

微生物目标片段中多拷贝区域的识别方法、装置及应用

技术领域

本发明涉及生物信息学领域，特别是涉及一种微生物目标片段中多拷贝区域的识别方法、装置及应用。

背景技术

由于生物样本中病原微生物的 DNA 浓度大多非常低，接近检测极限。而使用传统 PCR 或者实时 PCR 检测时常常缺乏检测灵敏度。其他方法如两步巢式 PCR 可以被用来提高灵敏度，但是方法耗时、成本高、准确性不好。因此，提高检测灵敏度至关重要。其中一种方式是在设计引物时寻找合适的模板区域，通常情况下会选用质粒和 16S rRNA。

但是，选用质粒做引物设计会产生一些问题：不是所有微生物都含有物种特异性质粒，并且有些微生物没有质粒。首先，质粒 DNA 的物种特异性不确定，有些物种质粒上的序列和其他物种质粒上的序列高度相似，那么基于质粒的 PCR 检测会产生假阳性或假阴性结果的高风险，许多临床实验室仍然需要使用其他 PCR 引物对来进行验证性实验。其次，质粒不具有普遍性，有些物种本身并没有质粒，那么就不能使用质粒来检测该物种，更不能通过在质粒上设计引物来提高检测灵敏度。例如，有研究报道，大约有 5% 的淋病奈瑟氏菌菌株因为没有质粒而无法检测到。

同样地，选用 rRNA 基因区域作为检测 PCR 的模板也存在一些问题：rRNA 基因虽然存在于所有微生物物种基因组中，并且常常有多个拷贝能够提高检测灵敏度。但事实上，并不是所有 rRNA 基因都是多拷贝。例如，在结核分支杆菌 H37Rv 中 rRNA 基因只有 1 个拷贝。另外，有一些 rRNA 基因序列的变化并不适合做检测。例如，在亲缘性很近的物种之间甚至是相同物种不同亚型的菌株之间，rRNA 基因由于序列太保守而无法满足物种特异性甚至是亚种特异性的要求。

发明内容

鉴于以上所述现有技术的缺点，本发明的目的在于提供一种微生物目标片段中多拷贝区域的识别方法、装置及应用。

本发明第一方面提供一种微生物目标片段中多拷贝区域的识别方法，所述方法至少包括以下步骤：

S100：寻找候选多拷贝区域：对微生物目标片段进行内部比对，寻找相似性满足预设值

的待测序列对应的区域作为候选多拷贝区域，所述相似性是指待测序列的覆盖率与匹配率的乘积；

S200：验证获得多拷贝区域：获得候选多拷贝区域拷贝数的中值；若候选多拷贝区域拷贝数的中值大于 1，则记为多拷贝区域。

本发明第二方面提供一种微生物目标片段中多拷贝区域的识别装置，所述装置至少包括：

候选多拷贝区域寻找模块，用于对微生物目标片段进行内部比对，寻找相似性满足预设值的待测序列对应的区域作为候选多拷贝区域，所述相似性是指待测序列的覆盖率与匹配率的乘积；

多拷贝区域验证获得模块，用于获得候选多拷贝区域拷贝数的中值；若候选多拷贝区域拷贝数的中值大于 1，则记为多拷贝区域。

本发明第三方面提供一种计算机可读存储介质，其上存储有计算机程序，该程序被处理器执行时实现前述微生物目标片段中多拷贝区域的识别方法。

本发明第四方面提供一种计算机处理设备，包括处理器及前述的计算机可读存储介质，所述处理器执行所述计算机可读存储介质上的计算机程序，实现前述微生物目标片段中多拷贝区域的识别方法的步骤。

本发明第五方面提供一种电子终端，包括：处理器、存储器、及通信器；所述存储器用于存储计算机程序，所述通信器用于与外部设备进行通信连接，所述处理器用于执行所述存储器存储的计算机程序，以使所述终端执行前述的微生物目标片段中多拷贝区域的识别方法。

本发明第六方面提供前述的微生物目标片段中多拷贝区域的识别方法，前述的微生物目标片段中多拷贝区域的识别装置，前述的计算机可读存储介质，前述的计算机处理设备，或前述的电子终端，用于检测微生物目标片段中多拷贝区域的用途。

