(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 115997255 A
(43)申请公布日 2023.04.21

(21)申请号 202180028019.5

(22)申请日 2021.03.10

(30)优先权数据
20162655.3 2020.03.12 EP

(85)PCT国际申请进入国家阶段日
2022.10.12

(86)PCT国际申请的申请数据
PCT/EP2021/056013 2021.03.10

(87)PCT国际申请的公布数据
WO2021/180771 FR 2021.09.16

(71)申请人 生物梅里埃公司
地址 法国迈合西-兰托阿嘞

(72)发明人 M·杰拉德·丹赛特 P·马埃

(74)专利代理机构 永新专利商标代理有限公司 72002
专利代理师 王健 林晓红

(51)Int.Cl.
G16B 30/00(2006.01)

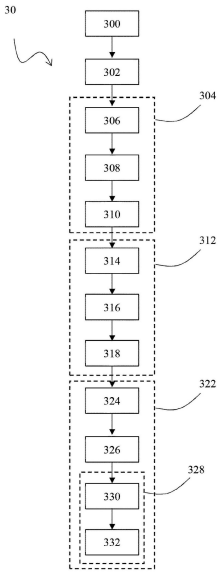
权利要求书3页 说明书22页
序列表2页 附图20页

(54)发明名称

从基因组预测细菌表型性状的分子技术

(57)摘要

确定细菌菌株的表型性状的方法包括对所述菌株的部分或全部基因组进行测序(400),并将用于预测所述性状的预定模型应用(402,404)于经测序的基因组,所述模型具有基因组序列组作为变量。根据本发明,选择所述组(304,312),以使得:所述组的组成基因组序列在所述细菌物种的基因组中的共现率高于预定阈值(304);和根据其在所述细菌物种的基因组中的共现率,所述组是所述基因组序列的聚类(312)。



CN 115997255 A

1.一种确定细菌菌株表型性状的方法,包括:

-对所述菌株的基因组进行部分或全部测序(400);

-将用于预测所述表型性状的预定模型应用于(402,404)经测序的基因组,所述模型具有基因组序列组作为变量,

其特征在于,选择(304,312)所述组,以使得:

-所述组的组成基因组序列在所述细菌物种的基因组中的共现率高于预定阈值(304);和

-根据其在所述细菌物种的基因组中的共现率,所述组是所述基因组序列的聚类(312)。

2.权利要求1的方法,其中所述基因组序列的第一子集是第一表型性状预测模型的未聚类变量的集合(314)。

3.权利要求2的方法,其中所述基因组序列的第二子集是与所述第一集合具有大于预定阈值的共现率的基因组序列的集合(316)。

4.权利要求2或3的方法,其中所述第一预测模型是自属于所述细菌物种的细菌菌株的基因组和表型性状的数据库训练的模型,所述学习属于简约学习类,特别是基于逻辑回归。

5.权利要求4的方法,其中所述学习是通过LASSO类型的逻辑回归来执行的。

6.前述权利要求中任一项的方法,其中所述基因组序列对应于从恒定长度的并出现在属于所述细菌物种的菌株的基因组中的基因组序列获得的压缩图的节点(306,308)。

7.权利要求6的方法,其中所述压缩图通过以下计算:

-从恒定长度的所述基因组序列计算字母A、T、G、C的De Bruijn图;

-压缩所述De Bruijn图的线性路径以获得所述压缩图。

8.前述权利要求中任一项的方法,其中基因组序列的聚类包括:

-通过在属于所述细菌物种的菌株的基因组中应用具有完美共现的序列聚类来获得第一组基因组序列(310);

-通过基于所述基因组中所述第一聚类的共现率应用所述第一组的第二聚类来获得所述预测模型的组(312)。

9.前述权利要求中任一项的方法,其中基因组序列的聚类(312)包括(318):

-计算树状图作为属于所述细菌物种的菌株的基因组中基因组序列的共现率的函数;和

-在预定高度聚类所述树状图以获得基因组序列的组。

10.权利要求8和9的方法,其中对第一组基因组序列计算所述树状图。

11.前述权利要求中任一项的方法,其中所述预测模型在属于细菌物种的细菌菌株的基因组和表型性状的数据库上进行训练,所述学习属于简约学习的类别,特别是基于逻辑回归(322)。

12.权利要求11的方法,其中所述学习是通过LASSO型逻辑回归执行的。

13.权利要求11或12的方法,其中计算简约学习的一组基因组序列的值(328):

-对于属于所述组的每个基因组序列,计算所述序列在数据库基因组中出现的向量;

-计算出现向量的平均向量,所述平均向量构成所述组的值。

14.前述权利要求中任一项的方法,其中应用用于所述细菌菌株的预测模型的一组基

因组序列的值等于所述细菌菌株的基因组中存在的所述组的基因组序列的百分比。

15.权利要求8的方法,其中应用用于所述细菌菌株的预测模型的一组基因组序列的值等于所述细菌菌株的基因组中存在的所述组的所述第一组的百分比(328)。

16.权利要求8的方法,其中通过以下计算应用用于所述细菌菌株的预测模型的一组基因组序列的值(328):

-对于所述组的每个第一组,计算所述细菌菌株的基因组中存在的所述第一组的基因组序列的百分比;和

-计算所述值,其等于所述组的第一组的百分比的平均值。

17.前述权利要求的方法,其中当构成所述序列的恒定长度的基因组序列的百分比大于取决于所述序列长度的预定阈值时,基因组序列存在于所述细菌菌株的基因组中。

18.一种鉴定预测细菌物种表型性状的基因组签名的方法,所述方法包括:

-建立(300)属于所述细菌物种的多个细菌菌株的基因组和表型性状的数据库;

-计算(306,308,310)描述所述数据库基因组的基因组序列的集合,并为每个所述序列计算所述序列在所述数据库基因组中出现的向量;

-从所述基因组序列集合中选择(314,316)具有高于预定阈值的共现率的序列,计算所述共现率作为出现向量的函数;

-根据其共现率对选择的基因组序列进行聚类(318),从而获得基因组序列组;

-训练(326)模型,用于预测细菌物种对抗生素的表型性状,以基因组序列组作为变量;

-鉴定(328)所述签名,其等于基因组序列组的联合。

19.权利要求18的方法,其中基因组序列的选择包括:

-选择由所述表型性状的第一预测模型的未聚类变量组成的基因组序列第一集合;

-从在第一选择中未被选择的基因组序列中选择与第一集合的基因组序列的共现率大于预定阈值的基因组序列。

20.权利要求18的方法,其中所述第一预测模型是自属于细菌物种的细菌菌株的基因组和表型性状的数据库训练的模型,所述学习属于简约学习类。

21.权利要求20的方法,其中所述学习是通过逻辑回归执行的,特别是LASSO类型的。

22.权利要求18至21任一项的方法,其中所述基因组序列对应于从恒定长度的并出现在属于所述细菌物种的菌株的基因组中的基因组序列获得的压缩图的节点。

23.权利要求22的方法,其中所述压缩图通过以下计算:

-从恒定长度的所述基因组序列计算字母A、T、G、C的De Bruijn图;

-压缩所述De Bruijn图的线性路径以获得所述压缩图。

24.权利要求18至23任一项的方法,其中所选基因组序列的聚类包括:

-通过在属于所述细菌物种的菌株的基因组中应用具有完美共现的序列聚类,获得第一组基因组序列;

-根据所述基因组中第一组序列的共现率,通过应用第二聚类获得预测模型的组。

25.权利要求18至24任一项的方法,其中所选基因组序列的聚类包括:

-计算树状图作为属于所述细菌物种的菌株的基因组中基因组序列的共现率的函数;和

-在预定高度聚类所述树状图以获得基因组序列的组。

26.权利要求24和25的方法,其中对第一组基因组序列计算树状图。

27.权利要求18至26任一项的方法,其中所述预测模型在属于所述细菌物种的细菌菌株的基因组和表型性状的数据库上进行训练,所述学习属于简约学习类。

28.权利要求27的方法,其中所述学习通过逻辑回归执行,特别是LASSO类型的。

29.权利要求27或26的方法,其中用于逻辑回归的一组基因组序列的值通过以下计算:

-对于属于所述组的每个基因组序列,计算在数据库基因组中所述序列出现的向量;

-计算出现向量的平均向量,所述平均向量构成所述组的值。

30.前述权利要求中任一项的方法,其中所述表型性状是所述细菌菌株的抗生素敏感性。

31.根据权利要求18至29任一项的方法鉴定的基因组签名在用于预测表型性状的分子测试,特别是聚合酶链式反应类型测试或基于DNA芯片的测试中的用途,所述测试至少部分靶向所述基因组签名。

32.一种存储计算机可执行指令的计算机程序产品,用于应用预测细菌菌株的表型性状,特别是抗生素敏感性的模型,所述应用如权利要求1至17任一项所述。

33.一种确定细菌菌株的表型性状,特别是抗生素敏感性的系统,其包含:

-用于对所述菌株的基因组进行部分或全部测序(400)的测序平台;

-计算机单元,其被配置为如权利要求1至17任一项所述将预测表型性状的预定模型应用(402,404)于经测序的基因组。

34.一种存储计算机可执行指令的计算机程序产品,用于执行鉴定预测细菌菌株的表型性状,特别是对抗生素敏感性的基因组签名的方法,所述方法如权利要求18至29任一项所述。

从基因组预测细菌表型性状的分子技术

发明领域

[0001] 本发明涉及应用于细菌基因组学的分子生物学技术领域，尤其涉及从细菌基因组预测细菌表型性状的领域。本发明特别可用于预测生物学样品中存在的细菌的抗生素敏感性和毒力。

现有技术

[0002] A.表型预测的分子技术

[0003] 细菌菌株对抗生素的敏感性，即其在基于施用于人类或动物的抗生素的治疗的背景下的敏感性或耐药性，不能被人类直接观察到。具体来说，直接观察菌株，即使使用显微镜，也无法确定其对抗生素的行为。在细菌背景中的体外诊断本质上在于使这种表型性质可观察，因此最终可为临床医生所用。在20世纪，体外诊断技术基本上结合了基于培养的样品制备技术(特别是使样品中存在的细菌菌株可见且可操作)以及在抗生素存在下光学测量菌株行为的技术。例如，传统微生物学实验室工作流程包括首先将疑似细菌感染的患者的样品散布在培养基上以产生细菌菌落，这些菌落在培养后对人类操作员或自动化系统可见。在第二阶段，当菌落足够大时，技术人员或自动化系统采集菌落，将其与不同浓度的抗生素混合并将混合物引入测量每种混合物的光密度的设备中并从中推断对抗生素的敏感性。由于光密度表明细菌增殖，它明确地表征了细菌的敏感性或耐药性：如果密度增加，这意味着尽管存在抗生素，但它仍在增殖，因此它对考虑中的抗生素浓度具有耐药性。

[0004] 现今，样品制备技术和基于光密度的测量技术的结合在面对全球快速进展的原核生物王国中存在重大限制，即获得对抗生素的多药耐药性，据估计这是到2050年造成超过癌症的更多死亡的原因。首先，这些技术与细菌无关。具体来说，根据所选择的培养基，某些菌株会生长而其它菌株不会，因此这些技术无法表征所有细菌物种的抗生素敏感性。其次，这些技术非常缓慢，因为其基于需要很长时间的细菌培养。因此，从采集样品开始，获得细菌的抗菌谱至少需要30小时。这种延迟不允许对系统性给予广谱抗生素混合物作为一线治疗的患者进行有效治疗。除了对患者造成的后果外，这种不恰当和大量施用抗生素还会加剧多重耐药细菌的选择压力，从而有助于其扩张。因此，目前认为传统的体外诊断技术越来越不适合治疗患者，在一定程度上是产生多药耐药的原因之一。

[0005] 分子生物学技术的成熟，特别是细菌DNA和/或RNA鉴定技术，如聚合酶链反应(PCR)、DNA芯片或测序，正在为实验室抗生素耐药性分析带来范式转变。首先，这些技术对细菌种类更加无关。例如，宏基因组技术使得处理生物学样品中的细菌DNA成为可能，而与存在的细菌种类无关。其次，这些技术的目标是在几个小时内提供结果，其中一些例如PCR甚至在不到20分钟内提供结果。另一方面，用于表征抗生素敏感性的分子技术基于表征所述敏感性的基因组签名(基因的不存在/存在，遗传突变，预测模型等)。图1以简化且非限制性的方式示例说明了两种细菌DNA表征技术，即PCR技术10和全基因组测序(WGS)技术20，其应用于微生物学工作流程以治疗疑似患有细菌感染的患者。这两种工作流程都起于患者生物学样品的收集12，然后应用PCR 10或WGS 20技术，每种技术都产生表征对一种或多种抗

生素敏感性的基因组签名的结果106、210，基于此结果，在14中由临床医生选择抗生素治疗并将其施用于患者。基本上，每种分子技术都需要在应用PCR本身104之前的所取样品的准备102、202，例如使用Biofir公司的Filmarray平台进行的嵌套PCR，或应用测序204，例如使用Illumina公司的MiSeq平台进行SBS测序。

[0006] 这两种分子技术之间的主要区别之一是基因组签名的性质。在PCR的情况下，基因组签名是分子的，因此是有形的：其被翻译成引入反应混合物中的引物，这些引物特异性靶向引入混合物中的细菌DNA序列，并且其检测通常是通过测量光信号来实现的。相比之下，在WGS的情况下，基因组签名是数字的，因为测序会产生数字基因组并且所述基因组的处理是计算机化的。尽管WGS技术至少允许实施PCR基因组签名，但最重要的是其允许复合利用数字基因组和使用通过PCR技术无法实施的抗生素敏感性预测模型。因此，WGS技术有利地基于基因组签名30的计算机设计，该设计借助于复合分析工具有利地利用大型基因组学和表型知识库，例如机器学习技术如简约约束逻辑回归(parsimoniously constraint logistic regression)。

[0007] 这是这样的情况，即所有分子技术都基于相同的技术基础，即测量来自细菌菌株的基因组信息并处理所述信息以提取有关菌株在抗生素存在下的行为的信息。此外，虽然这些计算机技术，尤其是测序技术，在本质上与那些与更传统的微生物学技术的光密度分析相关的技术不同，但它们并没有改变其进行的体外诊断的技术性质。例如，在感染性诊断的情况下，仍然需要应用处理生物学样品的技术以确定患者是否有细菌感染并了解感染性细菌在抗生素存在下的行为以便施用适当的抗生素治疗。

[0008] B.用于表型预测的分子技术的可解释性

[0009] 更特别地关注基因组签名，这是包括鉴定细菌基因组中先前确定的抗生素耐药性标志物的第一种方法，称为“直接关联”方法。虽然这些方法在导致耐药性的遗传机制是众所周知且简单的情况下是有效的，但就像在结核分枝杆菌(Mycobacterium tuberculosis)等物种中的大多数耐药机制一样，存在重大局限性：对许多物种和抗生素的耐药机制了解不完整，导致例如在不完整的数据库中难以考虑标志物预测能力的差异和抗生素敏感性的多因素方面(例如上位性、多个突变的组合等)等。面对这些困难，抗生素敏感性的遗传确定通过基于先进计算机技术的新方法、特别是通过监督机器学习技术而更有效地解决，其学习和应用架构可总结如下：

