(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 115497567 A
(43)申请公布日 2022.12.20

(21)申请号 202210359541.8

(22)申请日 2022.04.07

(71)申请人 天津大学
地址 300072 天津市南开区卫津路92号

(72)发明人 曲冠锦 吴华明

(74)专利代理机构 天津市三利专利商标代理有限公司 12107
专利代理师 张义

(51)Int.Cl.
G16B 30/10(2019.01)
G16B 40/00(2019.01)
G06K 9/62(2022.01)

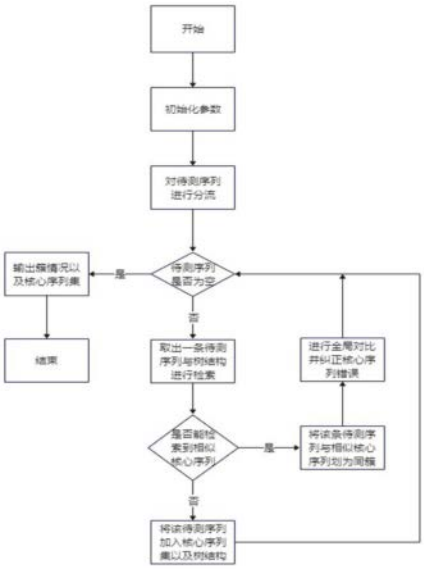
权利要求书2页 说明书8页 附图2页

(54)发明名称

核酸序列聚类方法、装置、计算机可读存储介质、终端

(57)摘要

本发明公开了一种核酸序列聚类方法、装置、计算机可读存储介质、终端通过构建了多个分支的树结构来对核酸序列的指定区间进行检索，进而避免了传统计算编辑距离所消耗的大量时间。此外，本申请采用节点漂移算法以对抗核酸序列发生错误所带来的干扰。与目前已有的核酸聚类算法相比，本申请提供的方法可以对未识别的大量核酸序列进行聚类的同时，还具备对聚类后的核酸序列进行自动纠错与比对的功能，可以直接输出纠错后的核酸原始序列，从而大大减少测序读取后的处理时间。



CN 115497567 A

1.一种核酸序列聚类方法,其特征在于,包括如下步骤:

步骤a:初始化参数;

步骤c:判断待测序列是否为空,若是跳到步骤d;若否跳到步骤f;

步骤d:输出簇情况以及核心序列集;

步骤e:结束;

步骤f:取出一条待测序列与树结构进行检索;

步骤g:判断是否能检索到相似核心序列,若是跳到步骤h;若否跳到步骤j;

步骤h:将该条待测序列与相似核心序列划为同簇;

步骤i:可选地进行全局比对并纠正核心序列错误;

步骤j:将该待测序列加入核心序列集以及树结构,跳到步骤c。

2.根据权利要求1所述的一种核酸序列聚类方法,其特征在于,步骤a中,所述初始化参数包括初始化树结构的数量、树结构的长度、树结构选取的区间位置、纵向漂移值、横向漂移值阈值、进程数、输入文件格式、输出文件格式、核心序列集和树结构。

3.根据权利要求1所述的一种核酸序列聚类方法,其特征在于,若为多进程模式,在步骤a之后,步骤c之前,还包括步骤b:对待测序列进行分流。

4.根据权利要求1所述的一种核酸序列聚类方法,其特征在于,步骤f-j具体包括:

取出一条待测序列与树结构进行检索,判断是否能检索到相似核心序列;首先,比较首段区间与首段树结构是否能够匹配,若能匹配到则停止后续树结构的检索,将该条待测序列与相似核心序列划为同簇,若开启了全局比对功能,则此时将该序列与所匹配到的序列进行全局比对,全局比对后将会标记序列中不同的碱基位,若核心序列集中某条序列的某个碱基位被频繁标记则将会被视为错误碱基位进而进行纠正;若首段区间无法匹配到,则再进行末端区间与末端树结构的匹配,若末端区间可以成功匹配,则停止后续树结构的检索,将该条待测序列与相似核心序列划为同簇,若开启了全局比对功能,则此时将该序列与所匹配到的序列进行全局比对,全局比对后将会标记序列中不同的碱基位,若核心序列集中某条序列的某个碱基位被频繁标记则将会被视为错误碱基位进而进行纠正;若末端也无法成功匹配,则进行中间区间的匹配在选取中间区间时,将会在原指定区间基础上,允许横向漂移纵向漂移数个碱基位置,进而允许选择多个区间,碱基区间检索树结构后,将选择最小横向漂移值的区间与匹配序列,若此时横向漂移值小于阈值则匹配成功,若开启了全局比对功能,则此时将该序列与所匹配到的序列进行全局比对,全局比对后将会标记序列中不同的碱基位,若核心序列集中某条序列的某个碱基位被频繁标记则将会被视为错误碱基位进而进行纠正;若仍无法匹配,则将该条测序序列视为一个新的簇,将其加入到核心序列集中,并将其指定区间加入到树结构中。