如上所述，本发明的微生物目标片段中多拷贝区域的识别方法、装置及应用，具有以下有益效果：

与文献数据库对比，本发明的微生物目标片段中多拷贝区域的识别方法准确性高，灵敏度高，识别出未发现的多拷贝区域；可在组装不完整的基序中寻找重复序列；与 16srRNA 相比更加全面，16srRNA 不都是多拷贝；本系统不受限于是否存在全基因组序列，可以通过提供目标菌株和对比菌株的名称或者本地上传序列文件来提交运算任务。

附图说明

图 1 是本发明实施例的方法的流程图。

图 1-1 是比对序列的覆盖率与序列匹配率的计算结果图。

图 1-2 是本发明的多拷贝区域验证获得模块示意图。

图 2 是本发明实施例的装置示意图。

图 3 是本发明实施例中电子终端示意图。

具体实施方式

以下通过特定的具体实例说明本发明的实施方式，本领域技术人员可由本说明书所揭露的内容轻易地了解本发明的其他优点与功效。本发明还可以通过另外不同的具体实施方式加以实施或应用，本说明书中的各项细节也可以基于不同观点与应用，在没有背离本发明的精神下进行各种修饰或改变。

此外应理解，本发明中提到的一个或多个方法步骤并不排斥在所述组合步骤前后还可以存在其他方法步骤或在这些明确提到的步骤之间还可以插入其他方法步骤，除非另有说明；还应理解，本发明中提到的一个或多个步骤之间的组合连接关系并不排斥在所述组合步骤前后还可以存在其他步骤或在这些明确提到的两个步骤之间还可以插入其他步骤，除非另有说明。而且，除非另有说明，各方法步骤的编号仅为鉴别各方法步骤的便利工具，而非为限制各方法步骤的排列次序或限定本发明可实施的范围，其相对关系的改变或调整，在无实质变更技术内容的前提下，当亦视为本发明可实施的范畴。

请参阅图 1 至图 3。需要说明的是，本实施例中所提供的图示仅以示意方式说明本发明的基本构想，虽图式中仅显示与本发明中有关的组件而非按照实际实施时的组件数目、形状及尺寸绘制，其实际实施时各组件的型态、数量及比例可为一种随意的改变，且其组件布局型态也可能更为复杂。

如图 1 所示，本发明的微生物目标片段中多拷贝区域的识别方法，至少包括以下步骤：

S100：寻找候选多拷贝区域：对微生物目标片段进行内部比对，寻找相似性满足预设值的待测序列对应的区域作为候选多拷贝区域，所述相似性是指待测序列的覆盖率与匹配率的乘积；

S200：验证获得多拷贝区域：获得候选多拷贝区域拷贝数的中值；若候选多拷贝区域拷贝数的中值大于 1，则记为多拷贝区域。

相似性预设值可以根据需要确定。相似性预设值推荐应超过 80%，例如 85%，90%，95%，96%，97%，98%，99%，或 100%。

覆盖率=(相似序列的长度/(待测序列末端值-待测序列起始值+1))%

匹配率即为待测序列与另一序列比对时的 identity 值。两序列比对的 identity 值可以利用 needle、water 或 blat 等软件获得。

相似序列的长度是指：待测序列与另一序列比对时，相匹配的片段在待测序列中所占的碱基数，即相匹配的片段长度。

例如，一候选多拷贝区域对应的待测序列的数据情况如图 1-1 所示，其中，

序列 A 为待测序列，将序列 A 和 B 进行对比，两者相匹配的片段长度为 187，A 序列的起始值（即起始位置）为 1，末端值（即结束位置）为 187，则：

A 序列的覆盖率 = $(187 / (187 - 1 + 1)) * 100\% = 100\%$

序列 A 和序列 B 的匹配率对应 identity 为 98.4%。

则 A 和 B 的相似性 = $100\% * 98.4\% = 98.4\%$ 。相似性预设值设置为 80%，A 和 B 的相似性满足预设值，因此作为候选多拷贝区域。

两个进行对比的序列之间的各个碱基的位置不发生交叉（即两比对序列在微生物目标片段中是完全分离的，没有重合部分）。可以在比对前去除有区域重叠的比对序列对，也可以在比对后，去除有区域重叠的比对序列对获得的相似性值。例如，如图 1 所示，所述序列 A 位置 1-187，则序列 B 中各个碱基的位置不会出现在 1-187 之间。可以在计算完覆盖率和匹配率后，利用 uniq 函数进行去重。

