(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 107533587 B
(45)授权公告日 2021.10.08

(21)申请号 201680020777.1

(22)申请日 2016.04.08

(65)同一申请的已公布的文献号
申请公布号 CN 107533587 A

(43)申请公布日 2018.01.02

(30)优先权数据
62/145,026 2015.04.09 US

(85)PCT国际申请进入国家阶段日
2017.10.09

(86)PCT国际申请的申请数据
PCT/EP2016/057799 2016.04.08

(87)PCT国际申请的公布数据
WO2016/162504 EN 2016.10.13

(73)专利权人 皇家飞利浦有限公司
地址 荷兰艾恩德霍芬

(72)发明人 林志衡 S·卡玛拉卡兰

(74)专利代理机构 永新专利商标代理有限公司 72002
代理人 王英 刘炳胜

(51)Int.Cl.
G16B 30/00(2019.01) (续)

(56)对比文件
CN 103827889 A,2014.05.28
CN 103627800 A,2014.03.12
CN 104252627 A,2014.12.31
EP 2390810 A2,2011.11.30
KR 20120139908 A,2012.12.28
US 2014249036 A1,2014.09.04
刘丽梅 等.食品中微生物危害的风险评估建模方法改进与应用.《农业工程学报》.2014,第30卷(第6期),第279-286页.
Rogan Carr 等.Reconstructing the Genomic Content of Microbiome Taxa through Shotgun Metagenomic Deconvolution.《PLOS》.2013,第1-15页.
Adam L Bazinet 等.A comparative evaluation of sequence classification programs.《BMC》.2012,第1-13页. (续)

审查员 张帅

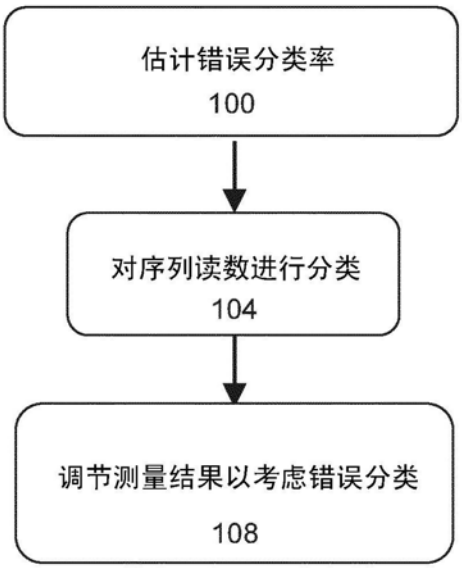
权利要求书2页 说明书8页 附图1页

(54)发明名称

用于估计样本中的分类单位内的微生物的量的方法和装置

(57)摘要

用于识别和量化存在于样本中的微生物的方法和装置。使用现有方法对序列读数进行分类,但是对分类结果进行校正以考虑如通过模拟所确定的预期被错误分类的读数的数量。利用对错误分类的读数的预期的数量的统计结果,能够使用线性最小二乘法(非负的或其他的)或其他相关技术来调节被分类至各种分类单位(例如,种类)的读数的数量,并且确定针对实际存在于样本中的那些分类单位的量的更准确的值,消除了被错误地确定为存在于样本中的分类单位中的微生物。



CN 107533587 B

[转续页]

[接上页]

(51)Int.Cl.

G16B 40/00(2019.01)

G16B 50/00(2019.01)

(56)对比文件

J Paul Brooks 等.The truth about metagenomics: quantifying and counteracting bias in 16S rRNA studies.《BMC》.2015,第1-14页.

Derrick E Wood 等.Kraken: ultrafast metagenomic sequence classification using exact alignments.《Genome Biology》.2014,第1-12页.

1.一种用于估计存在于样本中的分类单位内的微生物的量的计算机实施的方法,所述方法包括:

提供计算机处理器,所述计算机处理器被配置为:

(a)估计针对分类单位内的微生物的错误分类率,其中,所述计算机处理器被配置为利用经验性确定的读数长度和测序误差率使用所述分类单位内的所述微生物的基因组来模拟读数,并且对模拟的读数执行读数分类算法;

(b)接收样本中被分类至分类单位的列表的读数的数量的测量结果;

(c)使用所估计的错误分类率调节所接收的测量结果,以通过应用优化方法以由所估计的错误分类率和来自样本的被分类至分类单位的列表的读数的所述数量所确定的线性方程组找到所述样本中属于每个分类单位的读数的所述数量,来估计样本中属于每个分类单位的读数的所述数量;并且

(d)使用所估计的属于每个分类单位的读数的数量来估计来自存在于所述样本中的分类单位的微生物的数量。

2.根据权利要求1所述的计算机实施的方法,其中,所述计算机处理器还被配置为:通过使用所述分类单位中的微生物的基因组的基因组长度进行归一化、乘以基于所述基因组的鸟嘌呤-胞嘧啶含量的缩放因子、或者这两者,来调节对所述微生物的所述数量的所述估计。