[0010] A.对于一组学习细菌菌株：

[0011] a.对每个菌株进行测序和表型表征(例如测量其最小抑制浓度和/或测量其对于一种或多种抗生素的敏感性-耐药性、中等敏感性或敏感性)；

[0012] b.从基因组和表型数据中训练用于预测对抗生素敏感性的计算机模型。

[0013] B.对于从步骤(A.a)中寻求其抗生素敏感性的新菌株：

[0014] a.对所述菌株进行测序；

[0015] b.将计算机预测模型应用于其数字基因组以确定其敏感性。

[0016] 上述一般描述涉及首先定义描述细菌基因组的机器学习变量。有许多方式可以描述所述基因组，其中一种是以“k-mer”描述，即构成基因组的长度为k(即碱基数)的核酸序列的列表。如M.Jaillard Dancette论文“Vers une cartographie des polymorphisms liésàla résistance aux antimicrobiens[Toward a mapping of polymorphisms related

to antimicrobial resistance]”,2018所述,这个描述特别适用于是单倍体和与真核基因组相比高度可塑性的细菌基因组。换言之,这个描述有效描述了细菌抗生素敏感性的遗传机制的多样性。

[0017] 然而,这个描述有许多可能对机器学习技术产生负面影响的缺点,包括以下方面:

[0018] a.k-mer是高度冗余的:那些覆盖保守基因组区域的k-mer可以共现,即在一组基因组中以系统方式存在或不存在,因此在统计学上是等效的;

[0019] b.一些k-mer不是特异于基因组区域的,因此其难以注释,即难以在结构或功能上(基因、突变等)进行表征;

[0020] c.基因组-敏感性关联是一个非常高维的问题,每个基因组的k-mer数目大于十万甚至百万,因此冗余和非特异性导致学习工具的高度相关变量。

[0021] 对于高风险领域,尤其是人类健康领域,降低变量的维数以提高预测模型的可解释性非常重要。具体来说,机器学习工具敏感于学习数据中的偏差,例如缺乏基因组多样性,以及敏感于与作为细菌基因组函数的敏感性不完全形成相关的偏差,例如未考虑不同基因组之间的强相关性。通过降低维数,预测模型变得更容易为学习工具专家和细菌基因组学专家所解释,从而能够检测到偏差并因此能够构建适当的学习数据或重新制定学习工具旨在解决的问题。相似地,由于预测模型更易于解释,如果在其分析后没有明显的偏差,则所述预测模型更容易验证用于高风险领域。

[0022] 在允许大幅降低维数的工具中,简约自动学习工具例如惩罚套索回归或基于决策树的工具,可以获得一千甚至一百数目级的许多预测性k-mer,即保留在预测模型中的k-mer。然而,这些工具在其中变量高度相关的高维环境中是不稳定的。因此,他们可以选择共同构成基因组单元的预测器变量,所述单元非必须具有任何生物学现实,因此预测模型仍然难以解释。某些技术可以考虑变量之间的强相关性,特别是基于弹性网络惩罚的回归,其将L1类型的惩罚与L2类型的惩罚相结合,从而导致选择相关预测器变量组。然而,应该注意的是,这种聚类仍然主要是算法上的,并且保留的变量组仍然难以从生物学上解释。

[0023] 其它工具,例如组套索工具,可以将描述性变量先验聚类到基因组单元中。在这种情况下,一个单元中的所有变量或者是可预测的或者不是,这取决于组套索策略对所述单元的选择或不选择。然而,由于出于上述原因对于以k-mer的描述难以解释,因此解释单元中的先验定义是困难的。特别地,这假设高维空间中的高相关现象被细菌基因组学专家很好地理解和形式化,因为缺乏对这些现象的了解或不完全了解,这会转化为机器学习算法中的偏差。

[0024] C.分子预测技术在生物学变异性中的应用

[0025] 与上述问题并行,当预测基于基因组序列组时,如果构成所述聚类的所有序列均以相同方式存在于基因组中,那么就会出现一个问题,即这组序列是否存在于菌株的基因组中。如果应用这样的标准,假设学习数据是完整涵盖细菌物种的所有基因组变异性。除了在给定时间难以判断学习数据的完整性这一事实之外,由于其基因组的非常显著的可塑性,在某些细菌物种中所述数据经常随着时间的推移变得不完整。此外,应用过于严格的标准会导致高假阳性率或假阴性率。

[0026] 刚刚描述的关于细菌菌株对一种或多种抗生素的敏感性的基因组预测的问题同样出现在对于菌株的表型性状例如其毒力或其核糖型的任何基因组测定中。

[0027] 发明描述

[0028] 本发明的目的是提供表型性状的基因组预测,特别是对抗生素敏感性的基因组预测,其更代表了所述菌株所属的细菌物种的所述性状的生物学现实,而同时保持预测的性能。

[0029] 为此,本发明的一个主题是确定细菌菌株的表型性状的方法,包括:

[0030] -对所述菌株的基因组进行部分或全部测序;

[0031] -将用于预测所述性状的预定模型应用于测序的基因组,所述模型具有基因组序列组作为变量。

[0032] 根据本发明,选择所述组,由此:

[0033] -在所述菌株所属的细菌物种的基因组中,所述组的构成基因组序列的共现率大于预定阈值;和

[0034] -根据其在细菌物种基因组中的共现率,所述组是所述基因组序列的聚类。

[0035] 换句话说,本发明使机器学习在预测表型性状时将同等权重的基因组序列视为组而不是自变量。因此,本发明人注意到所述预测不会降低性能,同时具有最大预测权重的学习组对应于生物学现实,例如对应于整个基因或特定突变组。这种方法已经在敏感性基因组起源得到充分证明的细菌物种上得到了充分验证。

[0036] 在下文中,描述了对抗生素敏感性的应用,应理解所述应用也适用于任何表型性状。

[0037] 根据本发明的一个实施方案,基因组序列的第一子集是第一敏感性预测模型的未聚类变量的集合。换句话说,这种过滤可以通过有利地使用无约束学习工具进行聚类,从一开始就消除没有预测权重的序列,因此这是先验最有效的。

[0038] 特别地,第一预测模型是在属于所述细菌物种的细菌菌株的基因组和敏感性数据库上训练的模型,所述学习属于简约学习类,有利地是逻辑回归,特别是LASSO型。术语“简约学习”是指定义预测模型的学习,其变量构成问题描述变量的子集,旨在最小化所述子集的大小而不影响模型的预测性能。

[0039] 换句话说,简约方法允许选择高度预测性基因组序列的第一集合,这些基因组序列将与具有非常相似的共现单元的其它序列连接。更具体地,简约方法可以鉴定参与抗生素敏感性的基因组区域。有利地,基因组序列的第二子集是与第一集合具有高于预定阈值的共现率的基因组序列的集合。换句话说,虽然简约方法允许开始鉴定感兴趣的基因组区域,但由于变量之间的相关性,所述方法通常无法详尽地描述所述区域。通过用第二集合的共现序列简约学习扩展获得的第一集合序列,生成了可以对所述区域进行详尽描述的序列集合。

[0040] 根据一个实施方案,所述基因组序列对应于得自碱基数为长度k的恒定长度的基因组序列或“k-mer”的压缩图的节点,并且出现在属于所述细菌物种的菌株的基因组中。特别地,通过以下方式计算所述压缩图:a)从所述恒定长度的基因组序列计算字母A、T、G、C的De Bruijn图,b)压缩De Bruijn图的线性路径以获得所述压缩图。以此方式获得了独特基因组序列的集合,消除了恒定长度的描述序列即k-mer中存在的冗余。

[0041] 根据一个实施方案,所述基因组序列的聚类包括:

[0042] -通过在属于所述细菌物种的菌株的基因组中应用完美共现的序列聚类,获得第

一组基因组序列；

[0043] -通过基于所述基因组中所述第一组的共现率应用所述第一组的第二聚类，获得预测模型组。

[0044] 以这种方式，将序列聚类为单一基因组单元，减少了变量的数目，同时也隐藏了冗余，所述序列同时系统地存在于用于构建压缩图的细菌物种菌株的每个基因组中，或者同时不存在，或者同时一些存在而一些不存在。

[0045] 根据一个实施方案，应用用于所述细菌菌株的预测模型的一组基因组序列的值等于所述细菌菌株的基因组中存在的所述组的所述第一组的百分比(328)。

[0046] 根据一个实施方案，通过以下方式计算应用用于所述细菌菌株的预测模型的一组基因组序列的值：

[0047] -对于所述组的每个第一组，计算所述细菌菌株的基因组中存在的所述第一组基因组序列的百分比；和

[0048] -计算等于所述组的第一组的百分比的平均值的值。

[0049] 根据一个实施方案，基因组序列的聚类包括a)基于属于所述细菌物种的菌株的基因组中基因组序列的共现率计算树状图，以及b)在预定高度对树状图进行聚类以获得基因组序列组。特别地，树状图是在第一组基因组序列上计算的。这种基于出现的分层方法允许对序列组的数目进行全面调整。通过调整聚类高度，可以优化所选组的生物学值。

[0050] 根据一个实施方案，预测模型在属于所述细菌物种的细菌菌株的基因组和敏感性数据库上进行训练，学习算法属于简约学习类，特别是基于逻辑回归。特别地，所述学习是通过逻辑回归执行的，特别是LASSO型。通过简约方法，选择最具预测性的组，这些组构成敏感性的基因组签名。

[0051] 有利地，用于简约学习的一组基因组序列的值如下计算：

[0052] -对于属于所述组的每个基因组序列，计算所述序列在数据库基因组中出现的向量；

[0053] -计算出现向量的平均向量，所述平均向量构成所述组的值。

[0054] 根据一个实施方案，用于应用于细菌菌株的预测模型的一组基因组序列的值等于在所述细菌菌株的基因组中存在的所述组的基因组序列的百分比。特别地，当构成所述序列的恒定长度的基因组序列的百分比高于预定阈值时，所述基因组序列存在于所述细菌菌株的基因组中。

[0055] 本发明的一个主题还是一种鉴定预测细菌物种的抗生素敏感性的基因组签名的方法，所述方法包括：

[0056] -建立属于所述细菌物种的多个细菌菌株的基因组和敏感性数据库；

[0057] -计算描述数据库中基因组的基因组序列集合，并对于每个所述序列计算所述序列在数据库的基因组中出现的向量；

[0058] -从所述基因组序列集合中选择共现率高于预定阈值的序列，所述共现率根据出现向量计算；

[0059] -根据其共现率对选择的基因组序列进行聚类，从而获得基因组序列组；

[0060] -训练模型以基因组序列组作为变量来预测细菌物种对抗生素的敏感性；

[0061] -将所述签名鉴定为等于基因组序列组的联合。

[0062] 本发明的主题还是根据用于鉴定上述类型的基因组签名的方法鉴定的基因组签名在预测表型性状、特别是对抗生素的敏感性的分子测试、特别是聚合酶链反应类型测试或基于DNA芯片的测试中的用途，所述测试至少部分靶向所述基因组签名。

[0063] 本发明的一个主题也是一种计算机程序产品，其存储计算机可执行指令，以应用模型来预测上述类型的细菌菌株的表型性状，特别是对抗生素的敏感性。

[0064] 本发明的一个主题还是一种用于确定细菌菌株对抗生素的敏感性的系统，其包含：

[0065] -用于对所述菌株的基因组进行部分或全部测序的测序平台；

[0066] -计算机单元，其被配置为将预定模型应用于测序的基因组，以预测上述类型的表型性状。

[0067] 本发明的主题也是一种存储计算机可执行指令的计算机程序产品，用于执行鉴定预测上述类型的细菌菌株的表型性状、特别是对抗生素的敏感性的基因组签名的方法。

[0068] 附图简述

[0069] 阅读以下描述将更好地理解本发明，该描述仅作为示例给出，并与附图相关联，其中相同的附图标记表示相同或相似的元件，并且其中：

[0070] 图1示例说明了现有技术中用于预测细菌菌株对抗生素的敏感性的两种分子技术；

[0071] 图2A和2B是根据本发明的学习阶段和预测阶段的流程图；

[0072] 图3示例说明了根据本发明的MAF单元生成；

[0073] 图4示例说明了根据本发明的方法中使用的聚类；

[0074] 图5是示例说明根据本发明选择预测阈值的ROC曲线；

[0075] 图6示例说明了对于细菌物种肺炎克雷伯菌(Klebsiella pneumoniae)和抗生素美罗培南，根据套索技术和根据本发明的称为“聚类-套索”的方法获得的预测模型的系数，以及压缩图的子图中的两个套索和聚类-套索模型的预测变量的位置，注释为blaKPC基因；

[0076] 图7示例说明了对于细菌物种肺炎克雷伯菌和抗生素头孢西丁，涉及套索模型中最具预测性的扩展MAF单元的压缩图的子图(左侧部分)，以及涉及聚类-套索模型中最具预测性的聚类的压缩图的子图(右侧部分)；

[0077] 图8示例说明了对于沙门氏菌(Salmonella)和抗生素四环素，套索模型系数的绝对值、聚类-套索模型系数的绝对值以及聚类-套索模型的前10个最具预测性的聚类中包括的unitig的数目；

[0078] 图9示例说明了对于沙门氏菌和抗生素四环素，涉及套索模型中最具预测性的扩展MAF单元的压缩图的子图(左侧部分)，以及涉及聚类-套索模型中最具预测性的聚类的压缩图的子图(右侧部分)；

[0079] 图10示出对于沙门氏菌和抗生素庆大霉素，套索模型系数的绝对值、聚类-套索模型系数的绝对值以及聚类-套索模型的前10个最具预测性的聚类中包括的unitig的数目；

[0080] 图11示例说明了对于沙门氏菌和抗生素庆大霉素，涉及套索模型中最具预测性的扩展MAF单元的压缩图的子图(左侧部分)，以及涉及聚类-套索模型中最具预测性的聚类的压缩图的子图(右侧部分)；

[0081] 图12示例说明了对于淋病奈瑟氏菌(Neisseria gonorrhoeae)和抗生素头孢克

肟,涉及套索模型的最具预测性的扩展MAF单元的压缩图的子图(左侧部分),以及涉及聚类-套索模型中最具预测性的聚类的压缩图的子图(右侧部分);

[0082] 图13示例说明了对于淋病奈瑟氏菌和抗生素头孢克肟,套索模型系数的绝对值、聚类-套索模型系数的绝对值以及聚类套索模型的前10个最具预测性的聚类中包括的unitig的数目;

[0083] 图14示例说明了对于金黄色葡萄球菌(Staphylococcus aureus)和抗生素四环素,涉及套索模型中最具预测性的扩展MAF单元的压缩图的子图(左侧部分),以及涉及聚类-套索模型中最具预测性的聚类的压缩图的子图(右侧部分);

[0084] 图15示出对于金黄色葡萄球菌和抗生素四环素,套索模型系数的绝对值、聚类-套索模型系数的绝对值以及聚类-套索模型的前10个最具预测性的聚类中包括的unitig的数目;

[0085] 图16示例说明了在基因组中检测基因组序列的流程图;

[0086] 图17示出基因组序列分解为k-mer;

[0087] 图18示例说明了检测基因组序列必须达到的基因组中k-mer存在的百分比,这些百分比取决于所述序列的长度;

[0088] 图19A和19B示例说明了作为肺炎克雷伯菌菌株分离株的测序覆盖深度的函数的套索预测和聚类-套索预测的AUC;和

[0089] 图20和21示例说明了作为包含肺炎克雷伯菌菌株的宏基因组样品覆盖的测序深度的函数的聚类-套索预测的AUC。

[0090] 发明详述

[0091] A.本发明的实施方案

[0092] 参考图2A和2B,根据本发明的方法包括训练模型的第一部分30,以预测属于给定细菌物种的细菌菌株对抗生素的作为所述菌株的细菌基因组的函数的敏感性,以及将所述模型应用于所述细菌物种的菌株以预测其未知敏感性的第二部分40。