步骤 S200 中，候选多拷贝区域拷贝数的中值的获得方法为：确定各候选多拷贝区域在微生物目标片段上的位置，获得待验证候选多拷贝区域的每个碱基位置上覆盖的其他候选多拷贝区域的个数，计算该待验证候选多拷贝区域的拷贝数的中值。所述其他候选多拷贝区域是指除了待验证候选多拷贝区域以外的候选多拷贝区域。

具体的，例如，如图 1-2 所示，第一行代表微生物目标片段序列，在微生物目标片段序列中，框线内的片段为待验证的候选多拷贝区域，第二行的数字为待验证的候选多拷贝区域中，各碱基对应的多拷贝数，图中灰色片段代表待验证的候选多拷贝区域以外的候选多拷贝区域（之后简称为重复片段）。从左边开始，框线中第一行的第一个碱基 A，由于该碱基对应应在 5 个重复片段中出现（即被 5 个重复片段覆盖），因此认为，与其位置对应的重复片段的数量为 5，则该位置的多拷贝数为 5；如图中框线中最后一个碱基 G，与其位置对应的重复片段的数量为 4，则该位置的多拷贝数为 4。以此类推，统计待验证的候选多拷贝区域的每个碱基位置上覆盖的重复片段的个数。统计结果图中参见第二行的多拷贝数，结合各个位置的拷贝数的数值，即可计算获得候选多拷贝区域拷贝数的中值。中值是指：是指将统计总体当中的各个变量值按大小顺序排列起来，形成一个数列，处于变量数列中间位置的变量值。

进一步的，步骤 S100 中，微生物目标片段可以为一条链，也可以为不完整的多条基序。

当微生物目标片段为不完整的多条基序时，将各基序连接起来再寻找候选多拷贝区域。

基序连接的顺序没有特别限定，可以按照任意的顺序连接。例如，将各基序按随机顺序连接成一条链。相似性满足预设值的区域中若包含不同基序，则将该区域按照原来的基序连接点进行切割，分成两个区域，分别判断两个区域是否是候选多拷贝区域。

微生物目标片段为不完整的多条基序指的部分微生物目标片段序列不是连续的单条序列，而是由多个不同大小的基序组成。基序是现有二代测序条件下短读长拼接不完整导致的。该方法也同样适用于三代测序等新技术产生的全基因组序列数据。

步骤 S1 中的微生物目标片段均来源于公共数据库，所述公共数据库选自主要是 [ncbi \(https://www.ncbi.nlm.nih.gov\)](https://www.ncbi.nlm.nih.gov)。

进一步的，所述方法还包括以下步骤：S101，将选定的相邻微生物目标片段进行两两比对，若出现相似性低于预设值的比对结果，则发出警报并显示目标菌株对应的筛选条件。

可以过滤人为错误或其他原因导致的异常数据。

本发明所述方法不受限于是否存在全基因组序列，可以通过提供目标菌株和对比菌株的名称或者本地上传序列文件来提交运算任务。从检测范围上比较，本方法可以涵盖所有的致病微生物，包括但不限于细菌、病毒、真菌、变形虫、隐孢子虫、鞭毛虫、微孢子虫、梨形虫、疟原虫、弓形虫、毛滴虫、动质体等。

在优选的实施方式中，在步骤 S200 中，还可计算候选多拷贝区域拷贝数的 95%置信区间。置信区间是指由样本统计量所构造的总体参数的估计区间，即对这个目标区域的整体拷贝数的区间估计。它体现了该目标区域拷贝数的真实值有一定概率落在测量结果周围的程度，其给出的是被测量参数的测量值的可信程度。

在计算候选多拷贝区域拷贝数的 95%置信区间时，以候选多拷贝区域的碱基数为样本数，候选多拷贝区域中各碱基对应的拷贝数值为样本值计算。

如图 1-2 所示，在这个长度为 500bp 的多拷贝目标区域中，每个碱基对应 1 个拷贝数值，那么这是一组共计 500 个的拷贝数值。

除了上文提到的拷贝数中值外，本发明使用这 500 个拷贝数值的 95%置信区间来衡量在显著性水平为 0.05 时，置信度为 95%的情况下，该多拷贝目标区域整体拷贝数的区间估计。在置信水平相同时，样本量越多，置信区间越窄，越接近均值。

所述微生物目标片段可以是微生物的全基因组，也可以是微生物的基因片段。

本发明的机理为：正常情况下，代表这 500 个拷贝数值的中值和 95%置信区间可以反映出该候选多拷贝区域的真实情况。本模块的设计除了进一步验证多拷贝以外，也可以排除一些特殊情况。例如，这 500bp 的候选多拷贝区域中仅有 5 个碱基的拷贝数为 1000，而剩余 495