3.根据权利要求1所述的计算机实施的方法,其中,被配置为估计错误分类率的所述计算机处理器被配置为确定被分类至感兴趣的分类单位的列表的模拟的读数的百分比,并且

其中,所述读数是通过从具有已知组成的微生物的样本接收序列读数而获得的。

4.根据权利要求1所述的计算机实施的方法,其中,所述样本包括多个种类的微生物,并且针对所述样本中的所述种类中的每个来计算所述错误分类率。

5.根据权利要求1所述的计算机实施的方法,其中,针对包括多个种类的微生物的数据的数据库中的所述种类中的每个来计算所述错误分类率。

6.根据权利要求5所述的计算机实施的方法,其中,针对所述数据库中的所述种类中的每个来接收所接收的测量结果,并且其中,针对所述数据库中的所述种类中的每个来调节所接收的测量结果。

7.根据权利要求1所述的计算机实施的方法,还包括接收来自所述样本的测序数据。

8.根据权利要求1所述的计算机实施的方法,其中,针对选自感兴趣的一个或多个分类等级的分类单位来估计所述错误分类率。

9.一种计算机可读介质,其包含用于估计存在于样本中的分类单位内的微生物的量的计算机可执行指令,所述介质包括:

(a)用于基于使用所述微生物的已知基因组的模拟的读数来估计针对分类单位内的所述微生物的错误分类率的计算机可执行指令,包括用于利用经验性确定的读数长度和测序误差率使用所述分类单位内的所述微生物的基因组来模拟读数的计算机可执行指令,和用于对模拟的读数执行读数分类算法的计算机可执行指令;

(b)用于接收样本中被分类至分类单位的列表的读数的数量的测量结果的计算机可执行指令;

(c)用于使用所估计的错误分类率来调节所接收的测量结果以通过应用优化方法以由

所估计的错误分类率和来自样本的被分类至分类单位的列表的读数的所述数量所确定的线性方程组找到所述样本中属于每个分类单位的读数的所述数量，来估计样本中属于每个分类单位的读数的所述数量的计算机可执行指令；以及

(d)用于使用所估计的属于每个分类单位的读数的数量来估计来自存在于所述样本中的分类单位的微生物的数量的计算机可执行指令。

10.根据权利要求9所述的计算机可读介质，还包括用于通过使用所述分类单位中的微生物的基因组的基因组长度进行归一化、乘以基于所述基因组的鸟嘌呤-胞嘧啶含量的缩放因子、或者这两者来调节对所述微生物的所述数量的所述估计的计算机可执行指令。

11.根据权利要求9所述的计算机可读介质，其中，所述的用于估计错误分类率的计算机可执行指令包括用于确定被分类至感兴趣的分类单位的列表的模拟的读数的百分比的计算机可执行指令，并且

其中，所述读数是通过从具有已知组成的微生物的样本接收序列读数而获得的。

12.根据权利要求9所述的计算机可读介质，其中，所述样本包括多个种类的微生物，并且所述计算机可执行指令针对所述样本中的所述种类中的每个来计算所述错误分类率。

13.根据权利要求9所述的计算机可读介质，其中，所述计算机可执行指令针对包括多个种类的微生物的数据的数据库中的所述种类中的每个来计算所述错误分类率。

14.根据权利要求13所述的计算机可读介质，其中，所述计算机可执行指令接收针对所述数据库中的所述种类中的每个的读数的所述数量的测量结果，并且其中，所述计算机可执行指令针对所述数据库中的所述种类中的每个来调节所接收的测量结果。

15.根据权利要求9所述的计算机可读介质，还包括用于接收针对所述样本的测序数据的计算机可执行指令。

16.根据权利要求9所述的计算机可读介质，其中，针对选自感兴趣的一个或多个分类等级的分类单位来估计所述错误分类率。

用于估计样本中的分类单位内的微生物的量的方法和装置

技术领域

[0001] 本发明总体涉及识别和量化存在于微生物组样本中的分类单位，并且更具体涉及利用预测误差率对样本测量结果的校正。

背景技术

[0002] 近来的医学研究集中于分析人类微生物组、共生体的生态群落、共享我们的身体空间的共生的和致病的微生物，作为潜在的疾病起因。一种研究的方法涉及对来自多种多样环境(诸如口、肠等)的细菌、病毒和/或真菌的基因组测序，被称为宏基因组学的研究领域。

[0003] 被用于研究宏基因组样本的现有方法遭受错误分类的读数，其会错误地识别存在于样本内的确切种类(species)和/或产生对那些种类的丰度的不准确的估计。这些错误分类可能提供微生物组样本的不准确的视图，妨碍对患者状况的准确分析和诊断。