5.根据权利要求1所述的一种核酸序列聚类方法,其特征在于,在步骤c和步骤d之间,还包括步骤k,所述步骤k为设置一个阈值,若某个簇内所含序列较少,则认为该簇为一个噪声簇,将其从核心序列集中舍弃,此外若选择多进程模式,则将不同进程间的核心序列集进行合并,但是序列集中不同的序列不会合并,若输入文件为带标签的数据集,则会进行耗时、准确率、正确率的计算,最后输出簇的结果以及核心序列集。

6.一种核酸序列聚类装置,其特征在于,包括如下单元:

参数初始化单元,用于初始化参数;

分流单元，用于对待测序列进行分流；

待测序列判断单元，用于判断待测序列是否为空；

结果输出单元，用于输出簇情况以及核心序列集；

检索单元，用于取出一条待测序列与树结构进行检索；

检索结果判断单元，用于判断是否能检索到相似核心序列；

判断结果划分单元，用于将该条待测序列与相似核心序列划为同簇，或者，将该待测序列加入核心序列集以及树结构；

全局比对纠错单元，用于进行全局比对并纠正核心序列错误。

7.一种计算机可读存储介质，其特征在于，包括程序或指令，当所述程序或指令在计算机上运行时，实现如权利要求1-5中任一项所述的核酸序列聚类方法。

8.一种计算机终端，其特征在于，包括存储器，以及与所述存储器通信连接的一个或多个处理器；

所述存储器中存储有可被所述一个或多个处理器执行的指令，所述指令被所述一个或多个处理器执行，以使所述一个或多个处理器实现如权利要求1-5中任一项所述的核酸序列聚类方法。

核酸序列聚类方法、装置、计算机可读存储介质、终端

技术领域

[0001] 本发明属于数据存储技术领域，尤其涉及一种核酸序列聚类方法、装置和计算机可读存储介质、终端。

背景技术

[0002] 核酸是脱氧核糖核酸(DNA)和核糖核酸(RNA)的总称，是由许多核苷酸单体聚合成的生物大分子化合物，为生命的最基本物质之一。核酸的研究涉及到生物、医疗、计算机等多个领域。

[0003] 测序是研究核酸的基本手段之一。通过测序技术可以将核酸分子的信息读取到计算机等存储介质中进而进行进一步的使用与分析。近年来，随着第二代测序技术的成熟，相关领域研究进入高速发展。第二代测序(Next-generation sequencing,NGS)又称为高通量测序，其开创性的引入了可逆终止末端，从而实现边合成边测序，在核酸复制过程中通过捕捉新添加的碱基所携带的特殊标记(引物)来确定核酸序列。二代测序有两个重要特点：1.高通量，二代测序能一次并行对几十、几百万条核酸分子进行测序；2.序列长度短，由于测序过程随着读长增长，基因簇复制的协同性降低，会导致测序质量下降，因此二代测序的读长不超过500bp。对于较长的基因组、宏基因组需要被打断成小片段再测序，测序完毕后再拼接。

[0004] 由于第二代测序技术测序时所产生的核酸序列数量过多将会难以进行下一步分析，为此必须使用聚类算法对测序后的序列进行分类进而求得原始序列簇再进行分析以提高效率，图2展现了核酸序列测序读取的流程图。然而在例如DNA存储等领域，为了提高读取序列的准确性，往往会经过多轮分子链扩增与深度测序，其产生的DNA序列可能达到上亿条，目前已有的聚类方法将花费不可容忍的耗时与内存占用。因此对大量核酸序列的聚类分析方法仍有待改进。CN110111843A提供了一种对核酸进行聚类的方法，但由于其采用了计算序列间编辑距离的方式来进行聚类，因此仍会需要大量计算时间，对于复杂的测序数据来说难以快速处理。

[0005] 目前核酸聚类算法的难点主要集中在以下几个方面：

[0006] (1)需要聚类的核酸序列数量巨大，往往有上千万甚至更高数量级，且簇的数量极多，传统的聚类算法需要消耗大量的时间和内存。目前传统聚类算法针对十万条以上序列时，所需要的耗时将超过10h，而且随着序列数量的增多耗时急剧上升。对于动辄几千万条的DNA序列几乎无法处理。

[0007] (2)核酸序列作为一种文本形式的序列串，无法使用传统基于欧氏距离的聚类算法。目前已有的核酸聚类算法仍使用编辑距离进行判别距离，因此具有较高的计算复杂度。目前仍未有线性计算复杂度的核酸聚类算法。

[0008] (3)核酸序列在进行扩增、测序的过程中会随机产生错误，这就势必会给聚类增加难度，对于部分错误率高的序列，目前已有的算法难以进行正确的分类。

发明内容

[0009] 针对上述现有技术中存在的技术问题，本申请的目的在于提出一种核酸序列聚类方法、装置和计算机可读存储介质、终端，通过构建了多个分支的树结构来对核酸序列的指定区间进行检索，进而避免了传统计算编辑距离所消耗的大量时间。