[0093] 第一部分30起始于在300中建立所述物种的基因组和表型数据库。具体而言,例如从感染所述细菌物种的患者中收集一组菌株,对收集的每个菌株进行测序以获得其完整基因组,例如使用来自Illumina公司的MiSeq测序平台,并根据CLSI标准或EUCAST标准的临界浓度(或“断点”)建立抗菌谱以确定其对抗生素的敏感性,耐药性(“R”),中等(“I”)或敏感(“S”),例如使用来自bioMérieux公司的Vitek 2。优选地,基因组采用以本身已知的方式通过数字组装通过测序平台产生的数字序列(或“读取”)而产生的组装序列(或“重叠群”)的形式。每个菌株的完整数字基因组和抗生素敏感性存储在计算机数据库中以形成学习数据集和测试数据集。

[0094] 有利地但任选地,合并“耐药性”和“中间”状态以获得两种抗生素敏感性状态。以这种方式,定义了区分敏感(“S”)细菌菌株和非敏感(“NS”)细菌菌株的二元分类问题。例如,S状态用数字0编码,NS状态用数字1编码。

[0095] 步骤30的剩余部分由计算机执行,并在302开始将数据库聚类为两个,以获得学习数据库和测试数据库。优选地,所述聚类为“10倍”聚类,其中数据库的十分之九构成学习数据库,剩余的十分之一构成测试数据库。用于学习数据库的细菌菌株总数在下文中记为N。

[0096] 在下一步骤304中,确定细菌基因组的一组描述性变量。参考与图2平行的图3,在

306中，学习数据库的基因组G首先以k大小在15至50之间的k-mer描述，以便最佳捕获细菌物种的遗传变异性，同时限制k-mer的冗余性，例如k＝31。在随后的步骤308中，将k-mer转换成一组不同的基因组序列，而不丢失信息。为此，首先将k-mer转换为k阶和字母A、T、G、C的De Bruijn图DBG。然后通过自动对线性路径进行聚类将DBG图转换为压缩的De Bruijn图cDBG，因此没有分支。因此，k-mer被编码在一个图中，其节点彼此不同，对应于可变长度的序列，这些节点(以及等效地这些序列)被称为“unitig”。

[0097] 在随后的步骤310中，通过去重将unitig聚类成单元，该单元的次要等位基因频率(MAF)任选高于98％至99.5％之间的预定阈值“S_{MAF}”，例如99％。特别地：

[0098] a.压缩图cDBG的每个unitig都由一个二进制变量编码，该变量表示数据库的每个基因组中unitig的存在或不存在。因此获得了矩阵V，由此如果第j个unitig存在于数据库的第i个基因组中，则其元素$V_{i,j}$等于1，如果不存在则等于0。

[0099] b.然后为了计算unitig的MAF，对矩阵V的每个元素$V_{.,j}$，即与unitig相关联的每一列进行校正。如果第j个unitig的等位基因频率大于0.5，意味着它在数据库中超过50％的基因组中被观察到(即$\frac{1}{n}\sum_{i=1}^{n} V_{i,j} > 0.5$)，然后列$V_{.,j}$被转换，由此$\forall i, V_{i,j} = |1 - V_{i,j}|$。这种转换的优点是呈现矩阵V的最初是互补的相同两列，由此unitig的存在共现与其不存在共现相同。

[0100] c.然后针对相同的列过滤如此转换的矩阵V以便将完美共现的unitig分组成单元，即数据库的基因组中相同的单元不存在/存在。例如，如果转换的矩阵V的第一列$V._{,1}$与其第二列$V._{,2}$相同，则删除其中一列，其余列编码第一列和第二列的unitig联合，因此形成新的基因组单元；

[0101] d.频率低于(100-S_{MAF})％的单元，即对于S_{MAF}＝99％，由此$\frac{\sum_i V_{i,j}}{N} < 0.01$的单元然后被消除，意味着删除矩阵V的相应列。结果是矩阵X，由此如果数据库的第i个基因组中存在第j个MAF单元，则其元素$X_{i,j}$等于1，如果不存在则等于0。

[0102] 其余单元，在下文中称为“MAF单元”，是唯一的，彼此不同，并且随后在下文描述的机器学习工具中用作变量。这种聚类成单元和过滤显著减少了由k-mer中的基因组描述引起的冗余，而没有修改MAF单元关于其可能参与抗生素敏感性的固有信息值。步骤304-310更详细描述于M.Jaillard Dancette的论文“Vers une cartographie des polymorphismes liésàla résistance aux antimicrobiens[Towards a mapping of polymorphisms related to antimicrobial resistance]”,2018以及Jaillard M.et al.的文章“A fast and agnostic method for genome-wide association studies: Bridging the gap between k-mers and genetic events”,PLOS Genetics,2018中，并且例如使用M.Jaillard开发的软件DBGWAS(https://gitlab.com/leoisl/dbgwas)执行。

[0103] 学习部分30继续进行步骤312，将MAF单元聚类成有限数目的变量，这些变量既可预测抗生素敏感性又具有增强的生物学意义，在下文中称为“聚类”。

[0104] 步骤312开始于314，选择预测抗生素敏感性的MAF单元。有利地，使用LASSO型的惩罚逻辑回归工具执行该选择。更具体地，对于一组正值$\{\lambda_1, \lambda_2, \cdots, \lambda_m\}$中的每个值λ，解决以下优化问题：

[0105] $$\widehat{\boldsymbol{\beta}}(\lambda)=arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{p+1}}\sum_{i=1}^{N}\mathcal{L}\left(y_i,f\left(\boldsymbol{X}_{i,.}\right)\right)+\lambda\sum_{j=1}^{p}|\beta_j| \quad (1)$$

[0106] $$f\left(\boldsymbol{X}_{i,.}\right)=\beta_0+\sum_{j=1}^{p}\beta_j\boldsymbol{X}_{i,j} \quad (2)$$

[0107] 在该关系式中：

[0108] p是MAF单元的数目，因此是矩阵X中的列数；

[0109] N是用于制作学习数据库的细菌菌株数，因此是矩阵X中的行数；

[0110] y_i是与矩阵X的第i行相关的第i个细菌菌株对测量的抗生素的敏感性，即如果所述菌株是敏感的，则$y_i=0$，否则$y_i=1$；

[0111] $\mathcal{L}$是逻辑损失函数，量化测量的表型y_i和预测的表型$f(X_{i,.})$之间的差异，例如这两项的平方差或逻辑损失函数，由此$\mathcal{L}\left(y_i,f\left(\boldsymbol{X}_{i,.}\right)\right)=\log\left(1-e^{y_i-f(\boldsymbol{X}_{i,.})}\right)$。

[0112] 选择对于任何λ值是激活的并因此具有预测性的任何MAF单元，即如果$\exists\, k\in\{1,2,\ldots,m\}\setminus\widehat{\beta}_j(\lambda_k)\neq 0$，则选择第j个MAF单元。如已知的，LASSO工具沿着正则化路径$\{\lambda_1,\lambda_2,\cdots,\lambda_m\}$递增地激活其变量，横穿加入那些已经激活为路径$\{\lambda_1,\lambda_2,\cdots,\lambda_m\}$的一个或多个变量，以点亮最多N个变量。有利地选择$\{\lambda_1,\lambda_2,\cdots,\lambda_m\}$组以获得最大MAF单元的激活。例如，根据Friedman et al.的文章“Regularization Paths for Generalized Linear Models via Coordinate Descent”，Journal of Statistical Software，2010中描述的方法计算，或者例如通过在R中执行并且例如可从网站https://cloud.r-project.org/web/packages/glmnet/index.html可获得的package glmnet 3.0.2执行。特别地，这种方法允许预先选择正则化变量λ的数目m。例如，选择这个数字等于100。

[0113] 选择的MAF单元在下文中被称为“活性MAF单元”，活性MAF单元的数目被标注为p_a。标注为a_i的其秩(即其在矩阵X中的列索引)存储在$a=\{a_1,a_2,\ldots,a_i,\ldots,a_{p_a}\}$集中。

[0114] 在316中步骤310然后包括鉴定在步骤314中未选择但在基因组中与活性MAF单元具有最小共现率的MAF单元，然后将如此鉴定的MAF单元加入活性MAF单元列表。两个单元的共现率有利地通过其在矩阵X的相应列之间的相关性来测量。为此，计算维数$p_a\times p$的矩阵G，使得：

[0115] $$\forall(i,j)\in[1,p_a]\times[1,p],\boldsymbol{G}_{i,j}=cor\left(\boldsymbol{X}_{.,a_i},\boldsymbol{X}_{.,j}\right) \quad (3)$$

[0116] 其中cor是例如Bravais-Pearson线性相关，$\boldsymbol{X}_{.,a_i}$和$X_{.,j}$分别表示矩阵X的第a_i列和第j列。

[0117] 然后，选择与活性MAF单元具有高于预定义阈值s_1的相关性的任何MAF单元，并将其加入活性MAF单元的列表，因此选择的MAF单元集合在下文中称为“扩展的活性MAF单元”。换句话说，如果$\exists i\ /\ \boldsymbol{G}_{i,j}>s_1$，则选择第j个MAF单元，$j\in[1,p]$。$s_1$是一个数字，用于设置扩展的活性MAF单元之间所需的共现率。其在0.5和1之间，优选在0.8和1之间，并且有利地在0.9和0.95之间，例如0.95。扩展的活性MAF单元的总数，标注为p_e，然后远低于10^3而不是10^5到10^6级别的k-mer的初始数目$p(p_a<p_e<<p)$。扩展的活性MAF单元的秩，标注为e_1(即其在

矩阵X中的列索引)存储在$e=\{e_1,e_2,\ldots,e_l,\ldots,e_{p_e}\}$集合中。

[0118] 在随后的步骤318中,通过实施聚类分析工具明确定义高度共现的扩展的活性MAF单元的组或"聚类"。优选地,层次聚类基于从扩展的活性MAF单元之间的共现率计算的距离矩阵来实施。层次聚类例如在Bühlmann P.et al.的文章"Correlated variables in regression:Clustering and sparse estimation",Journal of Statistical Planning and Inference,2013中描述,或通过在R中执行的并且例如可得自网站https://www.rdocumentation.org/packages/stats/versions/3.6.2/source的Stats 3.6.2软件包的"hclust"函数执行,使用基于最近距离(或"单链接")的聚集标准。

[0119] 更具体地,层次聚类使用的距离矩阵是矩阵D,维数为$p_e\times p_e$,如下计算,其中$\boldsymbol{X}_{.,e_i}$和$\boldsymbol{X}_{.,e_j}$分别表示矩阵X的第e_i列和第e_j列:

[0120] $$\forall(i,j)\in[1,p_e]\times[1,p_e],\boldsymbol{D}_{i,j}=\left|1-cor\left(\boldsymbol{X}_{.,e_i},\boldsymbol{X}_{.,e_j}\right)\right| \quad (4)$$

[0121] 因此获得了扩展的活性MAF单元的树状图。然后将该树状图聚类在如图4所示的高度$1-s_2$,其中s_2是0和1之间的数字,固定聚类中的共现率,优选地在0.5和1之间,优选地在0.8和1之间,并且有利地在0.9和0.95之间,例如0.95。因此,树状图的"较低"部分定义了扩展的活性MAF单元根据其共现分布在其中的聚类。每个聚类也是唯一的,并且不与任何其他聚类共享任何MAF单元,特别是任何unitig。以数目p_c的聚类标注为$C_1,C_2,\ldots,C_j,\ldots C_{p_c}$,并且每个聚类$C_j$是包括在$c_j=\{c_{j,1},c_{j,2},\ldots,c_{j,l},\ldots,c_{j,p_j}\}$集合中的$p_j$扩展的活性MAF单元秩(即其在矩阵X中的列索引)的聚类。

[0122] 在随后的步骤320中,对于每个聚类C_j,构成其的unitig在unitig上通过以步长1滑动长度为k的窗口被分解成k-mer,k在15至50之间,例如k=31。对于每个聚类C_j,如果q_j用于表示构成其的unitig的数目,则存储分别与unitig相关联的k-mer的q_j集。为清楚起见,与聚类相关的k-mer和unitig分别称为"聚类k-mer"和"聚类unitig"。

[0123] 在随后的步骤322中,使用预测抗生素敏感性的训练模型,其变量是聚类$C_1,C_2,\ldots,C_j,\ldots C_{p_c}$。

[0124] 在324中通过计算学习数据库中每个基因组的每个聚类的值开始步骤322。有利地,这个值等于构成其的MAF单元的平均值。因此获得了维数$N\times p_c$的矩阵Y,由此:

[0125] $$\forall j\in[1,p_c],\boldsymbol{Y}_{.,j}=\frac{1}{p_j}\times\sum_{l=1}^{p_j}\boldsymbol{X}_{.,c_{j,l}} \quad (5)$$

[0126] 在326中,从学习数据库中训练几个抗生素敏感性预测模型,下一步328包括根据预定标准选择具有最佳性能的模型。

[0127] 有利地,使用惩罚逻辑回归训练预测模型,这使得减少模型最终保留的预测聚类的数目,同时保持高水平的性能。特别地,预测模型是根据以下关系的模型:

[0128] $$\begin{cases}\widehat{Su}=S\ if\ G(g)<S_p\\ \widehat{Su}=NS\ if\ G(g)\geq S_p\end{cases} \quad (6)$$

[0129] $$G(g)=\hat{B}_0+\sum_{j=1}^{p_c}\hat{B}_j\boldsymbol{Y}_{C_j} \quad (7)$$

[0130] 在所述关系中：

[0131] $\widehat{Su}$是属于所述细菌物种的细菌菌株的预测敏感性；

[0132] g是所述菌株的基因组；

[0133] $\hat{B}$是$\mathbb{R}^{p_c+1}$的参数向量；

[0134] $\boldsymbol{Y}_{C_j}$是基因组g的第j个聚类C_j的值；

[0135] S_p是预定的阳性阈值。

[0136] 步骤326通过使用简约逻辑回归工具例如LASSO型惩罚回归在学习数据库上训练模型G开始。具体地，通过根据以下关系求解优化问题，针对一组正值$\{\lambda_1^c, \lambda_2^c, \ldots, \lambda_i^c, \ldots, \lambda_M^c\}$的每个λ值计算模型G：

[0137] $$\widehat{\boldsymbol{B}}(\lambda) = arg \min_{\boldsymbol{B}\in\mathbb{R}^{p_c+1}} \sum_{i=1}^{N} \mathcal{L}\left(y_i, G\left(\boldsymbol{Y}_{i,.}\right)\right) + \lambda \sum_{j=1}^{p_c} |B_j| \quad (7)$$

[0138] 在所述关系中：

[0139] y_i是与矩阵Y的第i行相关的第i个细菌菌株对测量的抗生素的敏感性，即如果所述菌株是敏感的则$y_i=0$，否则$y_i=1$；

[0140] $\mathcal{L}$是逻辑损失函数，量化测量的表型y_i和预测的表型$G(Y_{i,.})$之间的差异，例如这两项的平方差或逻辑损失函数，由此$\mathcal{L}\left(y_i, f\left(\boldsymbol{X}_{i,.}\right)\right) = \log\left(1 - e^{y_i - f(\boldsymbol{X}_{i,.})}\right)$。

[0141] 例如，它是根据上文结合步骤314描述的方法计算的，数目M＝100。因此获得了100个模型，表示为$G_{\lambda_i^c}$，及因此根据关系(5)-(6)的100个预测模型，均取决于阈值S_p，在下文中表示为$G_{\lambda_i^c}(S_p)$。

[0142] 然后在328中从测试数据库中计算每个预测模型$G_{\lambda_i^c}(S_p)$的性能估计。所述性能评估使得可以并行计算阈值S_p。

[0143] 特别地，在步骤330中，对于测试数据库中的每个基因组，执行以下操作：

[0144] a.通过完全同源性检测基因组中存在的聚类k-mer，即如果k-mer在基因组中以相同形式存在，则检测到其；

[0145] b.如果在前面步骤中确定存在于基因组中的构成聚类的聚类k-mer的百分比高于第一预定检测阈值s_{uni}，则检测到该聚类unitig存在于基因组中；