[0004] 对存在于样本内的种类的更准确的识别以及对其丰度的更准确的量化能够产生对个人的疾病的状况或起因的更准确的识别。因此，存在对准确地识别和量化存在于微生物组样本中的种类和其他分类单位的方法和系统的需求。

发明内容

[0005] 提供该发明内容从而以简化的形式来介绍将在下文具体实施方式部分进一步描述的概念的选择。本发明内容并非旨在标识或排除所主张保护的主题的主要特征或基本特征，也并非旨在被用于辅助确定所主张保护的主题的范围。

[0006] 本发明的实施例总体涉及用于识别和量化存在于样本中的分类单位(例如，种类)的方法和装置。使用现有方法对序列读数进行分类，并且对分类结果进行校正以考虑预期被错误地分类的读数的数量(number)，如通过模拟或者利用已知量的微生物的测序实验所确定的。利用关于被错误分类的读数的预期数量的统计结果，线性最小二乘法(非负的或其他的)或者其他技术能够被用于确定针对实际存在于样本中的分类单位的量的更准确的值，并且消除被错误地确定为存在于样本中的分类单位。

[0007] 在一个方面中，本发明的实施例涉及一种用于估计存在样本中的分类单位内的微生物的量的计算机实施的方法。所述方法包括：提供计算机处理器，所述计算机处理器被配置为：估计针对分类单位内的微生物的错误分类率；接收样本中被分类至分类单位的列表的读数的数量的测量结果；并且使用所估计的错误分类率来调节所接收的测量结果以估计属于样本中的每个分类单位的读数的数量；并且使用所估计的属于每个分类单位的读数的数量来估计来自存在于样本中的分类单位的微生物的数量。在一个实施例中，所述计算机处理器还被配置为使用分类单位中的(一种或多种)微生物的(一个或多个)基因组的长度、GC含量、或者这两者来估计分类单位内的微生物的数量。

[0008] 在一个实施例中，被配置为估计错误分类率的所述计算机处理器被配置为：利用经验性确定的读数长度和测序误差率使用分类单位内的(一种或多种)微生物的(一个或多

个)基因组来模拟读数(或者从具有已知组成的微生物的样本来接收序列读数);对模拟的读数执行读数分类算法;并且确定被分类至感兴趣的分类单位的列表的模拟的读数的百分比。在一个实施例中,被配置为调节所接收的测量结果的所述计算机处理器被配置为:通过将最小二乘法(非负的或其他的)应用到由所估计的错误分类率和来自样本的被分类至分类单位的列表的读数的数量所确定的线性方程组,来调节所接收的测量结果。

[0009] 在一个实施例中,所述样本包括多个种类的微生物,并且针对样本中的种类(其被怀疑在样本中)中的每个并且针对具有相似基因组的密切相关的种类来计算错误分类率。在一个实施例中,针对包括多个种类的微生物的数据的数据库中的所述种类中的每个来计算错误分类率。可以针对数据库中的种类中的每个来接收所接收的测量结果,并且针对数据库中的种类中的每个来调节所接收的测量结果。

[0010] 在一个实施例中,所述方法还包括接收来自样本的测序数据。在一个实施例中,针对感兴趣的各种分类等级的分类单位来估计错误分类率,包括,但不限于种错误分类、属(genus)错误分类、以及亚种错误分类。

[0011] 在另一方面中,本发明的实施例涉及一种包含用于估计存在于样本中的分类单位内的微生物的量的计算机可执行指令的计算机可读介质。所述介质包括:用于估计针对分类单位内的微生物的错误分类率的计算机可执行指令;用于接收样本中被分类至分类单位的列表的读数的数量的测量结果的计算机可执行指令;以及用于使用所估计的错误分类率来调节所接收的测量结果以估计属于样本中的每个分类单位的读数的数量的计算机可执行指令;以及用于使用所估计的属于每个分类单位的读数的数量来估计存在于样本中的分类单位的微生物的数量的计算机可执行指令。在一个实施例中,所述介质还包括用于使用分类单位中的(一种或多种)微生物的(一个或多个)基因组的基因组长度、GC含量、或者这两者来估计分类单位内的微生物的数量的计算机可执行指令。

[0012] 在一个实施例中,用于估计错误分类率的计算机可执行指令包括:用于利用经验性确定的读数长度和测序误差率使用分类单位内的(一种或多种)微生物的(一个或多个)基因组来模拟读数(或者从具有已知组成的微生物的样本来接收序列读数)的计算机可执行指令;用于对模拟的读数执行读数分类算法的计算机可执行指令;以及用于确定被分类至感兴趣的分类单位的列表的模拟的读数的百分比的计算机可执行指令。在一个实施例中,用于调节所接收的测量结果的所述计算机可执行指令包括:用于通过将最小二乘法应用到由所估计的错误分类率和来自样本的被分类至分类单位的列表的读数的数量所确定的线性方程组来调节所接收的测量结果的计算机可执行指令。