[0010] 为实现本申请的目的，本申请提供的技术方案如下：

[0011] 第一方面

[0012] 本申请提供了一种核酸序列聚类方法，包括如下步骤：

[0013] 步骤a：初始化参数；

[0014] 步骤c：判断待测序列是否为空，若是跳到步骤d；若否跳到步骤f；

[0015] 步骤d：输出簇情况以及核心序列集；

[0016] 步骤e：结束；

[0017] 步骤f：取出一条待测序列与树结构进行检索；

[0018] 步骤g：判断是否能检索到相似核心序列，若是跳到步骤h；若否跳到步骤j；

[0019] 步骤h：将该条待测序列与相似核心序列划为同簇；

[0020] 步骤i：可选地进行全局比对并纠正核心序列错误；

[0021] 步骤j：将该待测序列加入核心序列集以及树结构，跳到步骤c。

[0022] 其中，步骤a中，所述初始化参数包括初始化树结构的数量、树结构的长度、树结构选取的区间位置、纵向漂移值、横向漂移值阈值、进程数、输入文件格式、输出文件格式、核心序列集和树结构。

[0023] 其中，若为多进程模式，在步骤a之后，步骤c之前，还包括步骤b：对待测序列进行分流。

[0024] 其中，步骤f-j具体包括：

[0025] 取出一条待测序列与树结构进行检索，判断是否能检索到相似核心序列；首先，比较首段区间与首段树结构是否能够匹配，若能匹配到则停止后续树结构的检索，将该条待测序列与相似核心序列划为同簇，若开启了全局比对功能，则此时将该序列与所匹配到的序列进行全局比对，全局比对后将会标记序列中不同的碱基位，若核心序列集中某条序列的某个碱基位被频繁标记则将会被视为错误碱基位进而进行纠正；若首段区间无法匹配到，则再进行末端区间与末端树结构的匹配，若末端区间可以成功匹配，则停止后续树结构的检索，将该条待测序列与相似核心序列划为同簇，若开启了全局比对功能，则此时将该序列与所匹配到的序列进行全局比对，全局比对后将会标记序列中不同的碱基位，若核心序列集中某条序列的某个碱基位被频繁标记则将会被视为错误碱基位进而进行纠正；若末端也无法成功匹配，则进行中间区间的匹配在选取中间区间时，将会在原指定区间基础上，允许横向漂移纵向漂移数个碱基位置，进而允许选择多个区间，碱基区间检索树结构后，将选择最小横向漂移值的区间与匹配序列，若此时横向漂移值小于阈值则匹配成功，若开启了全局比对功能，则此时将该序列与所匹配到的序列进行全局比对，全局比对后将会标记序列中不同的碱基位，若核心序列集中某条序列的某个碱基位被频繁标记则将会被视为错误碱基位进而进行纠正；若仍无法匹配，则将该条测序序列视为一个新的簇，将其加入到核心序列集中，并将其指定区间加入到树结构中。

[0026] 其中，在步骤c和步骤d之间，还包括步骤k，所述步骤k为设置一个阈值，若某个簇

内所含序列较少，则认为该簇为一个噪声簇，将其从核心序列集中舍弃，此外若选择多进程模式，则将不同进程间的核心序列集进行合并，但是序列集中不同的序列不会合并，若输入文件为带标签的数据集，则会进行耗时、准确率、正确率的计算，最后输出簇的结果以及核心序列集。

[0027] 第二方面

[0028] 本申请提供了一种核酸序列聚类装置，包括如下单元：

[0029] 参数初始化单元，用于初始化参数；

[0030] 分流单元，用于对待测序列进行分流；

[0031] 待测序列判断单元，用于判断待测序列是否为空；

[0032] 结果输出单元，用于输出簇情况以及核心序列集；

[0033] 检索单元，用于取出一条待测序列与树结构进行检索；

[0034] 检索结果判断单元，用于判断是否能检索到相似核心序列；

[0035] 判断结果划分单元，用于将该条待测序列与相似核心序列划为同簇，或者，将该待测序列加入核心序列集以及树结构；

[0036] 全局比对纠错单元，用于进行全局比对并纠正核心序列错误。

[0037] 第三方面

[0038] 本申请提供了一种计算机可读存储介质，包括程序或指令，当所述程序或指令在计算机上运行时，实现如上述的任一项核酸序列聚类方法。

[0039] 第四方面

[0040] 本申请提供了一种计算机终端，包括存储器，以及与所述存储器通信连接的一个或多个处理器；

[0041] 所述存储器中存储有可被所述一个或多个处理器执行的指令，所述指令被所述一个或多个处理器执行，以使所述一个或多个处理器实现如上述的核酸序列聚类方法。

[0042] 与现有技术相比，本发明的有益效果为，本发明方法绕过了常规的计算编辑距离的方式，基于多个核酸序列之间指定区间内碱基排列的差异，对所述多个核酸序列进行分类，以确定正确的原始簇集合。不同核酸序列进行比对分类时，方法允许检索在存储介质中树结构的节点间移动来抵抗相同簇内序列轻微差异进而提高分类的准确性。本发明允许对分类的序列进行全部碱基的比对以提高序列分类时的准确率，同时允许分类结束后输出原始序列。在此基础上进一步提供了对核酸序列进行聚类的装置和计算机可读存储介质、终端。采用本发明的方法和设备可以快速将大量核酸序列进行分类，并得到原始簇与原始序列，进而进行核酸序列的后续分析。

