的预取结果的搜索计算。根据实际测试，当批量读取测序数据的数量为 8~16 时本方法可以获得最佳性能。

S103、重复获取下一次搜索的搜索区间的碱基数据并基于获取结果进行搜索以得到目标子序列；下一次搜索的搜索区间根据每一测序数据下一个待搜索碱基和上一次搜索的搜索区间确定。

以下通过举例对上述方法的步骤进行具体说明。在一种可选的实施方式中，步骤 S101 包括：

S1011、利用缓存池的空队列从 IO 线程中接收测序数据；空队列由未存储数据的缓存区排列而成，缓存区存储数据后进入缓存池的满队列；

S1012、利用满队列将测序数据传输到比对线程以进行子序列搜索；比对线程完成子序列搜索后释放缓存区存储的数据。

在一个方案中可以通过以下步骤对测序数据的解析：

- 1、解压 FASTQ 文件，将 reads 数据写入管道；
- 2、比对阶段是多线程，多个线程竞争管道读取权限，竞争失败的线程需要等待；
- 3、竞争成功的线程从管道读取一定数量的 reads 信息，将其拷贝至线程本地缓存，然后释放读取权限，唤醒本轮竞争失败的线程；
- 4、本轮竞争失败的线程将再次竞争管道，竞争失败的线程将继续等待；
- 5、若比对线程完成比对任务，则将加入下一轮竞争管道读取权限的线程列表；
- 6、当 FASTQ 文件处理完毕，比对线程结束。

在这个对测序数据的解析方法中，Reads 数据存在多次拷贝，从解压缓存拷贝至管道，从管道拷贝至线程本地缓存池；每次只有一个线程竞争成功，其余线程全部等待，线程执行效率低。

如图 4 所示，本实施方式通过以下步骤对测序数据的解析：

1、IO 线程负责解析文件，从缓存池的空队列中获取一个缓存区，将一定数量的 reads 记录写入该缓存区，然后加入满队列；

2、比对线程从满队列中获取缓存区，完成 MMP 搜索、MMP 扩展后将缓存区加入空队列；

可见，如果比对线程消耗缓存区的速度比 IO 线程产生缓存区的速度快，会导致比对线程等待“满”缓存区，性能瓶颈由 IO 解析线程决定；如果比对线程消耗缓存区的速度低于 IO 线程产生缓存区的速度，则会导致 IO 线程等待空缓存区，这时，性能瓶颈是由对齐算法的效率决定的。在本实施方式中只需要一次 reads 信息拷贝，减少了数据拷贝次数，多个线程并行执行，提高了线程执行效率。缓冲区的大小可以调整，实际测试表明，一个缓冲区可以容纳 120k（千）条 reads。

在一种可选的实施方式中，在步骤 S102 之前，子序列搜索方法还包括：

根据每一测序数据对应的搜索方向和起始位置，为每一测序数据配置多个搜索阶段；

具体地，从满队列中获取一个缓存区，根据 reads 格式解析，记录缓存区中每次 reads 的各参数的起始位置。对缓冲区中的所有 reads 进行预处理，例如将序列中的碱基转换为 0、1、2、3，并初始化每个 reads 的参数。每次处理固定数量的 reads，每次 reads 只搜索一个碱基，并更新对应的搜索任务。

每条 reads 有多个 MMP 搜索子过程，每次搜索的起始位置不同，并且还涉及复杂的逻辑，例如两个方向的搜索和起始位置的选择。为了简化处理流程，本实施例设计了一个 stage(搜索阶段)结构：一个 stage 代表一个 reads 的搜索事务。根据搜索方向和起始位置，一条 reads 可以配置多个阶段。每个 stage-(off, dir)标识 MMP 搜索的起始位置和方向。

对于每个阶段，都会有多个 MMP 搜索。为了识别每个 MMP 搜索状态，本实施方式中设计了一个 task（搜索任务）结构；task-(sa1,sa2,off,len,dir,flag)标识当前 MMP 搜索的起始位置（off）、搜索方向（dir）、匹配长度（len）、

MMP 的范围在搜索区间 SA [sa1,sa2]上; task 使用 FMIndex(一种索引结构,应用于基于序列搜索,特点是每次反向搜索一个字符)一次搜索一个碱基。如果搜索成功,则更新(sa1,sa2,len),如果搜索失败,则任务结束并记录一个 MMP。

在本实施方式中,步骤 S103 可以包括:

S1031、当下一次搜索不存在对应的搜索区间时,以上一次搜索的结束位置为起始位置在搜索阶段内重新开始搜索区间的碱基数据的预取。具体地,如果一个 task 结束,剩下的匹配碱基足够了,在不匹配的位置开始一个新的 task; 每一个新创建的 task 会被 pre-index 结构进行初始化,目的在于缩减搜索区间提升 FMIndex 的搜索效率。

此外,在步骤 S1031 之后,步骤 S103 还包括:

S1032、当当前搜索阶段的待搜索碱基序列对应的搜索区间的碱基数据均完成预取,在下一搜索阶段内重新开始搜索区间的碱基数据的预取;具体地,当一个 task 结构结束时,需要在不匹配的位置开始一个新的 task,或者在 stage 转移后开始一个新的 task; 如果一个 task 结束后剩余的匹配碱基太少,直接切换到下一个 stage。

S1033、当所有搜索阶段的搜索区间的碱基数据均完成预取,获取待搜索基因序列的下一个测序数据。具体地,当一个 reads 的所有 stage 都执行完毕后,从缓冲区队列中重新加载一条 reads,直到缓冲区中的所有 reads 处理完;如图 5 所示,一条 reads 的 MMP 搜索涉及到多个 stage 之间的转移,如 stage-0 完成后,切换到 stage-1,以此类推,直到该 reads 的所有 stage 完成,当前 reads 的 MMP 搜索完成,即获取待搜索基因序列的下一个测序数据。