[0013] 在一个实施例中,所述样本包括多个种类的微生物,并且所述计算机可执行指令针对被怀疑在样本中的每种种类并且针对具有相似基因组的密切相关的种类来计算错误分类率。在一个实施例中,所述计算机可执行指令针对包括多个种类的微生物的数据的数据库中的所述种类中的每个来计算错误分类率。可以针对所述数据库中的种类中的每个来接收所接收的测量结果,并且针对数据库中的种类中的每个来调节所接收的测量结果。

[0014] 在一个实施例中,所述计算机可读介质还包括用于接收针对所述样本的测序数据的计算机可读指令。在一个实施例中,针对感兴趣的各种分类等级的分类单位来估计错误分类率,包括,但不限于:种错误分类、属错误分类、以及亚种错误分类。

[0015] 根据对后续的详细描述的阅读以及对相关附图的回顾,表征当前非限制性实施例

的这些特征和优点以及其他特征和优点将是显而易见的。应当理解，前文的一般性描述以及以下的详细描述都仅仅是说明性的，而并非是对所主张保护的非限制性实施例的约束。

附图说明

[0016] 参考以下附图描述了非限制性和非穷举的实施例，在附图中：

[0017] 图1描绘了根据本发明的用于识别存在于样本中的微生物的方法的一个实施例的范例；并且

[0018] 图2图示了根据本发明的用于宏基因组样本分析的示范性系统的框图。

[0019] 在附图中，相似的附图标记通常贯穿不同的视图指代对应的部分。附图不一定按比例绘制，而是重点放在操作的原理和概念上。

具体实施方式

[0020] 下文参照随附的附图更完全地描述了各种实施例，附图形成其一部分并且示出了具体的示范性实施例。然而，实施例可以以许多不同的形式来实施，而不应当被解读为限于在本文中所阐述的实施例；而是提供这些实施例以使得本公开是透彻和完整的，并且将向本领域技术人员充分地传达实施例的范围。实施例可以被实现为方法、系统或设备。因此，实施例可以采取硬件实施、完全软件实施、或者组合软件和硬件方面的实施的形式。因此，以下详细描述不当被认为是限制性的。

[0021] 说明书中对“一个实施例”或“实施例”的引用意味着结合实施例所描述的具体特征、结构或特性被包括在本发明的至少一个实施例中。在说明书中各处出现的短语“在一个实施例中”不一定全部指代同一实施例。

[0022] 根据对被存储在计算机存储器中的非瞬态信号的操作的符号表示来呈现以下描述的一些部分。这些描述和表示是数据处理领域的技术人员用于最有效地将其工作的实质传达给本领域技术人员的手段。这样的操作通常需要对物理量的物理操纵。通常地但并非必须地，这些量采取能够被存储、转移、组合、比较以及以其他方式来操纵的电、磁或光学信号的形式。有时，主要出于公共使用的原因，将这些信号称为比特、值、元素、符号、字符、术语、数字等是方便的。此外，有时在不失一般性的情况下将需要对物理量的物理操纵的步骤的特定布置称作模块或代码设备也是方便的。

[0023] 然而，所有这些和相似的术语与适当的物理量相关联，并且仅仅是被应用于这些量的方便的标签。除非另有具体说明，否则如从以下论述中显而易见的，应当领会，在整个说明书中，利用诸如“处理”或“计算”或“运算”或“确定”或“显示”等术语的论述指代计算机系统或相似的电子计算设备的动作和过程，其操纵和变换被表示为计算机系统存储器或寄存器或者其他这样的信息存储、传输或显示设备内的物理(电子)量的数据。

[0024] 本发明的特定方面包括可以在软件、固件或硬件中实现的过程步骤和指令，并且当以软件实现时，所述指令可以被下载以驻留在由各种操作系统使用的不同平台上并且由其操作。

[0025] 本发明也涉及用于执行本文中的操作的装置。该装置可以是针对所要求的目的而特别构造的，或者其可以包括由被存储在计算机中的计算机程序选择性地激活或重新配置的通用计算机。这样的计算机程序可以被存储在计算机可读存储介质中，所述计算机可读

存储介质诸如但不限于任何类型的盘，包括软盘、光盘、CD-ROM、磁-光盘、只读存储器(ROM)、随机存取存储器(RAM)、EPROM、EEPROM、磁或光卡、固态存储器、专用集成电路(ASIC)、或者适用于存储电子指令的任何类型的介质，并且每个被耦合到计算机系统总线或企业服务总线。此外，在说明书中所提及的计算机可以包括单个处理器，或者可以是采用多处理器设计的架构从而获得以分布的方式的提高的计算能力。