个碱基的拷贝数为 1。那么这种情况下的拷贝数中值为 1，均值却为 10.99，95%置信区间为 (2.25-19.73)。很显然，虽然均值显示为多拷贝，但是中值已经不在这 95%置信区间范围内，候选多拷贝区域不可判为多拷贝。

如图 2 所示，本发明的微生物目标片段中多拷贝区域的识别装置，至少包括：

候选多拷贝区域寻找模块，用于对微生物目标片段进行内部比对，寻找相似性满足预设值的待测序列对应的区域作为候选多拷贝区域，所述相似性是指待测序列的覆盖率与匹配率的乘积；

多拷贝区域验证获得模块，用于获得候选多拷贝区域拷贝数的中值；若候选多拷贝区域拷贝数的中值大于 1，则记为多拷贝区域。

覆盖率=(相似序列的长度/(待测序列末端值-待测序列起始值+1))%。

匹配率即为待测序列与另一序列比对时的 identity 值。两序列比对的 identity 值可以利用 needle、water 或 blat 等软件获得。

相似性预设值可以根据需要确定。相似性预设值推荐应超过 80%，例如 85%，90%，95%，96%，97%，98%，99%，或 100%。

进一步的，两个进行对比的序列之间的各个碱基的位置不发生交叉。

可选的，所述候选多拷贝区域寻找模块还包括以下子模块：原始数据相似性比较模块，用于将选定的相邻微生物目标片段进行两两比对，若出现相似性低于预设值的比对结果，则发出警报并显示目标株种对应的筛选条件。用户可以根据反馈报告重新选择目标株种进入后台运算。

候选多拷贝区域寻找模块中，当微生物目标片段为不完整的多条基序时，将各基序连接起来再寻找候选多拷贝区域。

相似性满足预设值的区域中若包含不同基序，则将该区域按照原来的基序连接点进行切割，分成两个区域，分别判断两个区域是否是候选多拷贝区域。

所述基序按照任意的顺序连接。

所述多拷贝区域验证获得模块中还包括候选多拷贝区域拷贝数的中值的获得子模块，用于确定各候选多拷贝区域在微生物目标片段上的位置，获得待验证候选多拷贝区域的每个碱基位置上覆盖的其他候选多拷贝区域的个数，计算该待验证候选多拷贝区域的拷贝数的中值。

多拷贝区域验证获得模块还用于计算候选多拷贝区域拷贝数的 95%置信区间。

在计算候选多拷贝区域拷贝数的 95%置信区间时，以候选多拷贝区域的碱基数为样本数，候选多拷贝区域中各碱基对应的拷贝数值为样本值计算。

由于本实施例中的装置与前述方法实施例的原理基本相同，在上述方法和装置实施例中，对相同特征的定义、计算方法、实施方式的列举及优选实施方式的列举阐述可以互用，不再重复赘述。

需要说明的是，应理解以上装置的各个模块的划分仅仅是一种逻辑功能的划分，实际实现时可以全部或部分集成到一个物理实体上，也可以物理上分开。这些模块可以全部以软件通过处理元件调用的形式实现；也可以全部以硬件的形式实现；还可以部分模块通过处理元件调用软件的形式实现，部分模块通过硬件的形式实现。例如，获取模块可以为单独设立的处理元件，也可以集成在某一个芯片中实现，此外，也可以以程序代码的形式存储于存储器中，由某一个处理元件调用并执行以上获取模块的功能。其它模块的实现与之类似。此外这些模块全部或部分可以集成在一起，也可以独立实现。这里所述的处理元件可以是一种集成电路，具有信号的处理能力。在实现过程中，上述方法的各步骤或以上各个模块可以通过处理器元件中的硬件的集成逻辑电路或者软件形式的指令完成。

例如，以上这些模块可以是配置成实施以上方法的一个或多个集成电路，例如：一个或多个特定集成电路(Application Specific Integrated Circuit，简称 ASIC)，或，一个或多个微处理器(digital signal processor，简称 DSP)，或，一个或者多个现场可编程门阵列(Field Programmable Gate Array，简称 FPGA)或图形处理器(Graphics Processing Unit，简称：GPU)等。再如，当以上某个模块通过处理元件调度程序代码的形式实现时，该处理元件可以是通用处理器，例如中央处理器(Central Processing Unit，简称 CPU)或其它可以调用程序代码的处理器。再如，这些模块可以集成在一起，以片上系统(system-on-a-chip，简称 SOC)的形式实现。