附图说明

[0043] 图1为本申请提供的核酸序列聚类方法的流程图；

[0044] 图2为现有技术中核酸序列测序读取的流程图；

[0045] 图3为本申请中DNA所构成的树结构示意图。

具体实施方式

[0046] 需要说明的是，在不冲突的情况下，本申请中的实施例及实施例中的特征可以相

互组合。

[0047] 以下结合附图和具体实施例对本发明作进一步详细说明。应当理解，此处所描述的具体实施例仅仅用以解释本发明，并不用于限定本发明。

[0048] 需要说明的是，本申请涉及基因测序，具体解决的问题是对已有的测序数据进行一种聚类，进而还原出原有的序列信息，降低处理测序信息的难度。DNA 数据在分子状态时，需要用测序仪读取它们的信息，但是读取之前需要把DNA 分子链复制很多份，读取后就会产生大量的重复的DNA信息，为此就需要用一个聚类软件，将这些相同类的DNA信息分到一类里方便进行信息的读取。例如：有10条DNA链(分子状态)，要先把它们每一条扩增复制10份，然后丢到测序仪中把DNA链的信息读取出来。这种方式电脑中就有100条DNA链的信息了，但它们有很多链是重复的，影响使用。为此需要一个聚类方法，将它们重新聚成10类，方便接下来的使用。(实际中会有上千万甚至上亿条链)。本发明提供的核酸聚类方法，与目前已有的核酸聚类算法相比，可以对未识别的大量核酸序列进行聚类的同时，还具备对聚类后的核酸序列进行自动纠错与比对的功能，可以直接输出纠错后的核酸原始序列，从而大大减少测序读取后的处理时间。

[0049] 首先给出本申请中待测序列的定义，待测序列表示测序处理后还未被分类的核酸序列。本申请模型可以简述为：首先模型将会构造一个核心序列集(核心序列集一开始里面是空的)，然后将待测序列里的每一条与核心序列集进行比对，如果序列能与核心序列集中每一条比对上，则被成功划分到指定簇，否则作为一条新的核心序列加入到核心序列集。在将待测序列中的序列与核心序列集比对时，首先将核心序列集构造为一种树结构的索引，再与待测序列进行比较，进而避免了核心序列集增大所带来的比较时间增加的问题，此外允许检索时在树上进行节点漂移以减少序列错误所产生的影响。树结构和节点漂移算法是本发明的关键，以下将首先介绍这两点，之后将给出方法的全部流程与描述。

[0050] 树结构

[0051] 树结构是一种重要的非线性数据结构。它是数据元素按分支关系组织起来的结构，很像自然界中的树那样。本申请在这里给出其定义：

[0052] 一棵树(tree)是由n(n>0)个元素组成的有限集合，其中：

[0053] (1)每个元素称为结点(node)；

[0054] (2)有一个特定的结点，称为根结点或根(root)；

[0055] (3)除根结点外，其余结点被分成m(m≥0)个互不相交的有限集合，而每个子集又都是一棵树。

[0056] 图3为DNA序列构成树结构的示意图，由于DNA序列只能由{A,T,G,C}组成，所以每个根下的结点至多有4个。本申请定义一个树的深度为L，如果这个树是由n(n>0)个长度为L的序列构造的。因此本申请可以显然得到一个定理：

[0057] 对于一个由M(m≥0)条序列构成的深度为L的树，任意序列对于树进行检索时，其计算复杂度都为O(L)。

[0058] 由此可见不论树结构包含多少序列(节点)，都不影响检索该树节点所用的时间。

[0059] 节点漂移

[0060] 由于核酸分子进行扩增与测序时，会随机发生碱基丢失、增添、替换等错误。为了对抗测序后核酸序列中错误碱基所带来的干扰，本申请允许待测序列检索树结构的时候进

行一定程度的节点漂移来防止错误序列无法成功匹配到正确的核心序列。本申请将漂移分为横向漂移与纵向漂移：

[0061] 横向漂移：对于检索树某一根未存在指定节点时，将会对根下其余已存在节点进行检索，若存在其他节点，且节点作为子树时下一节点仍能匹配到，则允许漂移到其他节点。