[0026] 在本文中所提出的过程和显示并不固有地与任何特定的计算机或其他装置相关。也可以根据本文的教导与程序一起使用各种通用系统，或者其可以证明方便构造更专用的装置以执行所要求的方法步骤。根据下文的描述，针对这些系统的各种系统的所要求的结构将是明显的。另外，本发明不是参考任何特定的编程语言来描述的。将领会到，可以使用各种编程语言来实施如本文所述的本发明的教导，并且提供下面针对特定语言的任何参考以用于公开本发明的实现和最佳模式。

[0027] 此外，在本说明书中使用的语言主要是为了可读性和指导目的而选择的，而并非被选择用于描绘或限定本发明的主题。因此，本发明的公开内容旨在是说明性的而非限制在权利要求中所阐述的本发明的范围。

[0028] 概述

[0029] 本发明的实施例涉及用于量化宏基因组样本内的特定分类单位(例如，种类)的丰度的经改进的方法。可用于该任务的现有工具通常要么将读数映射到一组参考基因组，要么使用序列分析以特定分类级别(例如，族、属、种、亚种、系、亚系等)对读数进行分类。然而，这样的工具常常不正确地将一些读数错误地映射或错误地分类为属于不正确的分类单位。

[0030] 相比之下，本发明提供了通过以下操作来估计样本内的分类单位的丰度的方法和系统：通过模拟来量化所使用的读数分类方法(例如，Kraken方法)的典型错误分类率；并且应用优化技术(例如，线性最小二乘法)来考虑和校正通过模拟所确定的估计的错误分类率。该过程的结果是对样本中的种、亚种等的存在和/或丰度的更准确的估计。

[0031] 我们期望分类基于DNA或RNA数据的测序。对于基于DNA的输入，我们能够将读数分类至各种微生物的基因组，以量化不同分类单位的丰度。对于RNA数据，我们能够将读数分类到特定的基因(而非全基因组)，并且使用被分类至每个基因的读数的数量来表征宏基因组样本内的基因的表达水平。

[0032] 图1呈现了根据本发明的用于识别存在于样本(例如，微生物组样本)中的微生物的示范性方法。所述方法假设存在具有至少一种微生物(例如，细菌、真菌、病毒等)的样本，所述至少一种微生物具有这样的基因(如果被测序的话)，则所述基因将或多或少地对应于被存储在数据库中的基因组序列。所述数据库也可以存储难以获得完整基因组的一些微生物的部分或不完整的基因组序列，但是当在数据库中有不完整和完整两者的基因组序列时，也能够应用我们的方法。另外，在使用靶向测序方法的情况下，该数据库也可以被有意地仅填充部分基因组序列，以将分类方法限于特定的感兴趣的基因组区域(例如16S)。此外，所述数据库也可以存储针对感兴趣的基因的序列的列表，其可以被用于对来自基因的RNA读数进行分类并量化其表达水平。也假设了所述数据库存储被存储在所述数据库中的微生物的基因组(完全或部分)之间的分类关系。所述数据库可以是预先存在的数据库，或者可以是为了与本发明的实施例一起使用而专门创建的。如上文所提到的，为了准确地估

计样本中的微生物的存在和/或丰度,通常针对具有被包含在数据库中的基因组的每种微生物,所述方法估计针对与样本一起使用的读数分类方法的错误分类率(步骤100)。

[0033] 使用针对全基因组或靶向测序(例如,16S)的市售测序技术(例如,Illumina HiSeq或MiSeq)对样本进行测序。靶向16S测序对于对细菌样本进行测序可能更有效,而当样本被认为含有真菌或其他非细菌微生物时,全基因组测序可能更有利。

[0034] 在一个实施例中,将分类算法应用于测序过程的输出,以基于所提供的基因组数据库将每个读数分类为来自分类单位(步骤104)。用于在本发明的实施例中使用的一种合适的分类算法是Kraken,其从(在2015年2月17日访问的)http://ccb.jhu.edu/software/kraken/可获得。

[0035] 一旦已经对每个读数进行了分类,就可以计算统计结果,诸如来自样本中的感兴趣的分类单位的微生物的流行率。然而,众所周知,这样的统计结果包括由于基础读数分类中的误差和错误分类而导致的一些误差分量。本发明的实施例调节这些样本测量结果以考虑这些读数分类误差(步骤108)。

[0036] 校正序列错误分类

[0037] 由于针对读数分类方法的错误分类率可以在微生物之间变化,所以能够针对预期在样本中的每种微生物或者针对存在于基因组序列的数据库中的每种微生物执行量化错误分类的模拟过程。能够通过以下操作来确定对错误分类率的估计:使用针对所讨论的微生物的已知基因组(例如,作为从NCBI下载的.fasta基因组序列文件所获得的)和测序模拟器(诸如MetaSim,从(在2015年2月17日访问的)http:/ab.inf.uni-tuebingen.de/software/metasim/可获得))来模拟读数;将模拟的读数(例如,作为.fastq文件)提供至要被应用于实际样本的分类算法(例如,Kraken);并且通过针对模拟的读数的分类算法来计算错误分类率。备选地,也能够根据具有已知量的一种或多种微生物的测序实验来计算错误分类率。