[0146] c.计算一个指标，用于检测由几个聚类unitig组成的扩展的MAF单元存在于基因组中，根据以下任何选项定义：

[0147] i.如果在前述步骤中确定存在构成聚类的聚类unitig的百分比大于第二预定检测阈值s_{clus}，例如大于或等于20％例如25％的阈值，则该指标等于1。否则，该指标等于0。这个选项适用于以下实例；或者

[0148] ii.如果在前述步骤中确定存在所有构成聚类unitig，则该指标等于1，否则等于0；或者

[0149] iii.如果在前述步骤中确定存在至少一个构成聚类unitig，则该指标等于1，否则等于0。

[0150] iv.所述指标等于在前述步骤中已被检测到存在的构成其的聚类unitig的百分

比。

[0151] 有利地，检测阈值s_{uni}取决于聚类unitig的长度。特别地，对于在15至50之间的k，例如k＝31，注意试图了解其存在于基因组中的聚类unitig的长度L，如果K≤61，则s_{uni}＝90％，如果61<L≤100，则s_{uni}＝80％，以及如果100<L，则s_{uni}＝70％。

[0152] 在330中，步骤328通过计算等于在前述步骤中确定的构成其的扩展的MAF单元的检测指标的平均值的聚类值。应该注意的是，对于选项i、ii和iii，该平均值对应于检测到存在的扩展的MAF单元的百分比。一旦针对测试数据库中的所有基因组计算了聚类值，则步骤330继续使用最大化模型的敏感性、特异性和简约性的模型选择策略(即最小化实际用于预测敏感性的聚类的数目)。为此，对于每个模型$G_{\lambda_i^c}(S_p)$，阈值S_p是变化的，并且对于阈值S_p的每个值，根据以下关系计算敏感性和特异性：

[0153] $$\text{敏感性} = \frac{TP}{TP + FN} \tag{7}$$

[0154] $$\text{特异性} = \frac{TN}{FP + TN} \tag{8}$$

[0155] 其中TP、TN、FP和FN分别是表1中描述的真阳性、真阴性、假阳性和假阴性的数。

[0156] 表1

[0157]

	实际条件 y_i	
	NS	S

[0158]

预测$\widetilde{Su}$			
	NS	TP	FP
	S	FN	TN

[0159] 然后分别在y轴和x轴上在ROC曲线上绘制阈值S_p的敏感性值和(1减去特异性)值，计算并存储ROC曲线下面积，表示为“AUC”。

[0160] 然后将模型$G_{\lambda_i^c}(S_p)$的阈值S_p的最佳值$\widetilde{S_{p,\iota}}$计算为对应于ROC曲线上最靠近横坐标0和纵坐标1的点的值，如图5所示。然后，根据以下关系计算模型$G_{\lambda_i^c}(\widetilde{S_{p,\iota}})$的平衡精度(“bACC”)，存储模型$G_{\lambda_i^c}(\widetilde{S_{p,\iota}})$的bACC、敏感性和特异性：

[0161] $$bACC = \frac{\text{敏感性} + \text{特异性}}{2} \tag{9}$$

[0162] 模型$G_{\lambda_i^c}(\widetilde{S_{p,\iota}})$中最终保留的模型是最简约的模型，使bACC最大化至一个公差内，例如所述模型由此：

[0163] a.A是一组模型$G_{\lambda_i^c}(\widetilde{S_{p,i}})$，由此其bACC大于max(bACC)-0.01，或者max(bACC)是计算的bACC中的最大值；

[0164] b.选择的模型是组A中最简约的模型。

[0165] 将选择的模型存储在计算机存储器中以供后续使用40，包括预测所述细菌物种的细菌菌株的抗生素敏感性。保留的聚类及因此构成unitig由此形成抗生素敏感性的基因组签名。

[0166] 特别地，这种预测包括：

[0167] a.在400中，通过应用完整基因组序列例如以关于图1描述的方式对细菌菌株的基因组进行测序；

[0168] b.在402中，按照前述方式计算存储的预测模型中的每个聚类的值；

[0169] c.在404中，使用存储模型的关系(5)-(6)计算菌株的敏感性$\widehat{Su}$。

[0170] B.实施例

[0171] B.1.肺炎克雷伯菌

[0172] 进行图2A和2B中描述的方法以预测细菌物种肺炎克雷伯菌对不同抗生素的敏感性。表2列出了用于训练和验证预测模型的菌株数目、其NS/S表型和测试的各种抗生素。

[0173] 表3列出了根据本发明的方法训练的模型(称为“聚类-套索模型”或“聚类-套索”)的性能，以及仅基于称为“套索模型”或“套索”的现有技术套索逻辑回归训练的预测模型的性能。根据前面描述的方法，此表中呈现的性能对应于交叉验证的估计。对于后者，计算根据关系(1)和(2)的模型，阈值$\widehat{Su}$和最终模型以先前针对根据本发明的方法描述的方式选择，因此基于相同的性能标准。

[0174] “支持物”列表示对于预测模型保留的预测变量的数目，即针对$\hat{B}_j \neq 0$的聚类或聚类-套索的数目以及针对套索的$\hat{\beta}_j$的“活性MAF单元”的数目。“unitig”列表示对于基因组签名保留的unitig总数，括号中是最广泛的预测变量中的unitig数目。

[0175] 表2

[0176]

	学习数据库		测试数据库	
	NS	S	NS	S
阿米卡星	346	1319	191	160

[0177]

氨曲南	1426	216	250	10
头孢吡肟	961	608	235	53
头孢西丁	976	667	319	138
头孢他啶	1259	136	457	125
环丙沙星	1461	201	471	137
亚胺培南	504	1160	259	301
美罗培南	524	1134	297	86
哌他唑（pipertazo）	1228	432	382	146
四环素	928	737	273	155

[0178] 表3

[0179]

	套索				聚类-套索			
	bACC	AUC	支持物	unitig	bACC	AUC	支持物	unitig
阿米卡星	92.7	95.4	16	22(4)	92.3	95.7	11	93(36)
氨曲南	76.7	81.9	31	45(3)	76.9	82.3	28	425(125)
头孢吡肟	74.0	80.4	53	65(3)	73.6	79.8	34	385(111)
头孢西丁	82.4	88.7	134	155(5)	82.2	88.8	171	1052(221)
头孢他啶	91.6	95.8	51	69(5)	90.7	95.3	43	863(185)
环丙沙星	95.6	98.6	25	27(2)	95.5	98.6	35	422(139)
亚胺培南	93.1	93.6	10	10(1)	92.7	93.4	7	241(194)
美罗培南	91.7	94.0	8	8(1)	91.4	93.5	3	164(159)
哌他唑	81.6	89.6	127	144(4)	81.5	89.0	120	1220(226)
四环素	83.0	88.5	181	198(3)	82.9	87.7	109	640(104)

[0180] 可以看出,聚类-套索模型的性能类似于学习变量不受聚类约束的预测模型的性能。因此值得注意的是,这两个模型在平衡准确度bACC和AUC方面表现出相似的性能,证实考虑或不考虑性状之间的相关性对预测性能的影响有限。还值得注意的是,所述聚类-套索的模型支持物通常略小(对于10种抗生素中的8种),表明用套索单独选择的几个性状最终通过聚类-套索合并成单一聚类。如预期,聚类-套索模型中涉及的unitig总数要大得多。值得注意的是,这个数字在预测聚类中分布不均。例如,在预测对美罗培南敏感性的模型中,164个unitig中的159个存在于单一聚类中,表明基因作为预测基因组性状而存在。

[0181] 图6(A)显示了针对美罗培南的模型的系数的量级。可以看出,聚类-套索模型的签名基本上由单一的重要性状概括,而套索签名的4到5个签名具有可观的权重。事实证明,具有最大预测权重的聚类也是unitig数最大的聚类。如图6(C)所示,这个聚类在压缩的De Bruijn图cDBG中的可视化(例如,使用DBGWAS软件,如前所述)示出这个聚类的unitig在图中形成了一条长线性路径。这因此表明该聚类对应于整个基因。由DBGWAS软件提供的这个线性通路的注释表明其对应于blaKPC基因,该基因在文献中也充分记录了其在美罗培南耐药性中的作用。相反,针对套索签名获得的可视化结果表明,八个预测变量中的三个-变量1、2和4-也共同位于一个被注释为blaKPC基因的区域中。然而,套索在blaKPC基因内选择了这些特异性unitig的事实表明,所涉及的耐药性决定簇是该基因中的点突变,即SNP或插入缺失。虽然基因的注释与使用聚类-套索获得的注释相同,但根据遗传变异对签名的解释是完全不同的。对套索签名的更深入研究表明,位于blaKPC基因中的三个变量实际上是高度相关的。根据本发明,通过明确检测这些实体是相关的并将其与甚至不涉及套索签名的其他相关基因组单元合并成一个聚类,聚类-套索因此导致在两方面对基础预测模型更具生物学意义的解释。首先,关于所涉及的基因组决定簇的性质:基因内的获得或突变。其次,就其对预测敏感性的总体贡献而言,通过将套索签名中涉及的几个不同但相关的性状的贡献相加。

[0182] 相似地,图7示例说明了头孢西丁的两种预测模型的可解释性。着眼于两个最具预测性的聚类所在的cDGB图的子图,这两个区域的注释鉴定了这两种方法的相同耐药性基因(首先,已知参与外排泵的OmpK36基因,其次是blaKPC基因)。另一方面,基因组决定簇的性质(基因存在、SNP、插入缺失等)不能从套索签名中推断出来。

[0183] 可解释性甚至可以非常详细。例如,关于针对聚类-套索签名获得的OmpK36注释子图(图7的右上图),其包含两个聚类(聚类1和3),聚类了9个unitig。这些unitig示出可归因

于局部多态性的拓扑结构,即复杂气泡,用叉子分离敏感菌株和耐药性菌株,如Jaillard M.et al.的文章“A fast and agnostic method for genome-wide association studies:Bridging the gap between k-mers and genetic events”,PLOS Genetics,2018所述。相比之下,针对套索获得的相应子图(图7的左上图)包括四个具有四个不同$\widehat{\beta}_j$值的单元(单元1、2、32和56)。不同$\widehat{\beta}_j$值可导致关于相应unitig序列的各个重要性的错误结论。实际上,当考虑掺入OmpK36的附加注释序列的多重比对时,似乎活性MAF单元2和56代表野生型,单元1和32比对L3环中相同的两个氨基酸插入,如Novais A.et al.的文章“Spread of an OmpK36-modified ST15 Klebsiella pneumoniae variant during an outbreak involving multiple carbapenem-resistant Enterobacteriaceae species and clones”,European Journal of Clinical Microbiology and Infectious Diseases,2012所述。本发明替代地提供了每个单倍型的平均B_j值。

[0184] 对于套索签名获得的第二个子图(图7的左下图)仅包含一个签名性状(以黑色显示)和七个周围节点(以灰色显示),其中两个是注释的blaKPC。由于单个签名节点本身没有注释,因此子图可以解释为blaKPC基因启动子区域中的局部多态性。然而,聚类-套索子图(图7的右下图)表明这个单一unitig是通过套索从数百个高度相关的unitig中选择的:它们都属于聚类2,其包括完整blaKPC基因(显示在括号中)以及其插入的与基因序列高度共现的质粒序列。

[0185] 因此,由聚类-套索提供的附加信息可以得出结论,头孢西丁耐药性的第一个因果变量是OmpK36基因中的局部突变。有利地,用于预测头孢西丁耐药性的分子技术(PCR、NGS等)将特异性靶向这种突变。此外,第二个因果变量是完整blaKPC基因的获得,任何特异于blaKPC的DNA序列都可以有利地被此类技术用于预测头孢西丁耐药性。

[0186] 测试了其他细菌物种/抗生素对。在不详细介绍与肺炎克雷伯菌有关的情况下,以下是针对沙门氏菌、金黄色葡萄球菌和淋病奈瑟菌的描述:

[0187] -分别与上述表2和表3相似的第一表和第二表;

[0188] -第一个、第二个和第三个图,分别示例说明了对于考虑的抗生素,套索模型系数的绝对值、聚类-套索模型系数的绝对值以及聚类-套索模型的前10个最具预测性的聚类中包含的unitig的数目;

[0189] -图,示例说明:

[0190] o在其左侧,压缩的cDBG图的子图,涉及套索模型的最具预测性的扩展的MAF单元。子图首先由最具预测性的单元鉴定并且当子图中存在其他单元时,其也被表示;

[0191] o在其右侧,压缩的cDBG图的子图,涉及聚类-套索模型中最具预测性的聚类。子图首先由最具预测性的聚类鉴定,当子图中存在其他聚类时,其也被表示。

[0192] B.2.沙门氏菌

[0193] 表4和5,四环素见图8和9,庆大霉素见图10和11。

[0194] 与套索模型不同,其鉴定了TetB基因中可能的一组点突变以获得四环素耐药性,鉴于这将暗示大量突变,否则这不是结论性的,基于聚类-套索的本发明鉴定了存在TetA基因(聚类1)、TetB/TetD基因(聚类2)的耐药性获得,以及TetR基因的获得。

[0195] 关于庆大霉素耐药性,本发明的结论是获得AAC3基因(聚类1)并且OXA、IMP和TEM

基因参与耐药性机制，而套索模型未能鉴定OXA和IMP基因。

[0196] 表4

[0197]

	学习数据库	
	NS	S
四环素	2597	1901
庆大霉素	637	3862

[0198] 表5

[0199]

	套索			聚类-套索		
	bACC	AUC	支持物	bACC	AUC	支持物
四环素	97.4	98.2	29	97.1	98.1	15
庆大霉素	96.8	98.2	48	96.5	98.2	34

[0200] B.3.淋病奈瑟氏菌

[0201] 表6和7，图12和13。

[0202] 关于淋病奈瑟氏菌的头孢克肟耐药性，本发明鉴定了penM基因中几种重组的获得。

[0203] 表6

[0204]

	学习数据库	
	NS	S
头孢克肟	110	554

[0205] 表7

[0206]

	套索			聚类-套索		
	bACC	AUC	支持物	bACC	AUC	支持物
头孢克肟	91.7	97.2	45	92.1	97.5	40

[0207] 金黄色葡萄球菌

[0208] 表8和9，图14和15。

[0209] 关于金黄色葡萄球菌的四环素耐药性，本发明鉴定了TetK基因(聚类1)的获得，但排除了作为庆大霉素耐药性的高度预测性基因组决定簇的TetM基因(聚类2和3)的获得，这是鉴于所涉及的聚类与套索模型不同的低系数，后者将TetM基因解释为高度预测性的。

[0210] 表8

[0211]

	学习数据库	
	NS	S
四环素	27	468

[0212] 表9

[0213]

	套索			聚类-套索		
	bACC	AUC	支持物	bACC	AUC	支持物
四环素	96.9	96.7	12	98.9	99.6	10

[0214] C.执行本发明的计算机方式

[0215] 步骤302、304、312、320，就像下面描述的步骤60和80，由计算机执行，例如包括一

个或多个处理器、存储空间和随机存取存储器的计算机单元，其能够存储计算机指令，当执行所述指令时执行前述的计算。所述计算单元例如是个人计算机、服务器或计算聚类。类似地，步骤402、404由计算机执行，例如如前所述的计算机单元。步骤302、304、312、320的单元与步骤402、404的单元是不同或相同的单元。有利地，预测的敏感性显示在计算机屏幕上，存储在实验室或医院计算机系统中，以在细菌菌株感染患者时补充患者的记录，或传输到临床医生的移动设备例如智能手机。