[0062] 纵向漂移：对于本应该检索树的序列特定区间[a,b](b>a≥0)，当序列纵向漂移为t(t≤a)时，则实际使用[a+t,b+t]至[a-t,b-t]之间的全部滑动窗口区间检索树。

[0063] 由定义可以知道，通过横向漂移可以减少碱基替换错误所带来的检索影响。通过纵向漂移可以减少前序序列碱基增添、缺失所带来的影响。

[0064] 本申请方法内设一个核心序列集。核心序列集在聚类之前为空集，本申请聚类完后核心序列集包含全部原始数据集。因此，本申请会以核心序列的前端、中间端、后端等指定区间分别构造多个树结构，对于每一个进入核心序列集的序列，会将其指定区间添入到树结构中。未分类序列将会逐条使用指定区间在树结构上进行检索，若成功检索到树的某一序列的全部节点，则被匹配到核心序列集中的该核心序列，若无法被检索到，则作为一条新序列加入到核心序列集中，并将指定区间增添至树结构。另外，在检索的时候指定了节点漂移的参数，从而对于发生错误的序列仍可以成功检索到其同簇序列。

[0065] 由于树结构的特点，不论核心序列集如何扩大，都不影响待测序列检索树结构所用的时间。显然，本申请方法的时间复杂度是线性的，即数据集中每个序列都会被执行一遍算法，且第一条序列和最后一条序列处理时间理论上是相等的。此外由于算法对于未分类序列读取完后就可以释放内存，因此内存复杂度只与树的深度以及原始序列的大小有关，这大大减少了内存的损耗。为了提高本模型的实用性，还允许未分类序列与核心序列匹配上之后进行一步全局对比以提高核心序列的准确性，进而可以在聚类完后直接输出纠错后的核心序列集以简化读取数据的难度。此外，允许进行多进程运行以提高算法的执行速度，具体多进程的方法为：根据待测序列的首端碱基将待测序列进行分流至相同首段碱基的进程中进而进行一步初始筛分。

[0066] 如图1所示，给出了本申请方法的流程图，包括如下步骤：

[0067] 步骤a：初始化参数；

[0068] 本申请提供的方法允许自定义多种模型参数，包括但不限于树结构的数量、树结构的长度、树结构选取的区间位置、纵向漂移值、横向漂移值阈值(即若横向漂移值大于该阈值则放弃检索)、进程数(若大于一则为多进程模式，该值只能为4的指数，例如4,16,64)、输入文件格式(允许输入带标签以及不带标签的文件，允许输入fasta、fastq、txt格式的文件)、输出文件格式(允许输出簇的分类信息、核心序列集等信息，若输入文件为带标签格式，则还允许输出准确率、耗时、覆盖率等信息。)构建核心序列集和树结构，构建好的核心序列集和树结构将为空集，只有等测序序列进入进行聚类操作才会逐渐增大。需要注意的是，若为多进程模式，则不同进程间的核心序列集与树结构相互无关联。只有在全部序列聚类完毕后，才会将不同进程间的核心序列集合并。

[0069] 步骤c：判断待测序列是否为空，若是跳到步骤d；若否跳到步骤f；

[0070] 步骤d：输出簇情况以及核心序列集；

[0071] 步骤e：结束；

[0072] 步骤f:取出一条待测序列与树结构进行检索;

[0073] 步骤g:判断是否能检索到相似核心序列,若是跳到步骤h;若否跳到步骤j;

[0074] 步骤h:将该条待测序列与相似核心序列划为同簇;

[0075] 步骤i:可选地进行全局比对并纠正核心序列错误;

[0076] 步骤j:将该待测序列加入核心序列集以及树结构,跳到步骤c。

[0077] 需要说明的是,对于每一条待测序列,将其与树结构进行检索,若成功检索到匹配的核心序列,则将其分到所匹配的核心序列的簇中,若开启了全局比对功能,则与所匹配的核心序列进行全序列的比对,并对有出入的位置进行标记,进而对核心序列进行纠正。若无法匹配到核心序列,则将其作为一条新的核心序列加入到核心序列集中,并对特定区间加入到树结构中。具体算法细节为:首先比较首段区间与首段树结构是否能够匹配(即匹配到的索引所产生的横向漂移值小于阈值),若能匹配到则停止后续树结构的检索,若开启了全局比对功能,则此时将该序列与所匹配到的序列进行全局比对(全局比对的算法与具体流程不是本专利所涉及的重点,目前已经有非常成熟的全局比对算法,本申请的程序允许提供一个接口,可以直接使用已有的全局比对算法),全局比对后将会标记序列中不同的碱基位,若核心序列集中某条序列的某个碱基位被频繁标记则将会被视为错误碱基位进而进行纠正;若首段区间无法匹配到,则再进行末端区间与末端树结构的匹配。若末端区间可以成功匹配,则跟上述内容一致,进行可选择的全局比对与纠错功能;若末端也无法成功匹配,则进行中间区间的匹配;在选取中间区间时,将会在原指定区间基础上,允许前后平移纵向漂移值个碱基位置,进而允许选择多个区间,例如原定中间区间为第40到第60个碱基,若纵向漂移值为2,则最终检索树结构的碱基区间为[38,58], [39,59],[40,60],[41,61],[42,62],这一系列的碱基区间检索树结构后,将会选择最小横向漂移值的区间与匹配序列,若此时横向漂移值小于阈值则匹配成功,进行进行可选择的全局比对与纠错功能;若仍无法匹配,则将该条测序序列视为一个新的簇,将其加入到核心序列集中,并将其指定区间加入到树结构中。