S1034、当在当前搜索阶段中当前测序数据的正向待搜索碱基序列对应的搜索区间的碱基数据均完成预取,获取待搜索基因序列的下一个测序数据;具体地,并非所有阶段都会执行,可能会跳过一个阶段。例如,如果正向待搜索碱基序列对应的搜索阶段(如 stage-0)中的 MMP 匹配长度等于 reads

的长度，则可以跳过反向待搜索碱基序列对应的搜索阶段（如 stage-4），即获取待搜索基因序列的下一个测序数据。

本实施例提供的基因序列的子序列搜索方法，通过批量获取待搜索基因序列的测序数据，并采用预取机制获取参考基因序列上每一测序数据对应的搜索区间的碱基数据，通过批量处理测序数据，使得 CPU 在等待预取结果时可以处理其他测序数据的计算，即通过批量处理使得 CPU 的计算时间填补了等待预取的时间，提高了基因搜索的效率。

实施例 2

请参考图 6，其为本实施例中的基因序列的比对方法的流程图。具体的，如图 6 所示，所述比对方法包括：

S201、获取待比对基因序列；

S202、利用实施例 1 的基因序列的子序列搜索方法得到待比对基因序列的目标子序列；

S203、扩展目标子序列以得到待比对基因序列的比对结果。

本实施例提供的基因序列的比对方法，通过利用上述子序列搜索方法，使得 CPU 在等待预取结果时可以处理其他测序数据的计算，即通过批量处理使得 CPU 的计算时间填补了等待预取的时间，提高了基因搜索的效率。

实施例 3

请参考图 7，其为本实施例中的基因序列的子序列搜索系统的结构示意图。具体的，如图 7 所示，所述子序列搜索系统包括：

测序数据获取模块 1，用于获取待搜索基因序列的预设数量的测序数据；具体地，如图 2 所示，以 FASTQ 为例，该文件每 4 行为一个记录（名称、序列、注释、质量值），该记录称为 reads。

搜索数据获取模块 2，用于依次获取参考基因序列上每一测序数据对应的搜索区间的碱基数据，并基于获取结果进行搜索；每一测序数据对应的搜索区间基于每一测序数据的待搜索碱基序列确定；具体地，在一种可选的实

施方式中,搜索数据获取模块2可以通过预取技术获取搜索区间的碱基数据;预取是CPU使用的一种技术,在指令或数据实际需要之前,将从其在较慢内存中提取到较快的本地Cache(高速缓冲存储器)中,从而提高执行性能;如果没有软件预取机制,由于数据之间存在依赖性,大部分时间都浪费在等待内存的相关数据,计算单元大部分时间处于空闲状态,采用预取机制使得计算单元和相关存取单元可以保持高效地运作,最大化地利用相关的硬件资源;本实施例中,通过批量处理测序数据,使得CPU在等待预取结果时可以处理其他reads的计算,即通过批量处理使得CPU的计算时间填补了等待预取的时间,相当于隐藏了CPU等待预取的时间,提高了基因搜索的效率。优选地,搜索数据获取模块2可以采用软件预取的方式。

在一种可选的实施方式中,待搜索基因序列的测序数据的数量基于搜索区间的碱基数据的预取时长和计算时长确定。如图3所示,当测序数据R4的计算结束时,CPU正好接收测序数据R1的预取结果,可以继续基于测序数据R1的预取结果的搜索计算。根据实际测试,当批量读取测序数据的数量为8~16时本方法可以获得最佳性能。

子序列确定模块3,用于重复调用搜索数据获取模块以获取下一次搜索的搜索区间的碱基数据并基于获取结果进行搜索以得到目标子序列;下一次搜索的搜索区间根据每一测序数据下一个待搜索碱基和上一次搜索的搜索区间确定。

以下通过举例对上述方法的步骤进行具体说明。在一种可选的实施方式中,测序数据获取模块1具体用于利用缓存池的空队列从IO线程中接收测序数据;空队列由未存储数据的缓存区排列而成,缓存区存储数据后进入缓存池的满队列;

测序数据获取模块1具体用于利用满队列将测序数据传输到比对线程以进行子序列搜索;比对线程完成子序列搜索后释放缓存区存储的数据。

在一个方案中可以通过以下方法对测序数据的解析:

- 1、解压 FASTQ 文件，将 reads 数据写入管道；
- 2、比对阶段是多线程，多个线程竞争管道读取权限，竞争失败的线程需要等待；
- 3、竞争成功的线程从管道读取一定数量的 reads 信息，将其拷贝至线程本地缓存，然后释放读取权限，唤醒本轮竞争失败的线程；
- 4、本轮竞争失败的线程将再次竞争管道，竞争失败的线程将继续等待；
- 5、若比对线程完成比对任务，则将加入下一轮竞争管道读取权限的线程列表；
- 6、当 FASTQ 文件处理完毕，比对线程结束。

在这个对测序数据的解析方法中，Reads 数据存在多次拷贝，从解压缓存拷贝至管道，从管道拷贝至线程本地缓存池；每次只有一个线程竞争成功，其余线程全部等待，线程执行效率低。

如图 4 所示，本实施方式的测序数据获取模块 1 具体通过以下步骤对测序数据的解析：

- 1、IO 线程负责解析文件，从缓存池的空队列中获取一个缓存区，将一定数量的 reads 记录写入该缓存区，然后加入满队列；
- 2、比对线程从满队列中获取缓存区，完成 MMP 搜索、MMP 扩展后将缓存区加入空队列；