[0216] D.本发明实施方案的教导的扩展

[0217] D.1.关于检测细菌基因组中k-mer、unitig和MAF单元的存在-步骤330

[0218] 步骤330描述了一种检测细菌基因组中某基因组序列、特别是unitig或一组基因组序列、特别是一组单元聚类unitig的存在或不存在的方式。一般而言，所述实施方案解决了某序列或序列组是否应在基因组中相同检测到的问题或者是否可以接受所述序列或序列组与基因组中序列或序列组之间的一定水平差异以决定其存在或不存在的问题。特别地，正如序言中所解释的，完美同源性假设学习数据是完整的，以涵盖生物物种的所有变异性，这实际上是困难的，特别是考虑到其基因组的可塑性。

[0219] 此外，测序的基因组可能会被错误污染，特别是当它是“读取”形式时，即在任何生物信息学处理如共有组装或过滤低质量读取之前在测序平台的输出端产生的序列。在这种情况下，一个序列虽然存在于基因组中，但可能由于测序错误而被检测为不存在，反之亦然。特别地，生物信息学处理通常包括过滤低质量读取，以及任选对过滤后的读取进行共有组装以获得组装的序列或“重叠群”。任选的组装性质通常取决于执行样品分析的背景。组装的效果是将重叠群中的测序错误显著减少至目前Illumina Inc.公司的平台中使用的SBS技术的10^{-5}水平以及Oxford Nanopore Technologies Ltd.公司的平台中使用的纳米孔技术的10^{-2}水平。另一方面，由于组装需要高计算能力和时间，因此它不太适合计算环境通常不是很有力和/或快速的“POC”(“即时检验”)基因组应用或者甚至实时应用。在这种情况下，直接对过滤或未过滤的读取进行基因组分析，如前所述例如确定样品中存在的一种或多种物种的身份和/或其对一种或多种抗生素的敏感性。然而，SBS技术的测序错误约为2％至3％的级别，纳米孔技术的高达12％。如果没有特别的防范措施，基因组分析会导致同样高的错误率。

[0220] 图16示例说明了用于更稳健地检测微生物特别是细菌菌株、酵母菌株或霉菌菌株基因组中基因组序列的存在或不存在的方法50，以解释基因组变异性和测序错误。这个方法虽然本身独立于图2A和2B的方法，但在其步骤330和/或步骤402中有利地执行。

[0221] 所述方法50包括对包含一种或多种微生物菌株的样品进行测序的步骤70，在下文中并且仅作为示例的一种或多种细菌菌株，并预处理由测序平台产生的读取，以及包括步骤80，其在所述细菌菌株之一的基因组中检测预定基因组序列。这个步骤80任选包括在所述基因组中检测至少一组预定基因组序列。

[0222] 步骤80使用基因组序列的分解以及一定数目的参数，这些参数在步骤60中计算，例如在所述方法50之前执行，并存储在数据库DB中。更具体地，参考图17，步骤60开始于步骤600，其中标记为“SEQ”的基因组序列通过以恒定步长、优选步长为1在序列SEQ上滑动长度为k的窗口W被分解成为恒定长度k的k-mer，其中k在15和50之间，例如k＝31。在窗口W的每个位置，因此存储了k-mer。对于长度为L的序列SEQ，因此产生了(L-k+1)个k-mer。在随后

的可选步骤602中，产生的k-mer集合被过滤掉其可能的重复，以仅保留由唯一k-mer组成的集合，称为KM＝{km_1,…,km_i,…,km_S}，该集合形成存储在DB中的序列SEQ的分解。

[0223] 步骤70涉及如本身已知的并且例如关于图1所描述的：

[0224] -在步骤700中，准备样品以对样品中包含的DNA进行测序并对制备的DNA进行测序，从而产生和存储读取；

[0225] -在步骤702中执行生物信息学处理，其一般包括过滤掉低质量读取，及任选通过共有组装过滤的读取以获得和存储组装的序列或“重叠群”。

[0226] 检测基因组序列SEQ的步骤80从800中的第一个测试开始，其在于了解是否对重叠群或读取执行该检测。如果检测是在重叠群上执行的，即其错误率足够低以允许通过检测k-mer使用完美同源性的基因组序列，则该方法在802中通过检测重叠群中KM集合的每个k-mer km_i的存在或不存在而继续。特别地，如果其相同地存在于至少一个重叠群中，则检测到k-mer km_i。

[0227] 在随后的步骤804中，进行测试以确定序列SEQ是否应该在细菌基因组中被相同地检测到。如果是，如果KM集合中的所有k-mer km_i都被检测为存在于重叠群中，则在806中检测到序列SEQ。在这方面要注意，由于测序技术和/或组装技术，序列SEQ不一定在一个重叠群中完全结束，而是可以分开在几个重叠群之间，这种分开的概率随着序列SEQ的长度L而增加。因此，分解成k-mer使得可以在基因组中鉴定后者，即使它在重叠群中未如此存在。

[0228] 如果在804中以相同方式没有寻找到序列SEQ，则在808中至少针对序列SEQ或其变体之一搜索基因组。例如，这些变体对应于原始序列SEQ中的突变或后者的不完美鉴定。如上所述，序列SEQ可以是基于先验数据或知识鉴定遗传决定簇(例如耐药性、毒力、身份等)的产物。如果后者不完美，则序列SEQ可能无法反映所述决定簇的全部多样性。通过可以鉴别序列SEQ或其变体之一，所述方法允许纠正数据和知识的初始缺陷，从而检测遗传决定簇。在较小程度上，它还允许考虑重叠群中可能的残余测序错误。此外，与纠错无关，还可以从单个序列SEQ检测一组变体中的至少一个成员是否存在于基因组中，而不必完美地检测每个所述变体。

[0229] 在第一变体中，在808中，如果其构成k-mer的百分比大于预定阈值s_{uni}，例如大于70％，则检测到序列SEQ或其变体之一。在优选的变体中，这个百分比取决于序列SEQ的长度L，并且更特别地作为L的函数而降低。具体地，优选保持足够长的k-mer长度，特别是大于15，优选大于30，以使后者对序列SEQ保持特异性。因此，随着长度L减小，在越来越大百分比的k-mer中发现与序列SEQ的差异，这使得可以校正随长度L降低的百分比s_{uni}。如图18所示，百分比s_{uni}例如是逐步递减的，且包含三个值。有利地，对于在15和50之间的k，例如k＝31，如果L≤61，则s_{uni}＝90％，如果61<L≤100，则s_{uni}＝80％，如果100<L，则s_{uni}＝70％。

[0230] 一旦检测到序列SEQ，任选在810继续所述方法，以下述方式检测标示为{SEQ_1,…,SEQ_i,…SEQ_E}的一组基因组序列。

[0231] 如果对读取进行序列SEQ的检测(测试800)，及因此在纠正测序错误的任何生物信息学处理之前，所述方法考虑这些错误以准确检测读取及因此基因组中k-mer的存在/不存在。直接从读取中检测的优势在于数据处理的速度，对于给定的计算环境，处理速度不到2-3分钟，而对于相同的环境，仅组装可能就需要一小时。

[0232] 在第一个变体中，如果读取包含最小数目的所述k-mer的拷贝，例如大于或等于3

的数目，则检测到k-mer。然而，这个变体具有未考虑测序覆盖深度的缺点。大致上，测序错误分布在整个基因组中，因此测序覆盖深度越大，检测到k-mer的概率就越高。然而，由于测序错误，很难确定k-mer是否真的存在于基因组中或者在读取中检测到的k-mer是否是带有测序错误的另一k-mer的产物。在优选的变体中，检测取决于在其中寻找序列SEQ的细菌菌株的实际测序覆盖深度。

[0233] 为此，执行测试812以确定样品是宏基因组样品，或更一般地是含有几种不同物种(几种细菌物种、人DNA或其他)的样品，还是从细菌菌株的分离株制备的样品。在后一种情况下，由于仅存在一种菌株，例如根据以下关系在814中计算属于所述菌株的所有读取及测序覆盖深度，记为“cov”：

[0234] $$cov = \frac{N_r}{N_g} \quad (10)$$

[0235] 在所述关系中，N_r是包括在读取中的碱基总数，N_g是细菌菌株所属的细菌物种的参考基因组中的碱基数，优先具有所述物种的平均大小或接近观察到的基因组大小的平均值(例如对于结核分枝杆菌，Ng＝4.4百万个碱基对(Mbp))。

[0236] 然后根据以下关系在816中计算以N_{cov}标示的拷贝数，其需要在读取中检测以确认基因组中确实存在k-mer：

[0237] $N_{cov}＝\tau\times cov$ (11)

[0238] 其中τ是预定参数，优选考虑所使用的测序技术的测序错误率，其有利地在5％至15％之间，优选地大于或等于10％，例如10％。对于后一个百分比和深度100，因此必须检测10个相同的k-mer拷贝以确定它实际上存在于读取中。10％的比率显著使k-mer的存在能够在Oxford Nanopore Technology Ltd公司的GridION平台使用同一公司的R9.4文库制备试剂盒产生的读取中被准确地检测到。这个比率还允许在通过SBS型测序技术产生的读取中进行精确检测，例如Illumina Inc.公司的MiSeq平台。

[0239] 然后在818中检测读取中KM集合的每个k-merkm$_i$的存在或不存在。特别地，如果读取中至少存在N_{cov}相同的拷贝，则检测到k-merkm$_i$。然后继续所述方法至之前描述的步骤804。

[0240] 如果样品包含若干物种(测试812)，则所述方法包括确定所考虑的细菌物种的测序覆盖深度。更具体地，在820中执行“分类分箱(taxonomic binning)”，这种分箱包括为每个读取分配样品中存在的物种中的起源。这种类型的分箱在现有技术中是熟知的并使用例如Wood D.E.et al.的文章“Kraken:ultrafast metagenome sequence classification using exact alignments”,Genome Biology,2014中描述的分类，或者通过可下载自网站https://github.com/DerrickWood/kraken2/releases的软件“Kraken2”执行。

[0241] 然后计算所考虑的所述细菌物种的测序覆盖深度，例如根据以下关系之一计算：

[0242] $$cov = \frac{N_r^m}{N_g} \quad (12)$$

[0243] $$cov = \rho \times \frac{N_r}{N_g} \quad (13)$$

[0244] 其中，关系N_r^m是分配给所考虑的细菌菌株所属的细菌物种的读取中包含的碱基

总数，N_g是细菌物种的中等大小基因组的碱基数，并且ρ是样品中细菌物种的相对比例。这个相对比例例如使用Wood D.E.et al.的文章“Kraken:ultrafast metagenome sequence classification using exact alignments”,Genome Biology,2014中描述的分类法计算，或者通过可下载自网站https://github.com/DerrickWood/kraken2/releases的软件“Kraken2”执行。所述方法继续进行以刚刚计算的测序覆盖深度为函数计算拷贝数的步骤816。

[0245] 再次提及检测以$\{SEQ_1,\cdots,SEQ_i,\cdots SEQ_E\}$表示的基因组序列集合的步骤810，这个步骤在以前述方式检测每个序列SEQ_i之后。更具体地，所述集合的检测根据以下选项之一实施：

[0246] i.如果确定为存在的SEQ_i的百分比大于第二预定检测阈值s_{clus}，例如大于或等于20％例如25％的阈值，则该集合被检测为存在于基因组中，否则是不存在；或者

[0247] ii.如果所有的SEQ_i都被确定为存在，则该集合被检测为存在于基因组中，否则是不存在；或者

[0248] iii.当至少一个序列SEQ_i被确定为存在时，则该集合被检测为存在于基因组中，否则是不存在；或者

[0249] iv.该集合被检测为存在于基因组中的概率等于确定为存在的SEQ_i的百分比。

[0250] 首先，注意到序列SEQ或SEQ集合的检测性能与通过k-mer识别直接获得的检测性能非常相似，如与前述基于套索的方法的性能相比的基于聚类-套索的方法的性能所证实。

[0251] 其次，检测方法50对于所采用的测序技术的类型是鲁棒性的，尤其是对于其测序错误。下表示例说明，对于分离形式的37个肺炎克雷伯菌测试菌株，使用针对unitig(等于SEQ_i)的变体检测(百分比70％、80％、90％)和根据上述选项i)的单元检测，通过MiSeq(“Illumina”)和GridION(“ONT”)测序对不同抗生素的耐药性的聚类-套索预测结果。这两种测序技术根据其读取与由其读取的组装产生的重叠群的函数进行测试。表10示出两种技术的结果相似，而其测序错误率显著不同，但读取和重叠群的结果也相似。

[0252] 表10

[0253]

	数据类型	bACC	敏感性	特异性	AUC
哌拉西林	Illumina 读取	86.5	81.5	100	91.1
	Illumina 重叠群	86.5	81.5	100	92.2
	ONT 读取	89.2	85.2	100	91.9
	ONT 重叠群	86.5	81.5	100	91.1
头孢吡肟	Illumina 读取	94.6	92.3	100	98.3
	Illumina 重叠群	97.3	96.2	100	98.6
	ONT 读取	94.6	92.3	100	96.5
	ONT 重叠群	97.3	96.2	100	98.3
美罗培南	Illumina 读取	86.5	81.0	93.3	92.7
	Illumina 重叠群	86.5	81.0	93.8	92.7
	ONT 读取	78.4	61.9	100	86.9
	ONT 重叠群	86.5	81.0	93.8	92.6
环丙沙星	Illumina 读取	94.6	96.0	91.7	95.7
	Illumina 重叠群	94.6	96.0	91.7	95.7
	ONT 读取	91.9	88.0	100	96.7
	ONT 重叠群	94.6	96.0	91.7	95.7

[0254] 此外，作为ONT技术的测序覆盖深度函数的预测性能(AUC)显示在图19A(套索预测)和19B(聚类套索预测)，样品从肺炎克雷伯菌菌株的分离株产生。值得注意的是，考虑到测序覆盖深度，及因此随着后者的拷贝数增加，可以从30的深度快速获得稳定性能。图20和21通过模拟研究表明，包含金黄色葡萄球菌菌株的宏基因组样品(此处模拟来自支气管肺泡灌洗液的临床样品)对作为Illumina技术的读取(图20)或重叠群(图21)的函数的聚类-套索预测的性能的影响。还注意到在高性能下非常快速的稳定取决于样品中存在的金黄色葡萄球菌菌株的实际测序覆盖深度。

[0255] D.2.关于实施方案的其它特征

[0256] 已经描述了本发明的特定实施方案。这个方法可以根据以下特征单独或组合进行修改：

[0257] -已经描述了对抗生素敏感性的预测。本发明适用于任何类型的表型，例如细菌菌株的毒力、其核糖型等；

[0258] -已经描述了本发明对取自患者的生物学样品的应用。本发明适用于任何类型的包含细菌的样品，特别是取自动物的样品或取自环境的样品；

[0259] -已经描述了细菌。本发明也适用于酵母和霉菌；

[0260] -已经描述了细菌基因组的完整测序。作为变体，所述基因组测序是部分测序并且靶向一个或多个已知涉及抗生素敏感性的特定区域；

[0261] -在所描述的实施方案中，基因组中k-mer和unitig的值是二元的(不存在或存在)，例如在矩阵X中编码。作为变体，k-mer或unitig的值等于其在基因组中的拷贝数；

[0262] -已经描述了二进制预测(S和NS状态)。作为变体，敏感性是有序的(更高状态数，例如S、R和I)或线性的(例如预测最小抑制浓度或“MIC”)。在这种情况下，回归是有序的或线性的；

[0263] -已经描述了通过LASSO类型的逻辑回归训练的预测模型。其它简约算法也是可能的，例如随机森林模型、梯度提升、集合覆盖机、聚集和蒙特卡洛或深度学习，或任何类型的