[0078] 其中,若为多进程模式,在步骤a之后,步骤c之前,还包括步骤b:对待测序列进行分流。1.若为多进程模式,则进行数据集分流,将已有的数据集分流给不同的进程。具体的分流方式为:由于测序序列首段错误率较低,我们将根据首段的信息进行分流,例如若进程数为4,则将测序序列第一个碱基根据“C”“G”“T”“A”进行分流;若进程数为16,则根据测序序列前两个碱基进行分流。根据这种分流方式,我们可以确保相同簇的序列分到相同进程中,且确保一个测序序列不会被分到两个不同的进程中。

[0079] 其中,步骤f-j具体包括:

[0080] 取出一条待测序列与树结构进行检索,判断是否能检索到相似核心序列;首先,比较首段区间与首段树结构是否能够匹配,若能匹配到则停止后续树结构的检索,将该条待测序列与相似核心序列划为同簇,若开启了全局比对功能,则此时将该序列与所匹配到的序列进行全局比对,全局比对后将会标记序列中不同的碱基位,若核心序列集中某条序列的某个碱基位被频繁标记则将会被视为错误碱基位进而进行纠正;若首段区间无法匹配到,则再进行末端区间与末端树结构的匹配,若末端区间可以成功匹配,则停止后续树结构的检索,将该条待测序列与相似核心序列划为同簇,若开启了全局比对功能,则此时将该序列与所匹配到的序列进行全局比对,全局比对后将会标记序列中不同的碱基位,若核心序

列集中某条序列的某个碱基位被频繁标记则将会被视为错误碱基位进而进行纠正;若末端也无法成功匹配,则进行中间区间的匹配在选取中间区间时,将会在原指定区间基础上,允许横向漂移纵向漂移数个碱基位置,进而允许选择多个区间,碱基区间检索树结构后,将选择最小横向漂移值的区间与匹配序列,若此时横向漂移值小于阈值则匹配成功,若开启了全局比对功能,则此时将该序列与所匹配到的序列进行全局比对,全局比对后将会标记序列中不同的碱基位,若核心序列集中某条序列的某个碱基位被频繁标记则将会被视为错误碱基位进而进行纠正;若仍无法匹配,则将该条测序序列视为一个新的簇,将其加入到核心序列集中,并将其指定区间加入到树结构中。

[0081] 其中,当全部待测序列全部聚类完毕,将会允许设置一个阈值,若某个簇内所含序列较少,则认为该簇为一个噪声簇,将其从核心序列集中舍弃,此外若选择多进程模式,则将不同进程间的核心序列集进行合并,但是序列集中不同的序列不会合并。若输入文件为带标签的数据集,则会进行耗时、准确率、正确率的计算。最后输出簇的结果以及核心序列集。

[0082] 在优选的实施例中,本申请提供了一种核酸序列聚类装置,包括如下单元:

[0083] 参数初始化单元,用于初始化参数;

[0084] 分流单元,用于对待测序列进行分流;

[0085] 待测序列判断单元,用于判断待测序列是否为空;

[0086] 结果输出单元,用于输出簇情况以及核心序列集;

[0087] 检索单元,用于取出一条待测序列与树结构进行检索;

[0088] 检索结果判断单元,用于判断是否能检索到相似核心序列;

[0089] 判断结果划分单元,用于将该条待测序列与相似核心序列划为同簇,或者,将该待测序列加入核心序列集以及树结构;

[0090] 全局比对纠错单元,用于进行全局比对并纠正核心序列错误。

[0091] 在优选的实施例中,本申请提供了一种计算机可读存储介质,包括程序或指令,当所述程序或指令在计算机上运行时,实现如上述的任一项核酸序列聚类方法。

[0092] 在优选的实施例中,本申请提供了一种计算机终端,包括存储器,以及与所述存储器通信连接的一个或多个处理器;