可见，如果比对线程消耗缓存区的速度比 IO 线程产生缓存区的速度快，会导致比对线程等待“满”缓存区，性能瓶颈由 IO 解析线程决定；如果比对线程消耗缓存区的速度低于 IO 线程产生缓存区的速度，则会导致 IO 线程等待空缓存区，这时，性能瓶颈是由对齐算法的效率决定的。在本实施方式中只需要一次 reads 信息拷贝，减少了数据拷贝次数，多个线程并行执行，提高了线程执行效率。缓冲区的大小可以调整，实际测试表明，一个缓冲区可以容纳 120k 条 reads。

在一种可选的实施方式中，子序列搜索系统还包括：

搜索阶段配置模块 4, 用于根据每一测序数据对应的搜索方向和起始位置, 为每一测序数据配置多个搜索阶段;

具体地, 从满队列中获取一个缓存区, 根据 reads 格式解析, 记录缓存区中每次 reads 的各参数的起始位置。对缓冲区中的所有 reads 进行预处理, 例如将序列中的碱基转换为 0、1、2、3, 并初始化每个 reads 的参数。每次处理固定数量的 reads, 每次 reads 只搜索一个碱基, 并更新对应的搜索任务。

每条 reads 有多个 MMP 搜索子过程, 每次搜索的起始位置不同, 并且还涉及复杂的逻辑, 例如两个方向的搜索和起始位置的选择。为了简化处理流程, 本实施例设计了一个 stage(搜索阶段)结构: 一个 stage 代表一个 reads 的搜索事务。根据搜索方向和起始位置, 一条 reads 可以配置多个阶段。每个 stage-(off, dir)标识 MMP 搜索的起始位置和方向。

对于每个阶段, 都会有多个 MMP 搜索。为了识别每个 MMP 搜索状态, 本实施方式中设计了一个 task(搜索任务)结构: task-(sa1, sa2, off, len, dir, flg)标识当前 MMP 搜索的起始位置(off)、搜索方向(dir)、匹配长度(len)、MMP 的范围在搜索区间 SA[sa1, sa2]上; task 使用 FMIndex(一种索引结构, 应用于基于序列搜索, 特点是每次反向搜索一个字符)一次搜索一个碱基。如果搜索成功, 则更新(sa1, sa2, len), 如果搜索失败, 则任务结束并记录一个 MMP。

在本实施方式中, 子序列确定模块 3 具体用于当下一次搜索不存在对应的搜索区间时, 调用搜索数据获取模块 2 以上一次搜索的结束位置为起始位置在搜索阶段内重新开始搜索区间的碱基数据的预取。具体地, 如果一个 task 结束, 剩下的匹配碱基足够了, 在不匹配的位置开始一个新的 task; 每一个新创建的 task 会被 pre-index 结构进行初始化, 目的在于缩减搜索区间提升 FMIndex 的搜索效率。

此外, 子序列确定模块 3 具体用于当当前搜索阶段的待搜索碱基序列对应的搜索区间的碱基数据均完成预取, 调用搜索数据获取模块 2 在下一搜索

阶段内重新开始搜索区间的碱基数据的预取；具体地，当一个 task 结构结束时，需要在不匹配的位置开始一个新的 task，或者在 stage 转移后开始一个新的 task；如果一个 task 结束后剩余的匹配碱基太少，直接切换到下一个 stage。

子序列确定模块 3 具体用于当在当前搜索阶段中当前测序数据的正向待搜索碱基序列对应的搜索区间的碱基数据均完成预取，调用测序数据获取模块 1 获取待搜索基因序列的下一个测序数据；具体地，并非所有阶段都会执行，可能会跳过一个阶段。例如，如果正向待搜索碱基序列对应的搜索阶段（如 stage-0）中的 MMP 匹配长度等于 reads 的长度，则可以跳过反向待搜索碱基序列对应的搜索阶段（如 stage-4），即获取待搜索基因序列的下一个测序数据。

子序列确定模块 3 具体用于当所有搜索阶段的搜索区间的碱基数据均完成预取，调用测序数据获取模块 1 获取待搜索基因序列的下一个测序数据。具体地，当一个 reads 的所有 stage 都执行完毕后，从缓冲区队列中重新加载一条 reads，直到缓冲区中的所有 reads 处理完；如图 5 所示，一条 reads 的 MMP 搜索涉及到多个 stage 之间的转移，如 stage-0 完成后，切换到 stage-1，以此类推，直到该 reads 的所有 stage 完成，当前 reads 的 MMP 搜索完成，即获取待搜索基因序列的下一个测序数据。

本实施例提供的基因序列的子序列搜索系统，通过批量获取待搜索基因序列的测序数据，并采用预取机制获取参考基因序列上每一测序数据对应的搜索区间的碱基数据，通过批量处理测序数据，使得 CPU 在等待预取结果时可以处理其他测序数据的计算，即通过批量处理使得 CPU 的计算时间填补了等待预取的时间，提高了基因搜索的效率。

实施例 4

请参考图 8，其为本实施例中的基因序列的比对系统的结构示意图。具体的，如图 8 所示，所述比对系统包括：

基因序列获取模块 5, 用于获取待比对的基因序列;

目标序列确定模块 6, 用于利用实施例 3 的基因序列的子序列搜索系统得到待比对的基因序列的目标子序列;

目标序列扩展模块 7, 用于扩展目标子序列以得到待比对基因序列的比对结果。

本实施例提供的基因序列的比对系统, 通过利用上述子序列搜索系统, 使得 CPU 在等待预取结果时可以处理其他测序数据的计算, 即通过批量处理使得 CPU 的计算时间填补了等待预取的时间, 提高了基因搜索的效率。