惩罚套索学习(弹性网络、组套索、融合套索、自适应套索等)；

[0264] -已经描述了使用逻辑回归套索选择MAF单元。其它选择也是可能的，例如在Friedman J.H.的文章“Greedy Function Approximation:A Gradient Boosting Machine”,The Annals of Statistics,2001中描述的以及例如通过得自website https://xgboost.readthedocs.io/en/latest/的软件“xGBoost”使用的选择，或者任何其它非线性选择；

[0265] -已经描述了基于Bravais-Pearson相关值的聚类。其它类型的共现测量也是可能的，例如Jaccard或Sørensen-Dice距离；

[0266] -已经描述了一个特定聚类。其它类型的聚类也是可能的，例如“标准”层次聚类；

[0267] -聚类值被描述为等于构成其的单元的平均值。其它值是可能的。例如，对每个聚类执行“套索组”类型的逻辑回归，以便为构成其的不同单元分配不同权重；

[0268] -已经描述了从组装基因组中学习算法的用途。作为变体，本发明直接应用于由测序平台产生的基因组，即以读取形式、任选从低质量读取中过滤的读取形式的基因组。

序列表

<110> 生物梅里埃公司

<120> 从基因组预测细菌表型性状的分子技术

<130> Cluster Lasso 2008

<160> 6

<170> PatentIn version 3.5

<210> 1
<211> 10
<212> DNA
<213> 人工序列

<220>
<223> 仅用于说明目的

<400> 1
ttcgctcgta 10

<210> 2
<211> 10
<212> DNA
<213> 人工序列

<220>
<223> 仅用于说明目的

[0001]

<400> 2
ttcgatcgta 10

<210> 3
<211> 10
<212> DNA
<213> 人工序列

<220>
<223> 仅用于说明目的

<400> 3
ttcgctcgta 10

<210> 4
<211> 10
<212> DNA
<213> 人工序列

<220>
<223> 仅用于说明目的

<400> 4
ttcgatcgta 10

<210> 5
<211> 10
<212> DNA
<213> 人工序列

<220>

<223> 仅用于说明目的

<400> 5
ttcgatcgta 10

[0002]

<210> 6
<211> 28
<212> DNA
<213> 人工序列

<220>
<223> 仅用于说明目的

<400> 6
gagtatggaa agaattagtt ttccgaaa 28

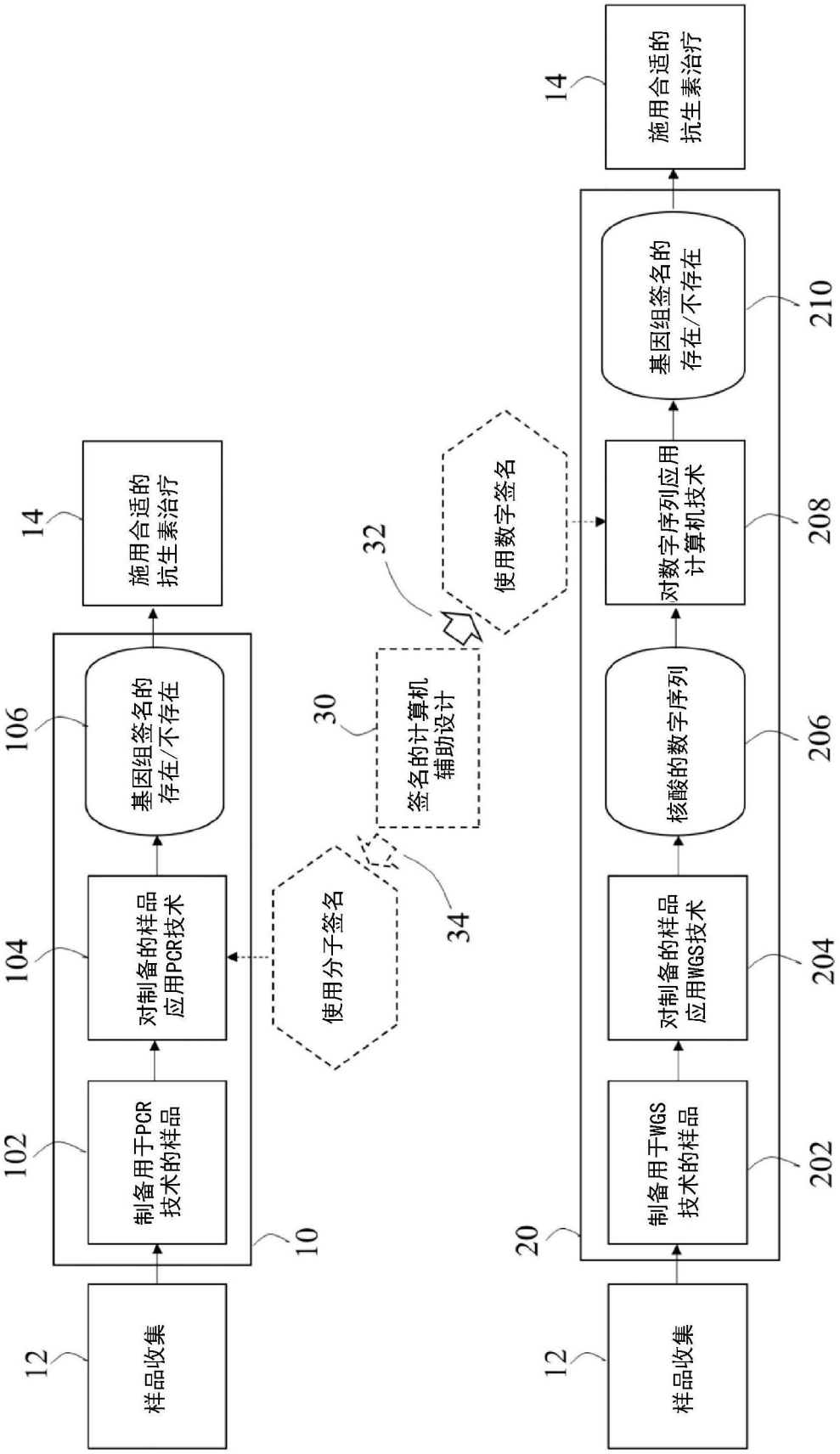

样品收集
制备用于PCR技术的样品
对制备的样品应用PCR技术
基因组签名的存在/不存在
施用合适的抗生素治疗
使用分子签名
签名的计算机辅助设计
使用数字签名
样品收集
制备用于WGS技术的样品
对制备的样品应用WGS技术
核酸的数字序列
对数字序列应用计算机技术
基因组签名的存在/不存在
施用合适的抗生素治疗
12
102
104
106
14
10
30
32
34
12
202
204
206
208
210
14
20