[0038] 作为对测序模拟器的输入而提供的读数长度和测序误差率能够是在实践中针对将与样本(例如,Illumina、454等)一起使用的特定测序技术观测到的或者以其他方式经验性确定的值。然后,能够将测序模拟器的输出提供至读数分类算法。

[0039] 在一个实施例中,针对分类单位i中的微生物的错误分类率可以被表达为通过读数分类算法被分类至分类单位j的针对微生物所模拟的读数的分数,我们将其表示为a(j,i),其中,来自分类单位i的微生物选自前述微生物基因组的数据库。我们通常假设针对感兴趣的每个分类单位都有一个基因组,并且该基因组将用作感兴趣的分类单位的所有微生物的表示。在另一实施例中,针对微生物i的错误分类率可以被表达为针对通过读数分类算法被分类为微生物i之外的其他物质的模拟微生物i的读数的分数。

[0040] 在另一实施例中,可以仅针对被认为存在于待分析的样本中的分类单位i,j以及与一些读数可能被错误地分类的密切相关的分类单位(具有相似的基因组)来计算所估计的错误分类率。可以例如通过从样本获得的测序结果,或者通过关于样本的源的信息等,来通知该确定。例如,该信息可以是抽取样本的产地,或者可以是其他临床信息,诸如对患者的初步诊断。

[0041] 出于概念性的目的,值a(0,i)将表示来自在感兴趣的分类级别上仍未通过算法分类的微生物i的读数的分数(例如,当考虑在种类级别被分类的读数时,那么a(0,i)将表示

在种类级别未分类的读数的数量)。当我们具有我们希望对我们的微生物进行分类的n个感兴趣的分类单位时,将a(j,i)的个体值汇总成矩阵A,针对{0,1,...,n}中的j以及{1,...,n}中的i,创建大小为n+1乘以n个条目的矩阵。

[0042] 来自真正对应于来自微生物基因组的数据库的特定微生物i的样本的读数的数量能够被定义为x_i。个体值x_i能够被向量化为列x,亦即,再次地,是大小为n个条目,即,正在考虑的分类单位的数量。

[0043] 来自测序过程的通过分类算法被分类为来自微生物基因组的数据库(真阳性和假阳性两者)的微生物i的读数的数量能够被定义为b_i。个体值b_i能够被向量化为大小为n+1个条目的列b,即,在考虑中的分类单位的数量加上一(由于感兴趣的分类级别的未分类读数的数量)。

[0044] 利用这些定义,我们将期望矩阵方程Ax=b成立。然而,由于该过程是随机的,所以我们仅期望Ax=b期望符合大数定律的大量读数。在实践中,由于测序过程中固有的随机性(诸如测序错误)以及由于测序读数的有限数量,Ax=b并不是严格为真的。尽管如此,能够计算表示来自被分类至来自前述数据库中的每个生物体的样本的读数的数量的向量b,以及表示来自数据库的每个生物体的模拟的错误分类率的矩阵A。方程中的未知量是向量x。

[0045] 在一个实施例中,x被求解为使得:

[0046] $$\min_{x} \|Ax - b\|,$$

[0047] 该优化问题可以使用线性最小二乘法来求解,即:

[0048] $X=(A^TA)^{-1}A^Tb$,

[0049] 在其他实施例中,能够使用优化方法,诸如最小绝对值、最小截平方和等,并且这些方法常常具有发现向量x必须是非负的(例如,非负的线性最小二乘法)、必须是整数、或者这两者的版本。我们优选向量x为非负的并且具有整数值,因为其表示来自每个分类单位的读数的数量,其不能够为负数。在其他实施例中,能够使用使向量x中的非零条目的数量最小化的方法。这样的过程的结果能够说是“最简单的”答案,因为其要求来自分类单位的最少数量的微生物来解释所观测到的测序结果。

[0050] 在已经计算了估计来自对应于每种微生物的样本的读数的数量的向量x之后,向量x能够被归一化以解决一些微生物比其他微生物具有更长基因组的事实。基因组长度的差异将可能使经分类的读数的数量偏向有利于具有较长基因组的微生物,并且能够通过将向量x的每个条目x_i除以微生物i的基因组的长度来解决,得到针对样本中的微生物i的数量的归一化的估计。