[0093] 所述存储器中存储有可被所述一个或多个处理器执行的指令,所述指令被所述一个或多个处理器执行,以使所述一个或多个处理器实现上述所述的核酸序列聚类方法。

[0094] 本申请使用了真实数据与模拟数据评估了本方法的性能。Erlich和Zielinski 等人提出了一种基于喷泉码的DNA编码技术,将可以复原的信息量比之前高了几个数量级。它们合成了152长的72000条DNA分子。其中DNA合成技术采用的是Twist公司技术,测序则采用了Illumina公司的MiseqV4技术。本申请采用了它们所合成的真实测序数据集ERR181698(共14654644条序列,隶属于 72000条原始序列)和ERR1817036(共34095791条序列,隶属于72000条原始序列)。此外选取了starcode、DBSCAN作为基准算法,其中starcode为目前公开算法中最快的DNA聚类算法,它主要通过编辑矩阵来求导朗文斯特距离。DBSCAN为传统聚类算法中具有较低复杂度的算法,它是基于密度的聚类算法具有小于二次计算复杂度。测试环境包含一台家用电脑、一台云服务器、以及一台超级计算机。为了得到高可信度的带标签数据集,我们使用了pear与bowtie 等生物软件对原始数据集与原始集合进了对比。

[0095]

数据集大小	多进程模式	单线程模式	starcode	DBSCAN
1 万条	0.01	0.08	1.16	128.15
10 万条	0.15	0.85	13.15	13500
100 万条	0.71	9.9	724.75	超过 10 小时
1000 万条	11.29	112	超过 10 小时	超过 10 小时

[0096] 表一:方法的耗时对比(单位:秒)

[0097] 对ERR181698数据集分别提取了不同数量级的数据集合作为基准数据集,以更为量化的展示聚类效果,表一展示了不同算法之间的耗时对比。从表中可以看出本申请方法的聚类速度越高于starcode以及DBSCAN,且多核运行将显著提高算法的聚类速度。此外,表二展现了本方法在真实数据集下的准确率情况,可以看出本方法在真实数据集上具有极高的准确率。

[0098]

数据集	多进程模式	单线程模式
1 万条数据基准集	1	1
10 万条数据基准集	0.9999	0.9999
100 万条数据基准集	0.9993	0.9999
1 万条数据基准集	0.9952	0.9994
ERR1816980 数据集	0.9935	0.9992
ERR1817036 数据集	0.9941	0.9993

[0099] 表二:方法的准确率

[0100] 为了展现本方法对于极大量数据集下的聚类效果,模拟了100亿条DNA测序序列集(测序深度为1000,错误率为千分之四),这也是目前已知DNA存储领域最大的模拟数据集。本申请在一台超算服务器上进行了多线程的实验,实验结果表明本方法在多进程下可以以约4h的耗时将一百亿条数据进行聚类,聚类准确率为99.99%。证明本方法面对大规模数据集仍可以在短时间内聚类完毕。

[0101] 需要说明的是,本申请中未详述的技术方案,采用公知技术。

[0102] 以上所述仅是本发明的优选实施方式,应当指出的是,对于本技术领域的普通技术人员来说,在不脱离本发明原理的前提下,还可以做出若干改进和润饰,这些改进和润饰也应视为本发明的保护范围。