在本发明的一些实施例中，还提供了一种计算机可读存储介质，其上存储有计算机程序，该程序被处理器执行时实现前述微生物目标片段中多拷贝区域的识别方法。

在本发明的一些实施例中，还提供了一种计算机处理设备，包括处理器及前述的计算机可读存储介质，所述处理器执行所述计算机可读存储介质上的计算机程序，实现前述微生物目标片段中多拷贝区域的识别方法的步骤。

在本发明的一些实施例中，还提供了一种电子终端，包括：处理器、存储器、及通信器；所述存储器用于存储计算机程序，所述通信器用于与外部设备进行通信连接，所述处理器用于执行所述存储器存储的计算机程序，以使所述终端执行实现前述微生物目标片段中多拷贝区域的识别方法。

如图3所示，展示本发明提供的一种电子终端的示意图。所述电子终端包括处理器31、

存储器 32、通信器 33、通信接口 34 和系统总线 35；存储器 32 和通信接口 34 通过系统总线 35 与处理器 31 和通信器 33 连接并完成相互间的通信，存储器 32 用于存储计算机程序，通信器 33、通信接口 34 用于和其他设备进行通信，处理器 31 和通信器 33 用于运行计算机程序，使电子终端执行如上图像分析方法的各个步骤。

上述提到的系统总线可以是外设部件互连标准(Peripheral Component Interconnect，简称 PCI)总线或扩展工业标准结构(Extended Industry Standard Architecture，简称 EISA)总线等。该系统总线可以分为地址总线、数据总线、控制总线等。为便于表示，图中仅用一条粗线表示，但并不表示仅有一根总线或一种类型的总线。通信接口用于实现数据库访问装置与其他设备(例如客户端、读写库和只读库)之间的通信。存储器可能包含随机存取存储器(Random Access Memory，简称 RAM)，也可能还包括非易失性存储器(non-volatile memory)，例如至少一个磁盘存储器。

上述的处理器可以是通用处理器，包括中央处理器(Central Processing Unit，简称 CPU)、网络处理器(Network Processor，简称 NP)等；还可以是数字信号处理器(Digital Signal Processing，简称 DSP)、专用集成电路(Application Specific Integrated Circuit，简称 ASIC)、现场可编程门阵列(Field-Programmable Gate Array，简称 FPGA)、图形处理器(Graphics Processing Unit，简称 GPU)或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件。

本领域普通技术人员可以理解：实现上述各方法实施例的全部或部分步骤可以通过计算机程序相关的硬件来完成。前述的计算机程序可以存储于一计算机可读存储介质中。该程序在执行时，执行包括上述各方法实施例的步骤；所述计算机可读存储介质可包括，但不限于，软盘、光盘、CD-ROM(只读光盘存储器)、磁光盘、ROM(只读存储器)、RAM(随机存取存储器)、EPROM(可擦除可编程只读存储器)、EEPROM(电可擦除可编程只读存储器)、磁卡或光卡、闪存、或适于存储机器可执行指令的其他类型的介质/机器可读介质。所述计算机可读存储介质可以是未接入计算机设备的产品，也可以是已接入计算机设备使用的部件。

在具体实现上，所述计算机程序为执行特定任务或实现特定抽象数据类型的例程、程序、对象、组件、数据结构等等。

前述的微生物目标片段中多拷贝区域的识别方法、微生物目标片段中多拷贝区域的识别装置、计算机可读存储介质、计算机处理设备或电子终端可应用在微生物 PCR 检测中。

具体的，用于模板序列的筛选。

前述的微生物目标片段中多拷贝区域的识别装置，前述的计算机可读存储介质，前述的计算机处理设备，或前述的电子终端，可以用于检测微生物目标片段中多拷贝区域的用途。

所述微生物选自细菌、病毒、真菌、变形虫、隐孢子虫、鞭毛虫、微孢子虫、梨形虫、疟原虫、弓形虫、毛滴虫或动质体中的一种或多种。

上述实施例仅例示性说明本发明的原理及其功效，而非用于限制本发明。任何熟悉此技术的人士皆可在不违背本发明的精神及范畴下，对上述实施例进行修饰或改变。因此，举凡所属技术领域中具有通常知识者在未脱离本发明所揭示的精神与技术思想下所完成的一切等效修饰或改变，仍应由本发明的权利要求所涵盖。