[0051] 除了其长度之外或者代替其长度,通过考虑微生物的基因组的鸟嘌呤-胞嘧啶(GC)含量,能够进一步细化存在于样本中的微生物的估计的量。特定的测序技术难以捕获具有不平衡GC含量的基因组序列,因此,在微生物组样本中可能少算了具有包含GC重/轻区域的基因组的微生物。调节过程能够通过例如将每个微生物计数乘以基于在数据库中反映的微生物基因组中的GC重/轻区域的频率而计算的缩放因子来考虑该系统性少算。

[0052] 对于普通技术人员将显而易见的是,在前述讨论中的步骤的次序不一定是规范的。例如,普通技术人员将认识到,能够在接收到测序结果之后计算针对分类算法的估计的误差,允许计算限于样本中所识别的分类单位的简化的误差矩阵。

[0053] 图2是根据本发明的用于宏基因组样本分析的示范性系统的框图。在该实施例中，计算单元200与微生物基因组数据源208和测序数据源204通信。

[0054] 计算单元200可以在各种实施例中采用各种形式。适于与本发明一起使用的示范性计算单元包括台式计算机、膝上型计算机、虚拟计算机、服务器计算机、智能电话、平板电脑、平板手机等。数据源204、208也可以采取各种形式，包括但不限于：结构化数据库(例如，SQL数据库)、非结构化数据库(例如，Hadoop聚类、NoSQL数据库)、或者运行在各种计算单元(例如，台式计算机、膝上型计算机、虚拟计算机、服务器计算机、智能电话、平板电脑、平板手机等)上的其他数据源。在本发明的各种实施例中，所述计算单元可以是异构的或同构的。在一些实施例中，数据源204可以是对样本中的至少一种微生物的基因组进行测序的测序仪器。在一些实施例中，数据源208可以是基因组数据的公共或私人可访问的数据库。

[0055] 可以在各种实施例中使用异构或同构的各种网络技术来互连系统的各部件。适合的网络技术包括但不限于：有线网络连接(例如，以太网、千兆以太网、令牌环等)和无线网络连接(例如，蓝牙、802.11x、3G/4G无线技术等)。

[0056] 在操作中，计算单元200向测序数据源204查询针对对来自微生物组样本的一种或多种微生物的测序数据。测序数据源204可以具有这样的信息，因为其已经对样本进行了这样的测试，或者其可以从执行这样的测试的一部仪器直接或间接地(即，通过数据输入或传输)接收这样的信息。

[0057] 在操作中，计算单元200向基因组数据源208查询关于由测序数据源204识别的一种或多种微生物的基因组的信息。基因组数据源208可以具有本地存储的这样的信息，或者其可以联系其他计算单元以根据需要获得相关的基因组信息。

[0058] 如上文所论述的，在已经接收到所请求的测序数据以及针对一种或多种微生物的基因组数据后，计算单元200继续以估计针对每种微生物的错误分类率。计算单元200通过利用经验性确定的读数长度和测序误差率使用针对微生物的基因组数据对读数进行模拟来这样做。备选地，也能够使用来自具有已知量的一种或多种微生物的真实测序实验的读数。读数分类算法被应用于模拟的读数或实验地生成的读数，并且然后，计算被分类至感兴趣的每个分类单位的模拟的读数的百分比，以确定错误分类率。

[0059] 计算单元200将读数分类算法应用于从测序数据源204所接收的实际读数，并且通过将诸如线性最小二乘法(非负的或其他的)的优化方法应用于如上文所论述的由经分类的读数的数量和所估计的错误分类率而确定的线性方程组，提供了对属于感兴趣的每个分类单位中的微生物的读数的数量的经改进的估计。如上文所论述的，感兴趣的分类单位能够被限于被怀疑存在于样本中或者存在于基因组数据208中的分类单位。

[0060] 计算单元200或者可以首先访问数据源204、208或者可以同时访问这两个数据源。在一些实施例中，计算单元200对于操作者是本地的，即，位于由操作者访问的局域网上。在其他实施例中，计算单元200由操作者通过诸如广域网或互联网的另一网络连接(未示出)进行访问，并且通过这样的网络连接将图形表示递送给操作者。在这些实施例中，计算单元200包括对这样的远程访问的设备通常的安全性和web服务器功能。

[0061] 尽管前述讨论集中在本发明的在种级别对样本中的微生物进行分类的实施例中，但是应当理解，一些分类算法也可以将序列读数分类(和错误分类)为属于属、亚种或其他分类等级。我们也可以选择将读数分类至任何任意的分类单位的集合，这可以基于由微生

物引起的诸如临床表型的特性。本发明的实施例通过向错误分类率矩阵A中添加附加条目a(1,i)来表示来自被分类至基因组数据库中的每个分类群组1的微生物i的读数的分数(其可以具有不同的分类等级,例如,属/亚种),以及通过添加针对基因组数据库中的每个分类群组1的额外条目b_1,来求解这些类型的分类算法。注意,除了这些条目之外,我们也可以添加表示不能够被分类至不同的分类等级的读数的数量的条目,这能够是有用的知识,因为例如一些读数可以在属级别进行分类,但不能在种级别进行分类。也可以在这些实施例中使用上文所论述的最小二乘法或其他方法,以找到与分类的和未分类的读数的观测到的数量最佳匹配的适当向量x。