实施例 5

图 9 为本发明实施例 5 提供的一种电子设备的结构示意图。所述电子设备包括存储器、处理器及存储在存储器上并可在处理器上运行的计算机程序, 所述处理器执行所述程序时实现实施例 1 的基因序列的子序列搜索方法或实施例 2 的基因序列的比对方法。图 9 显示的电子设备 30 仅仅是一个示例, 不应对本发明实施例的功能和使用范围带来任何限制。

如图 9 所示, 电子设备 30 可以以通用计算设备的形式表现, 例如其可以为服务器设备。电子设备 30 的组件可以包括但不限于: 上述至少一个处理器 31、上述至少一个存储器 32、连接不同系统组件(包括存储器 32 和处理器 31)的总线 33。

总线 33 包括数据总线、地址总线和控制总线。

存储器 32 可以包括易失性存储器, 例如随机存取存储器(RAM)321 和/或高速缓存存储器 322, 还可以进一步包括只读存储器(ROM)323。

存储器 32 还可以包括具有一组(至少一个)程序模块 324 的程序/实用工具 325, 这样的程序模块 324 包括但不限于: 操作系统、一个或者多个应用程序、其它程序模块以及程序数据, 这些示例中的每一个或某种组合中可能包括网络环境的实现。

处理器 31 通过运行存储在存储器 32 中的计算机程序, 从而执行各种功

能应用以及数据处理,例如本发明实施例1的基因序列的子序列搜索方法或实施例2的基因序列的比对方法。

电子设备30也可以与一个或多个外部设备34(例如键盘、指向设备等)通信。这种通信可以通过输入/输出(I/O)接口35进行。并且,模型生成的设备30还可以通过网络适配器36与一个或者多个网络(例如局域网(LAN),广域网(WAN)和/或公共网络,例如因特网)通信。如图所示,网络适配器36通过总线33与模型生成的设备30的其它模块通信。应当明白,尽管图中未示出,可以结合模型生成的设备30使用其它硬件和/或软件模块,包括但不限于:微代码、设备驱动器、冗余处理器、外部磁盘驱动阵列、RAID(磁盘阵列)系统、磁带驱动器以及数据备份存储系统等。

应当注意,尽管在上文详细描述中提及了电子设备的若干单元/模块或子单元/模块,但是这种划分仅仅是示例性的并非强制性的。实际上,根据本发明的实施方式,上文描述的两个或更多单元/模块的特征和功能可以在一个单元/模块中具体化。反之,上文描述的一个单元/模块的特征和功能可以进一步划分为由多个单元/模块来具体化。

实施例4

本实施例提供了一种计算机可读存储介质,其上存储有计算机程序,所述程序被处理器执行时实现实施例1的基因序列的子序列搜索方法或实施例2的基因序列的比对方法。

其中,可读存储介质可以采用的更具体可以包括但不限于:便携式盘、硬盘、随机存取存储器、只读存储器、可擦拭可编程只读存储器、光存储器件、磁存储器件或上述的任意合适的组合。

在可能的实施方式中,本发明还可以实现为一种程序产品的形式,其包括程序代码,当所述程序产品在终端设备上运行时,所述程序代码用于使所述终端设备执行实现实施例1的基因序列的子序列搜索方法或实施例2的基因序列的比对方法。

其中,可以以一种或多种程序设计语言的任意组合来编写用于执行本发明的程序代码,所述程序代码可以完全地在用户设备上执行、部分地在用户设备上执行、作为一个独立的软件包执行、部分在用户设备上部分在远程设备上执行或完全在远程设备上执行。

虽然以上描述了本发明的具体实施方式,但是本领域的技术人员应当理解,这仅是举例说明,本发明的保护范围是由所附权利要求书限定的。本领域的技术人员在不背离本发明的原理和实质的前提下,可以对这些实施方式做出多种变更或修改,但这些变更和修改均落入本发明的保护范围。

权利要求

1.一种基因序列的子序列搜索方法，其特征在于，所述子序列搜索方法包括：

获取待搜索基因序列的预设数量的测序数据；

依次获取参考基因序列上每一所述测序数据对应的搜索区间的碱基数据，并基于获取结果进行搜索；每一所述测序数据对应的搜索区间基于每一所述测序数据的待搜索碱基序列确定；

重复获取下一次搜索的搜索区间的碱基数据并基于获取结果进行搜索以得到目标子序列；所述下一次搜索的搜索区间根据每一所述测序数据下一个待搜索碱基和上一次搜索的搜索区间确定。

2.如权利要求1所述的子序列搜索方法，其特征在于，所述待搜索基因序列的测序数据的数量基于所述搜索区间的碱基数据的预取时长和计算时长确定。

3.如权利要求1所述的子序列搜索方法，其特征在于，在依次获取参考基因序列上每一所述测序数据对应的搜索区间的碱基数据的步骤之前，所述子序列搜索方法还包括：

根据每一所述测序数据对应的搜索方向和起始位置，为每一所述测序数据配置多个搜索阶段；

重复获取下一次搜索的搜索区间的碱基数据的步骤包括：

当下一次搜索不存在对应的搜索区间时，以上一次搜索的结束位置为起始位置在所述搜索阶段内重新开始搜索区间的碱基数据的预取。

4.如权利要求3所述的子序列搜索方法，其特征在于，在以上一次搜索的结束位置为起始位置在所述搜索阶段内重新开始搜索区间的碱基数据的预取的步骤之后，所述重复获取下一次搜索的搜索区间的碱基数据的步骤还包括：