图1

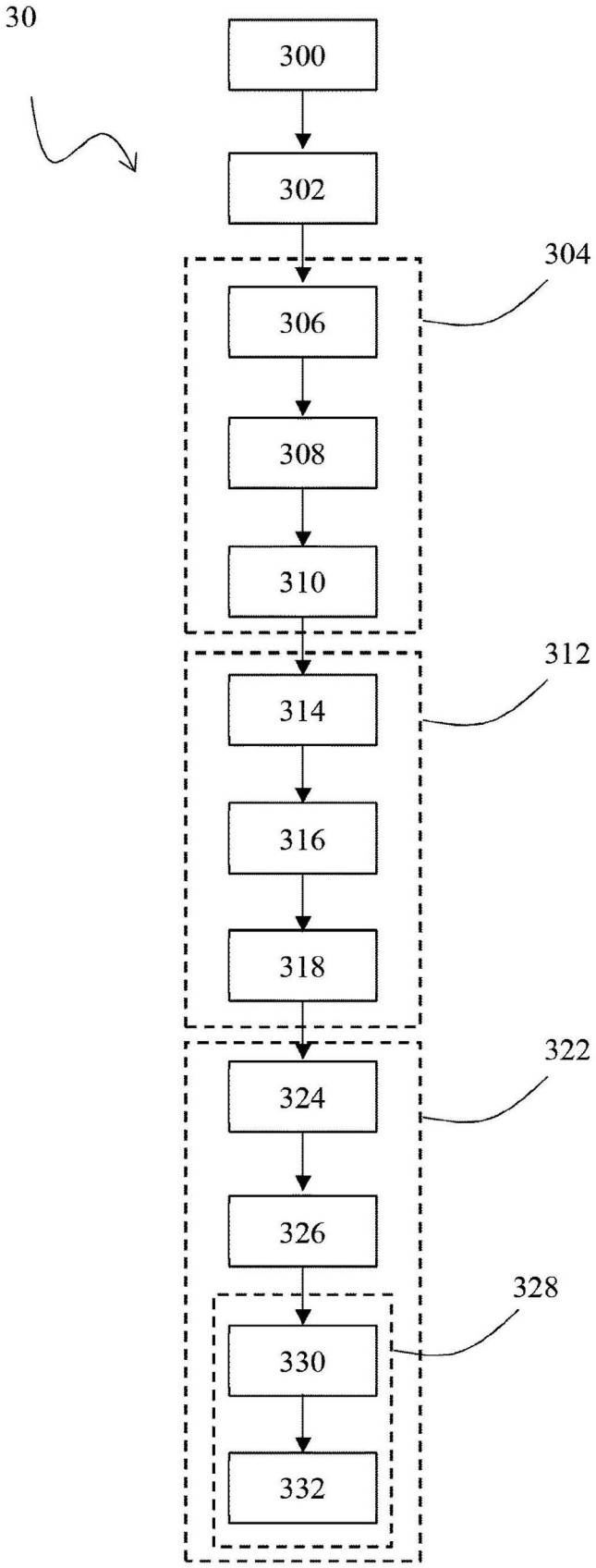

30
300
302
304
306
308
310
312
314
316
318
322
324
326
328
330
332

图2A

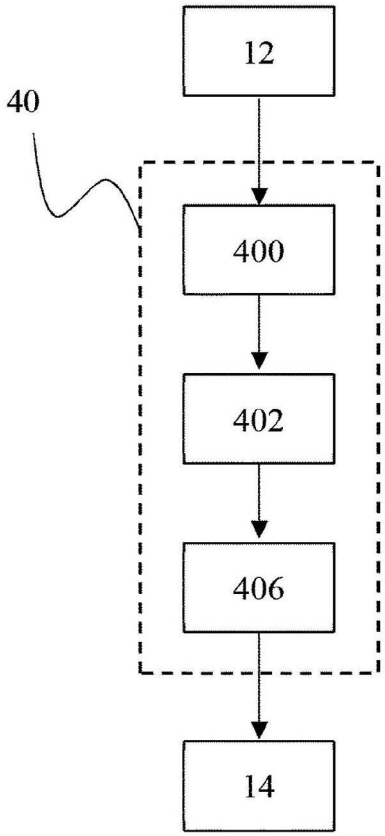


图2B

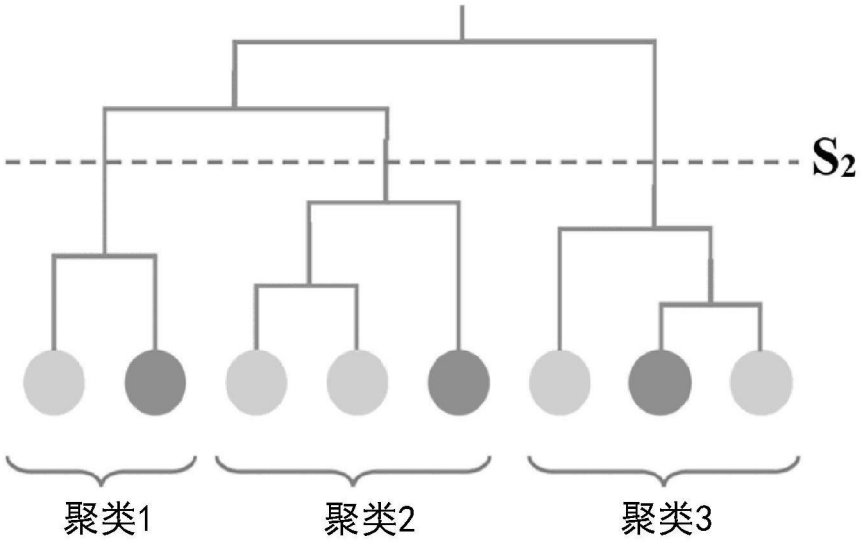


图4

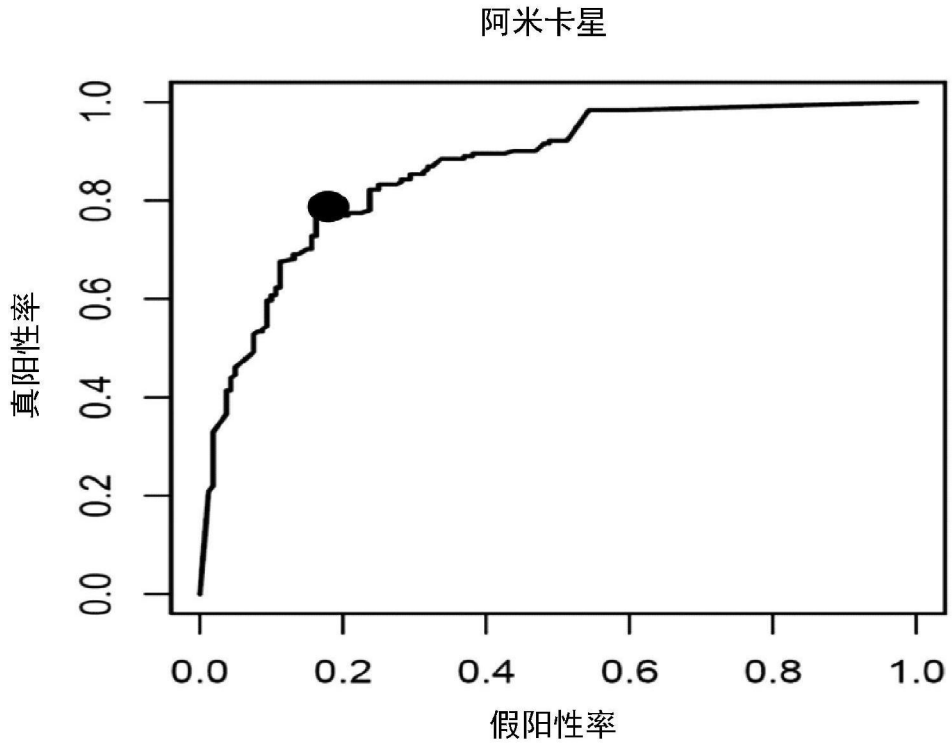


图5

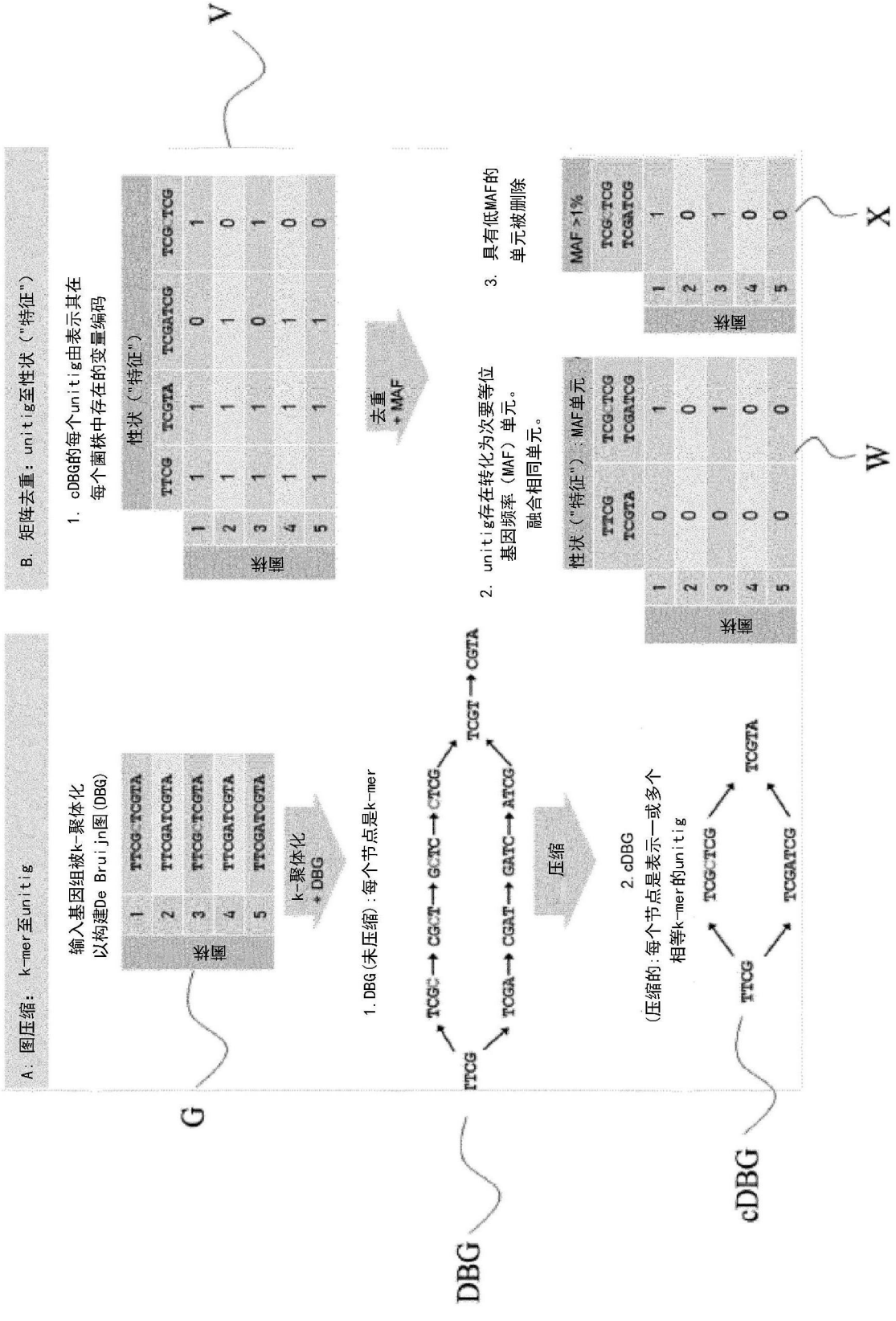

A. 图压缩： k-mer至unitig
B. 矩阵去重：unitig至性状（"特征"）
输入基因组被k-聚体化
以构建De Bruijn图(DBG)
样圃
1 TTCG-TCGTA
2 TTCGATCGTA
3 TTCG-TCGTA
4 TTCGATCGTA
5 TTCGATCGTA
k-聚体化
+DBG
1. DBG(未压缩)：每个节点是k-mer
TTCG
TCGC→CGCT→GCTC→CTCG
TCGA→CGAT→GATC→ATCG
TCGT→CGTA
压缩
2. cDBG
(压缩的:每个节点是表示一或多个
相等k-mer的unitig
TTCG
TCGCTCG
TCGATCG
TCGTA
1. cDBG的每个unitig由表示其在
每个菌株中存在的变量编码
性状（"特征"）
TTCG TCGTA TCGATCG TCG-TCG
1 1 1 0 1
2 1 1 1 0
3 1 1 0 1
4 1 1 1 0
5 1 1 1 0
去重
+MAF
2. unitig存在转化为次要等位
基因频率（MAF）单元。
融合相同单元。
性状（"特征"）：MAF单元
TTCG TCG-TCG
TCGTA TCGATCG
1 0 1
2 0 0
3 0 1
4 0 0
5 0 0
3. 具有低MAF的
单元被删除
MAF >1%
TCG-TCG
TCGATCG
1 1
2 0
3 1
4 0
5 0
G
DBG
cDBG
V
W
X

图3

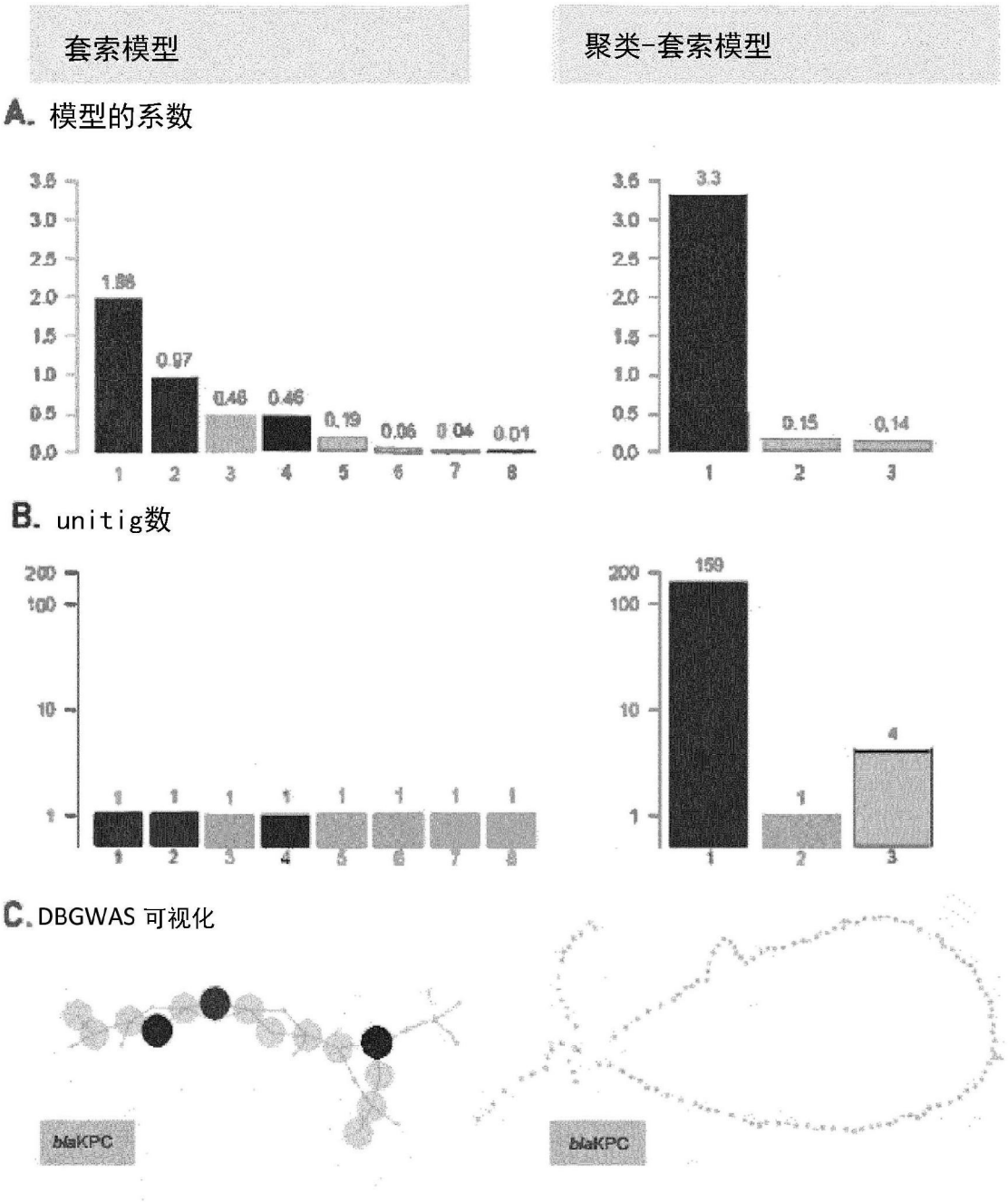

套索模型
聚类-套索模型
A. 模型的系数
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
1.88
0.97
0.48
0.46
0.19
0.06
0.04
0.01
1 2 3 4 5 6 7 8
3.3
0.15
0.14
1 2 3
B. unitig数
200
100
10
1
1 1 1 1 1 1 1 1
159
4
1 2 3
C. DBGWAS 可视化
blaKPC
blaKPC

图6

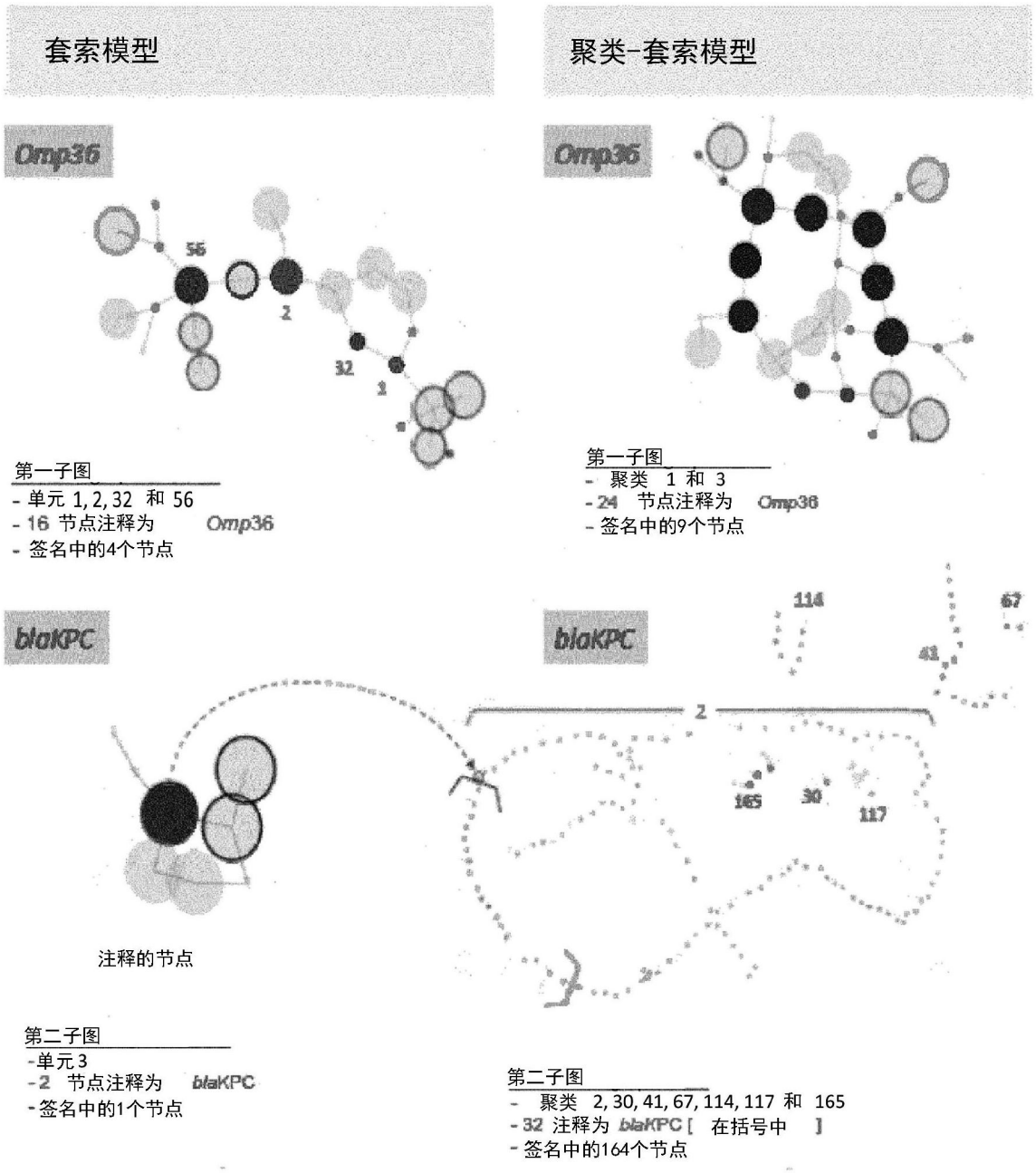
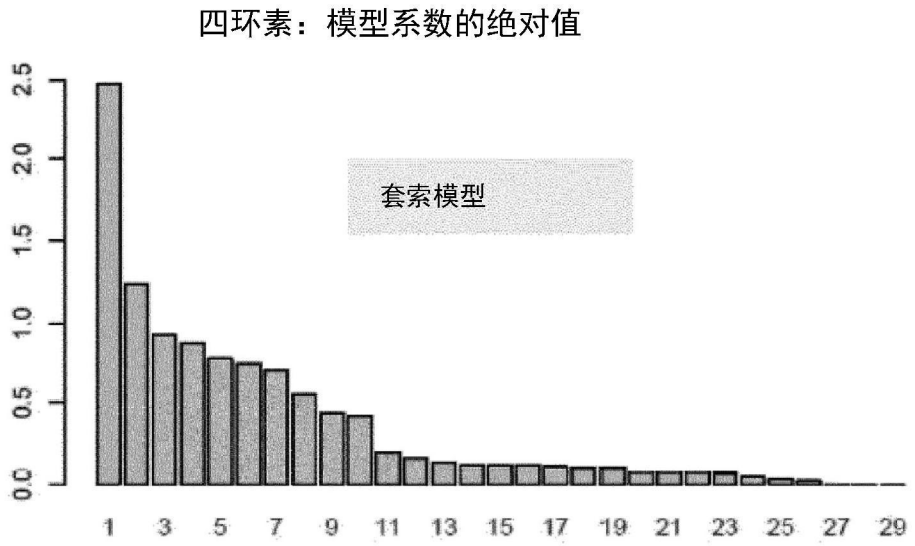
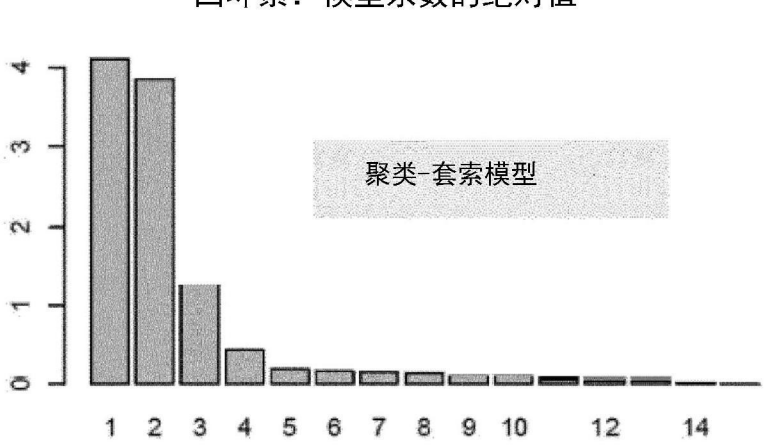

套索模型
聚类-套索模型
Omp36
56
2
32
1
第一子图
- 单元 1, 2, 32 和 56
- 16 节点注释为 Omp36
- 签名中的4个节点
Omp36
第一子图
- 聚类 1 和 3
- 24 节点注释为 Omp36
- 签名中的9个节点
blaKPC
注释的节点
第二子图
-单元3
- 2 节点注释为 blaKPC
- 签名中的1个节点
blaKPC
114
67
41
2
165
30
117
第二子图
- 聚类 2, 30, 41, 67, 114, 117 和 165
- 32 注释为 blaKPC [在括号中]
- 签名中的164个节点

图7


四环素：模型系数的绝对值
套索模型
0.0
0.5
1.0
1.5
2.0
2.5
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29