[0062] 在另一实施例中,微生物的错误分类误差和分类不仅基于分类单位,而且还能够基于微生物的任意随机分组。这些分组可以基于诸如对人类健康的影响的标准。即使在相同的种类内,亚组能够在分子水平上形成具有独特特性的系,这可能导致致病能力、使用独特的碳源的能力、或者对抗微生物剂的抗性的差异。可以基于这些系对人类健康的影响——即共生微生物与致病微生物——对这些系进行分组。在附加的实施例中,我们可以将微生物分类为严格的病原体(例如,结核分枝杆菌和淋病奈瑟氏球菌)和机会性病原体(例如,金黄色葡萄球菌、大肠杆菌)。

[0063] 本发明的实施例具有若干有用的商业应用,包括对存在于宏基因组样本内的种类的识别、量化样本内的种类的存在、样本分析、以及对感染性疾病的识别。

[0064] 例如,以上参考根据本公开的实施例的方法、系统和计算机程序产品的框图和/或操作说明来描述本公开的实施例。在框中标注的功能/动作可以不按照任何流程图所示的次序发生。例如,取决于所涉及的功能/动作,连续示出的两个框实际上可以基本上并发执行,或者有时可以以相反的次序执行。另外,并非任何流程图中示出的所有框都需要被执行和/或实现。例如,如果给定的流程图具有包含功能/动作的五个框,则可以是仅执行和/或实现五个框中的三个框。在该范例中,可以执行和/或实现五个框中的三个的任意的框。

[0065] 在本申请中所提供的一个或多个实施例的描述和说明并不意图以任何方式限制或约束本公开的范围。在本申请中提供的实施例、范例和细节被认为足以传达所有物,并且使得他人能够制造和使用所要求保护的实施例的最佳模式。所要求保护的实施例不应当被解释为限于本申请中所提供的任何实施例、范例、或细节。无论是组合还是单独进行显示和描述,各种特征(结构和方法两者的)意图被选择性地包括或省略以产生具有特定特征集合的实施例。在已经提供了本申请的描述和说明之后,本领域技术人员可以设想到落入在本申请中体现的总体发明概念的更宽泛方面的主旨中的变化、修改和替代实施例,而不背离所要求保护的实施例的更宽的范围。

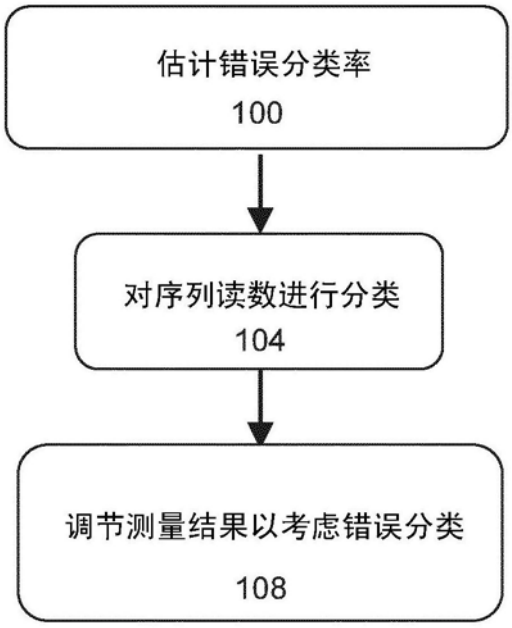

估计错误分类率
100
对序列读数进行分类
104
调节测量结果以考虑错误分类
108

图1

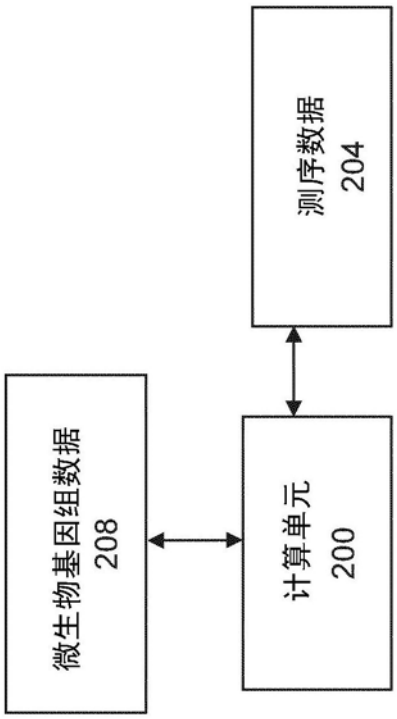

测序数据
204
计算单元
200
微生物基因组数据
208

图2