当当前搜索阶段的待搜索碱基序列对应的搜索区间的碱基数据均完成预取，在下一搜索阶段内重新开始搜索区间的碱基数据的预取；和/或，

当在当前搜索阶段中当前测序数据的正向待搜索碱基序列对应的搜索区间的碱基数据均完成预取，获取待搜索基因序列的下一个测序数据；和/或，

当所有搜索阶段的搜索区间的碱基数据均完成预取，获取待搜索基因序列的下一个测序数据。

5.如权利要求1所述的子序列搜索方法，其特征在于，获取待搜索基因序列的预设数量的测序数据的步骤包括：

利用缓存池的空队列从IO线程中接收所述测序数据；所述空队列由未存储数据的缓存区排列而成，缓存区存储数据后进入缓存池的满队列；

利用所述满队列将所述测序数据传输到比对线程以进行子序列搜索；所述比对线程完成子序列搜索后释放缓存区存储的数据。

6.一种基因序列的比对方法，其特征在于，所述比对方法包括：

获取待比对基因序列；

利用如权利要求1-5任一项所述的基因序列的子序列搜索方法得到所述待比对基因序列的目标子序列；

扩展所述目标子序列以得到所述待比对基因序列的比对结果。

7.一种基因序列的子序列搜索系统，其特征在于，所述子序列搜索系统包括：

测序数据获取模块，用于获取待搜索基因序列的预设数量的测序数据；

搜索数据获取模块，用于依次获取参考基因序列上每一所述测序数据对应的搜索区间的碱基数据，并基于获取结果进行搜索；每一所述测序数据对应的搜索区间基于每一所述测序数据的待搜索碱基序列确定；

子序列确定模块，用于重复调用所述搜索数据预取模块以获取下一次搜索的搜索区间的碱基数据并基于获取结果进行搜索以得到目标子序列；所述下一次搜索的搜索区间根据每一所述测序数据下一个待搜索碱基和上一次

搜索的搜索区间确定。

8.一种基因序列的比对系统，其特征在于，所述比对系统包括：

基因序列获取模块，用于获取待比对的基因序列；

目标序列确定模块，用于利用如权利要求7所述的基因序列的子序列搜索系统得到所述待比对的基因序列的目标子序列；

目标序列扩展模块，用于扩展所述目标子序列以得到所述待比对基因序列的比对结果。

9.一种电子设备，包括存储器、处理器及存储在存储器上并可在处理器上运行的计算机程序，其特征在于，所述处理器执行所述计算机程序时实现如权利要求1至5中任一项所述的基因序列的子序列搜索方法或权利要求6所述的基因序列的比对方法。

10.一种计算机可读存储介质，其上存储有计算机程序，其特征在于，所述计算机程序被处理器执行时实现如权利要求1至5中任一项所述的基因序列的子序列搜索方法或权利要求6所述的基因序列的比对方法。