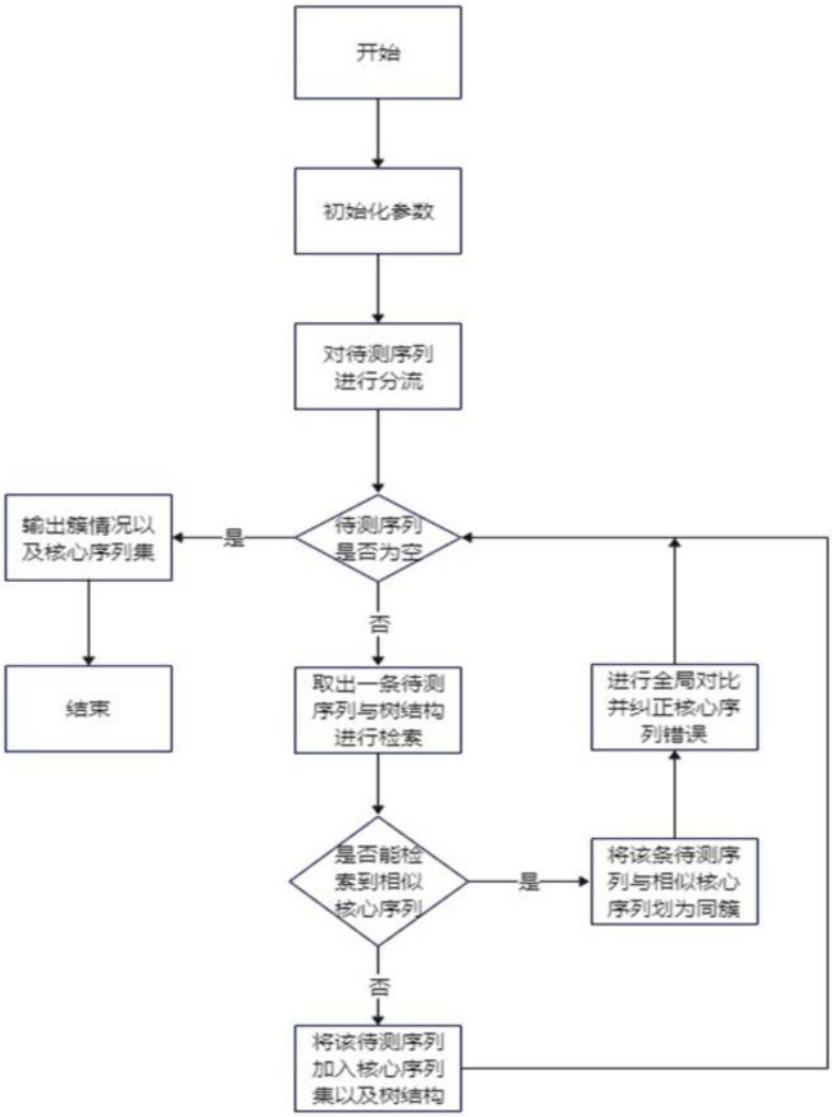

开始
初始化参数
对待测序列进行分流
待测序列是否为空
是
输出簇情况以及核心序列集
结束
否
取出一条待测序列与树结构进行检索
是否能检索到相似核心序列
是
将该条待测序列与相似核心序列划为同簇
进行全局对比并纠正核心序列错误
否
将该待测序列加入核心序列集以及树结构

图1

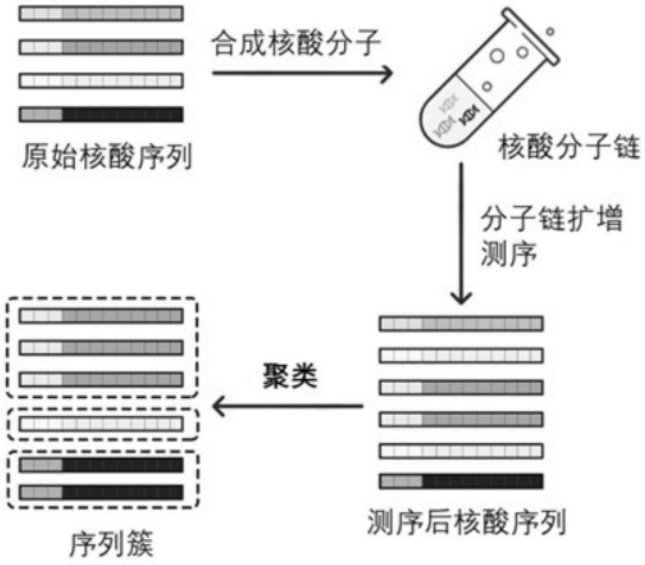

合成核酸分子
原始核酸序列
核酸分子链
分子链扩增
测序
聚类
测序后核酸序列
序列簇

图2

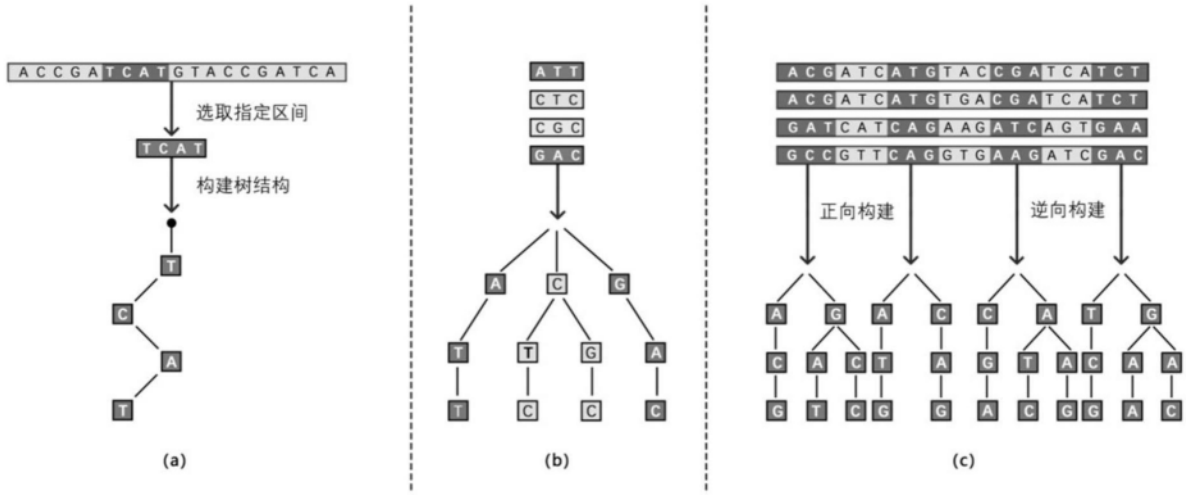

ACCGATCATGTACCGATCA
选取指定区间
TCAT
构建树结构
(a)
ATT
CTC
CGC
GAC
(b)
ACGATCATGTACGATCATCT
ACGATCATGTGACGATCATCT
GATCATCAGAAGATCAGTGAA
GCCGTTCAGGTGAAGATCGAC
正向构建
逆向构建
(c)

图3