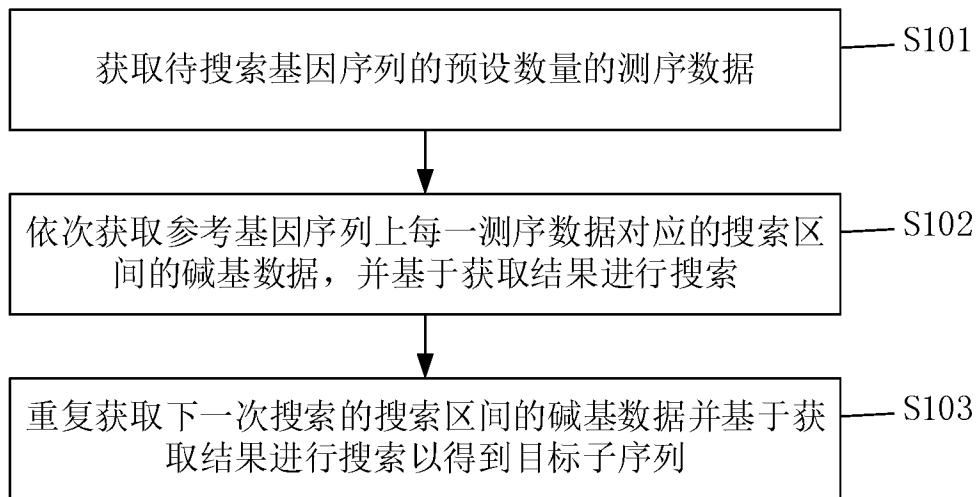


图 1



图 2

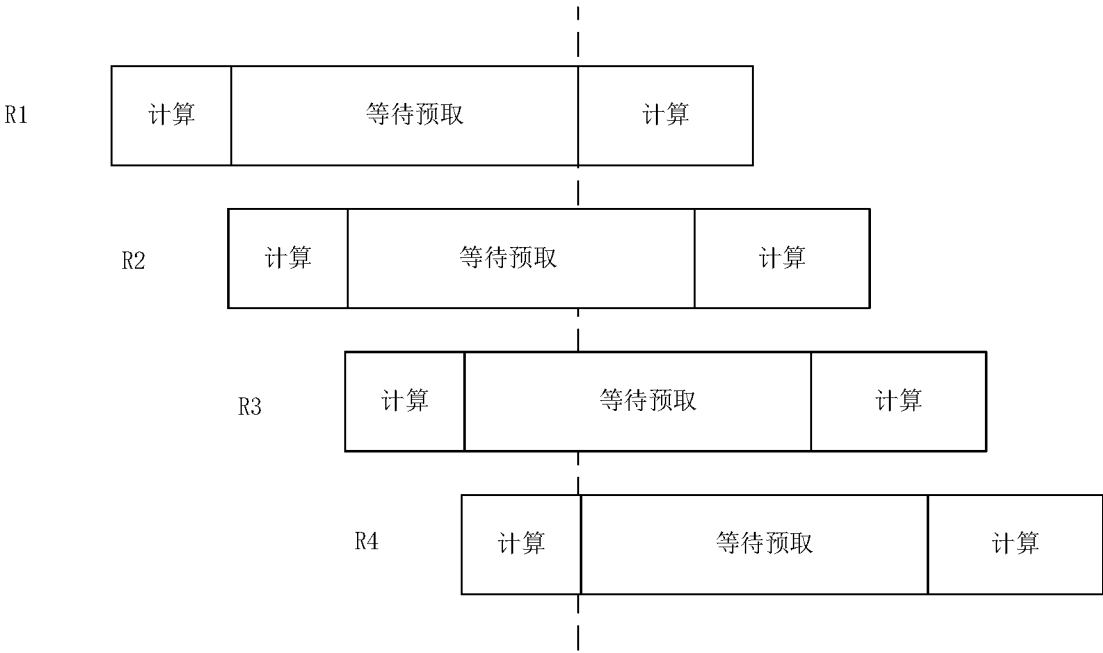


图 3

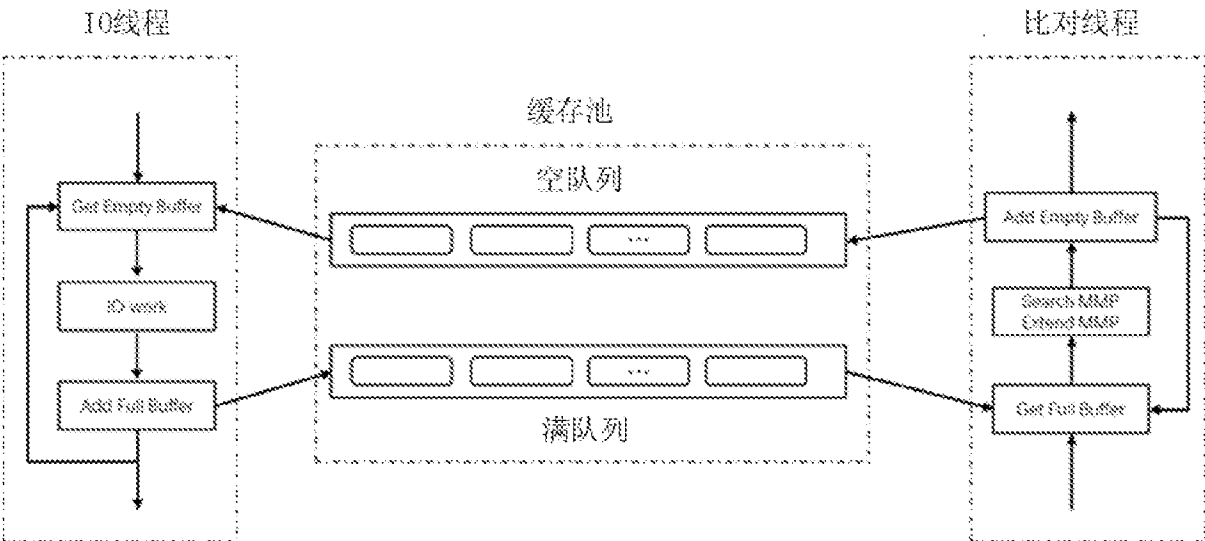


图 4

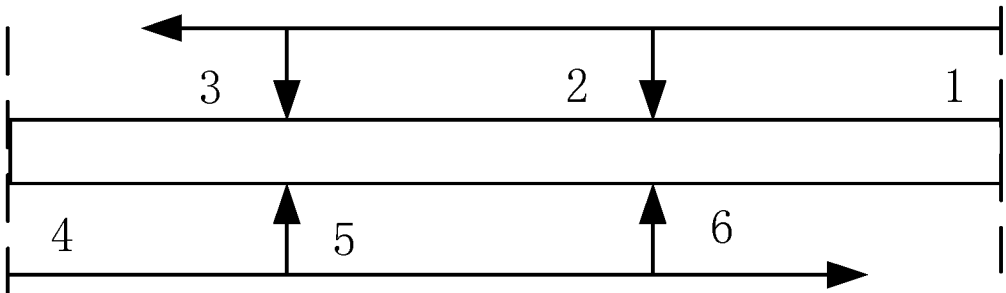


图 5

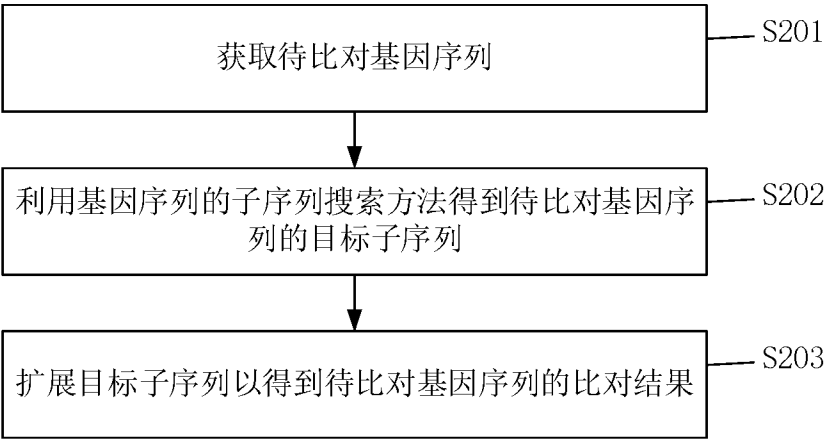


图 6

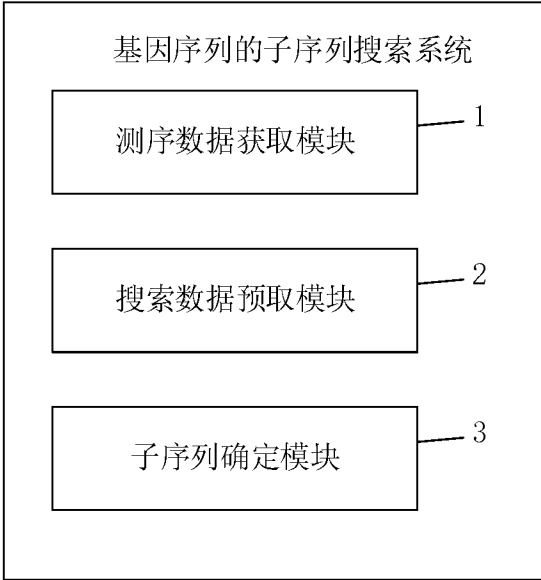


图 7

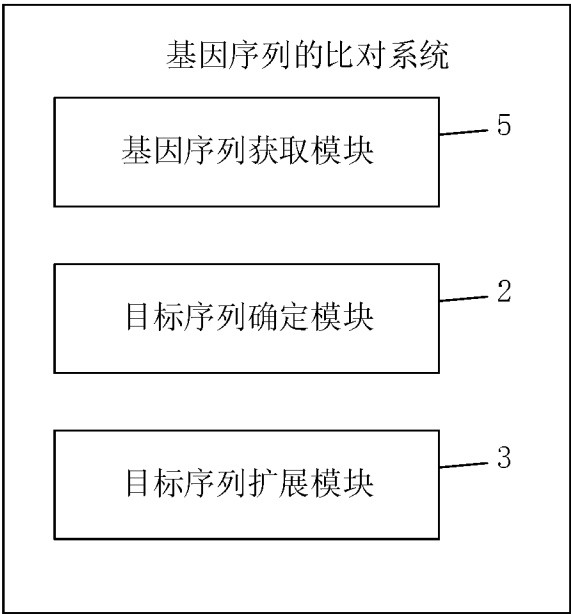


图 8

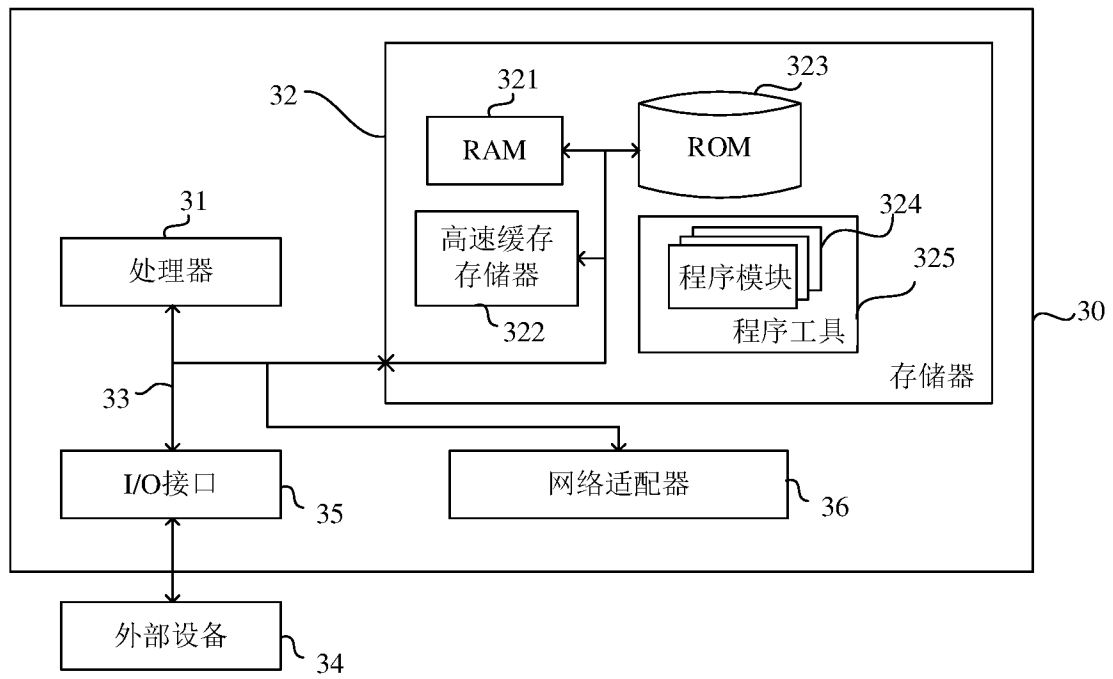


图 9

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2023/101916

A. CLASSIFICATION OF SUBJECT MATTER

G16B30/10(2019.01)i; G06F16/10(2019.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

IPC: G16B, G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

VEN, CNABS, CNTXT, WOTXT, EPTXT, USTXT, JPTXT, CNKI, IEEE: 基因, 碱基, 序列, 片段, 匹配, 对比, 检索, 查询, 搜索, 区间, 预取, 多线程, gene, base, sequence, fragment, match, compare, retrieve, query, search, interval, prefetch, multi, thread.

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 103793626 A (SAMSUNG SDS CO., LTD. et al.) 14 May 2014 (2014-05-14) description, paragraphs 28-123	1, 3, 6-10
Y	CN 103793626 A (SAMSUNG SDS CO., LTD. et al.) 14 May 2014 (2014-05-14) description, paragraphs 28-123	2, 4-5
Y	US 2017237445 A1 (ILLUMINA CAMBRIDGE LTD.) 17 August 2017 (2017-08-17) description, paragraphs 232-242 and 524-527	2, 4-5
A	WO 2022041881 A1 (HUAWEI TECHNOLOGIES CO., LTD.) 03 March 2022 (2022-03-03) entire document	1-10
A	CN 107844684 A (SHENZHEN BGI RESEARCH INSTITUTE) 27 March 2018 (2018-03-27) entire document	1-10
A	CN 114121153 A (GUANGZHOU KINGMED CENTER FOR CLINICAL LABORATORY CO., LTD.) 01 March 2022 (2022-03-01) entire document	1-10



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:

“A” document defining the general state of the art which is not considered to be of particular relevance

“D” document cited by the applicant in the international application

“E” earlier application or patent but published on or after the international filing date

“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

“O” document referring to an oral disclosure, use, exhibition or other means

“P” document published prior to the international filing date but later than the priority date claimed

“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

“&” document member of the same patent family

Date of the actual completion of the international search

10 August 2023

Date of mailing of the international search report

07 February 2024

Name and mailing address of the ISA/CN

China National Intellectual Property Administration (ISA/
CN)
China No. 6, Xitucheng Road, Jimenqiao, Haidian District,
Beijing 100088

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2023/101916

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 115527612 A (SICHUAN INNOVATION RESEARCH INSTITUTE OF TIANJIN UNIVERSITY) 27 December 2022 (2022-12-27) entire document	1-10

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2023/101916

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	103793626	A	14 May 2014	US	2014121986	A1	01 May 2014
				WO	2014069767	A1	08 May 2014
				KR	20140056560	A	12 May 2014
US	2017237445	A1	17 August 2017	WO	2016020682	A1	11 February 2016
				AU	2015298543	A1	16 March 2017
				US	9929746	B2	27 March 2018
				GB	201413877	D0	17 September 2014
				HK	1222744	A1	07 July 2017
				EP	3195480	A1	26 July 2017
				CN	106687966	A	17 May 2017
WO	2022041881	A1	03 March 2022	CN	114090840	A	25 February 2022
CN	107844684	A	27 March 2018	HK	1247368	A1	22 April 2022
CN	114121153	A	01 March 2022	None			
CN	115527612	A	27 December 2022	None			

A. 主题的分类

G16B30/10(2019.01)i; G06F16/10(2019.01)i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

IPC: G16B, G06F

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称, 和使用的检索词(如使用))

VEN, CNABS, CNTXT, WOTXT, EPTXT, USTXT, JPTXT, CNKI, IEEE: 基因, 碱基, 序列, 片段, 匹配, 对比, 检索, 查询, 搜索, 区间, 预取, 多线程, gene, base, sequence, fragment, match, compare, retrieve, query, search, interval, prefetch, multi, thread.

C. 相关文件

类型*	引用文件, 必要时, 指明相关段落	相关的权利要求
X	CN 103793626 A (三星SDS株式会社 等) 2014年5月14日 (2014 - 05 - 14) 说明书第28-123段	1、3、6-10
Y	CN 103793626 A (三星SDS株式会社 等) 2014年5月14日 (2014 - 05 - 14) 说明书第28-123段	2、4-5
Y	US 2017237445 A1 (ILLUMINA CAMBRIDGE LTD.) 2017年8月17日 (2017 - 08 - 17) 说明书第232-242、524-527段	2、4-5
A	WO 2022041881 A1 (HUAWEI TECHNOLOGIES CO., LTD.) 2022年3月3日 (2022 - 03 - 03) 全文	1-10
A	CN 107844684 A (深圳华大基因研究院) 2018年3月27日 (2018 - 03 - 27) 全文	1-10
A	CN 114121153 A (广州金域医学检验中心有限公司) 2022年3月1日 (2022 - 03 - 01) 全文	1-10

☒ 其余文件在C栏的续页中列出。

☒ 见同族专利附件。

* 引用文件的具体类型:

“A” 认为不特别相关的表示了现有技术一般状态的文件

“D” 申请人在国际申请中引证的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件(如具体说明的)

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性

“&” 同族专利的文件

国际检索实际完成的日期

2023年8月10日

国际检索报告邮寄日期

2024年2月7日

ISA/CN的名称和邮寄地址

中国国家知识产权局
中国北京市海淀区蓟门桥西土城路6号 100088

授权官员

徐卫锋

电话号码 (+86) 010-53961365

PCT/ISA/210 表(第2页) (2022年7月)

C. 相关文件		
类 型*	引用文件，必要时，指明相关段落	相关的权利要求
A	CN 115527612 A (天津大学四川创新研究院) 2022年12月27日 (2022 - 12 - 27) 全文	1-10

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2023/101916

检索报告引用的专利文件			公布日 (年/月/日)	同族专利			公布日 (年/月/日)
CN	103793626	A	2014年5月14日	US	2014121986	A1	2014年5月1日
				WO	2014069767	A1	2014年5月8日
				KR	20140056560	A	2014年5月12日
US	2017237445	A1	2017年8月17日	WO	2016020682	A1	2016年2月11日
				AU	2015298543	A1	2017年3月16日
				US	9929746	B2	2018年3月27日
				GB	201413877	D0	2014年9月17日
				HK	1222744	A1	2017年7月7日
				EP	3195480	A1	2017年7月26日
				CN	106687966	A	2017年5月17日
WO	2022041881	A1	2022年3月3日	CN	114090840	A	2022年2月25日
CN	107844684	A	2018年3月27日	HK	1247368	A1	2022年4月22日
CN	114121153	A	2022年3月1日	无			
CN	115527612	A	2022年12月27日	无			