四环素：模型系数的绝对值
聚类-套索模型
0
1
2
3
4
1 2 3 4 5 6 7 8 9 10 12 14

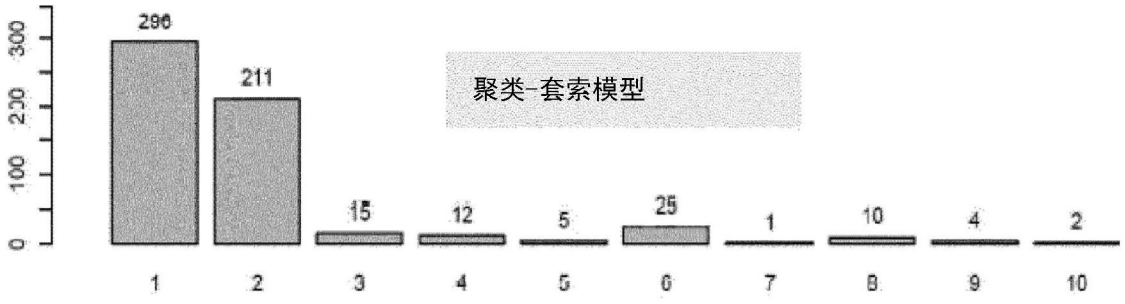

10个最大聚类的unitig数
聚类-套索模型
0
100
200
300
296
211
15
12
5
25
1
10
4
2
1 2 3 4 5 6 7 8 9 10

图8

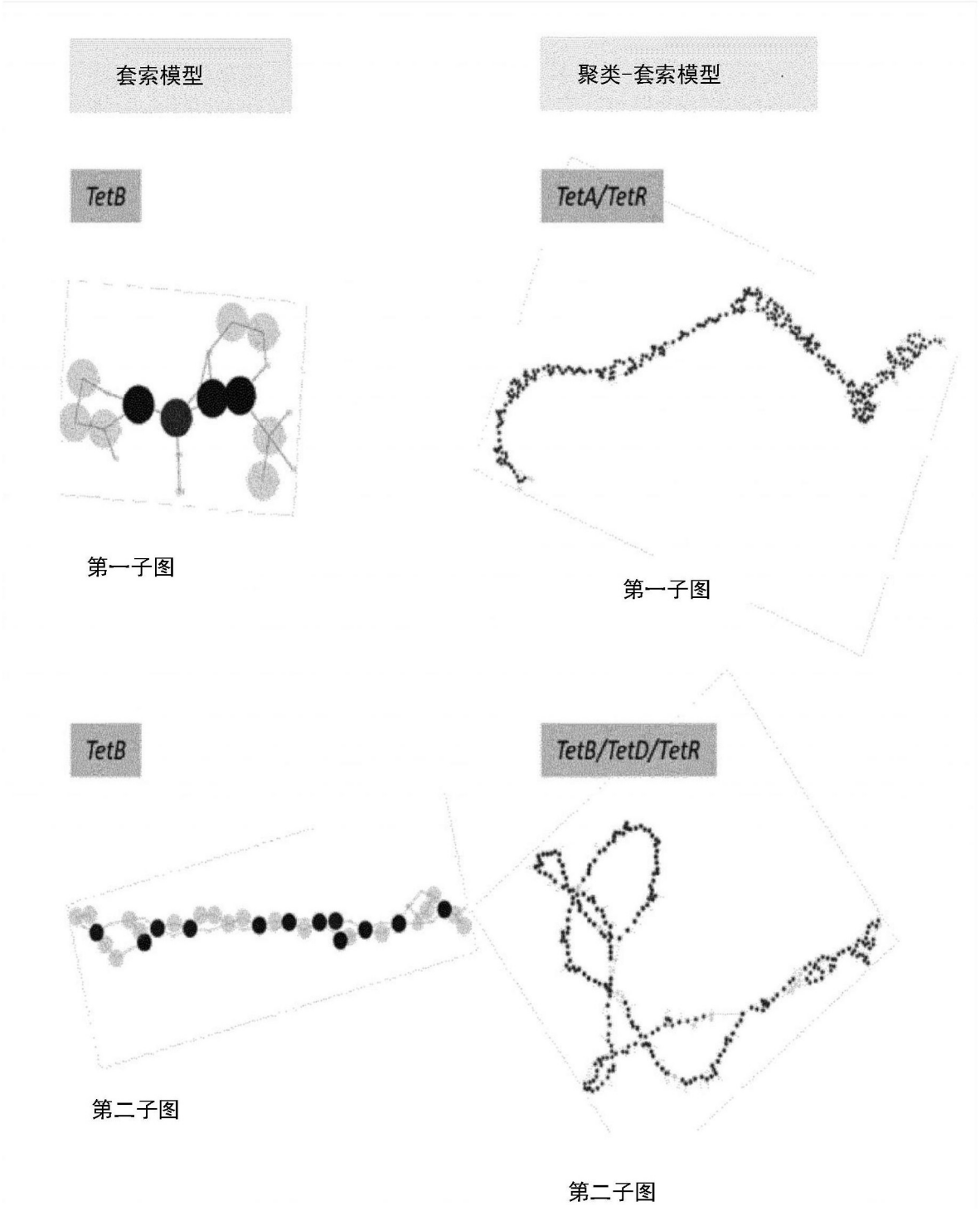
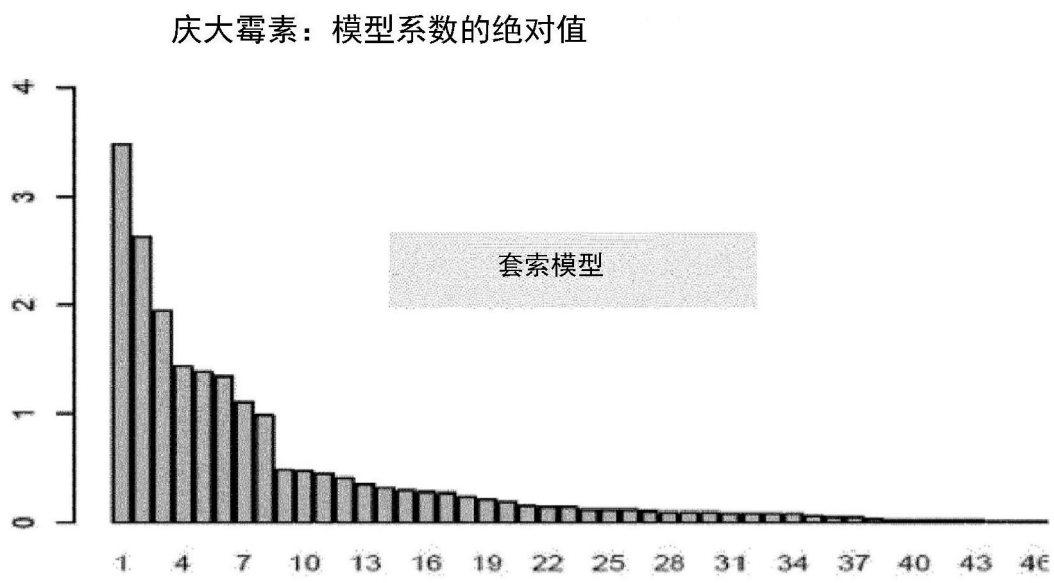
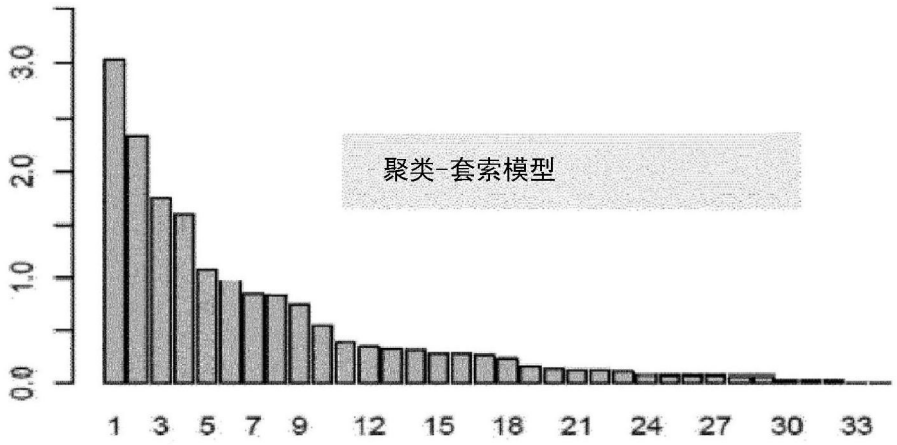

套索模型
聚类-套索模型
TetB
TetA/TetR
第一子图
第一子图
TetB
TetB/TetD/TetR
第二子图
第二子图

图9


庆大霉素：模型系数的绝对值
套索模型
0
1
2
3
4
1
4
7
10
13
16
19
22
25
28
31
34
37
40
43
46


庆大霉素：模型系数的绝对值
聚类-套索模型
0.0
1.0
2.0
3.0
1
3
5
7
9
12
15
18
21
24
27
30
33

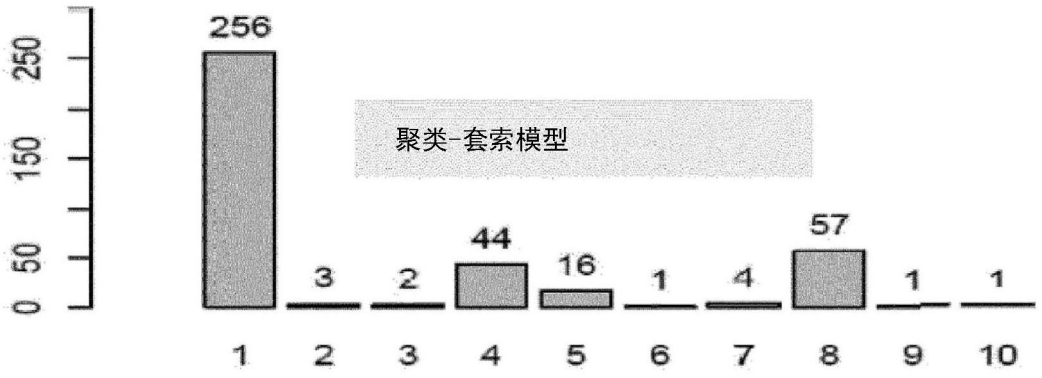

10个最大聚类的unitig数
聚类-套索模型
0
50
150
250
256
3
2
44
16
1
4
57
1
1
1
2
3
4
5
6
7
8
9
10

图10

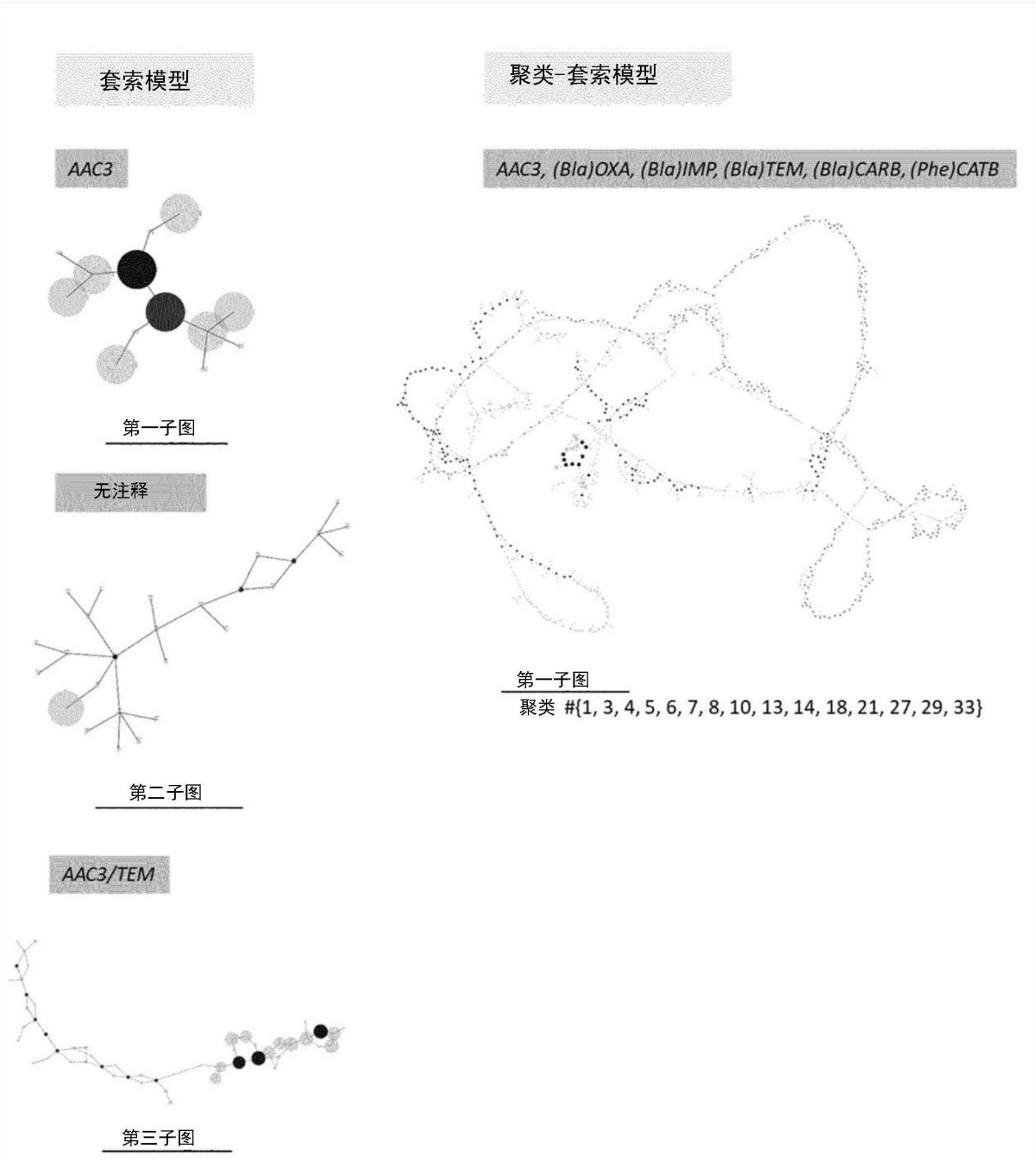
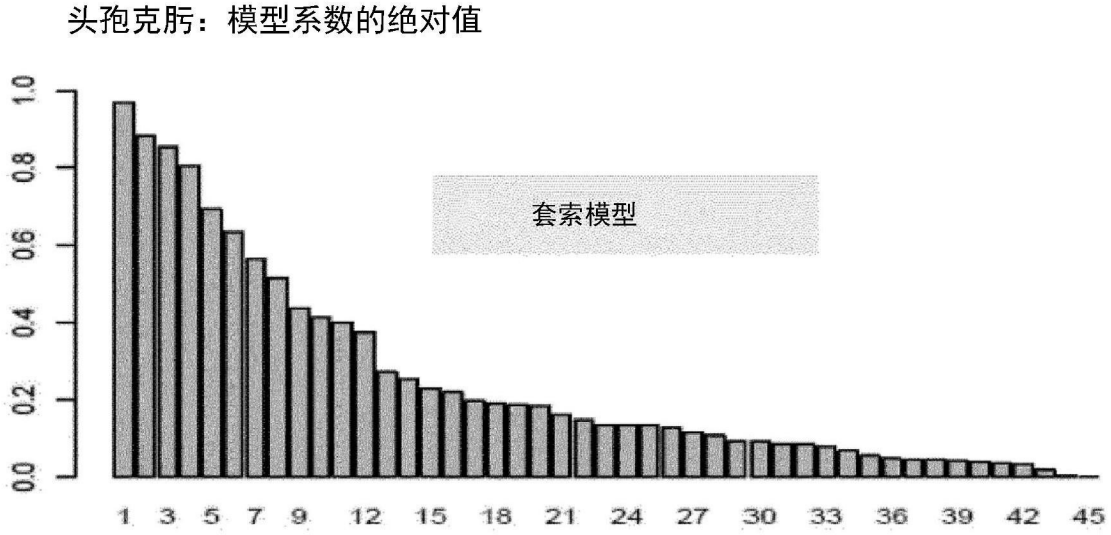
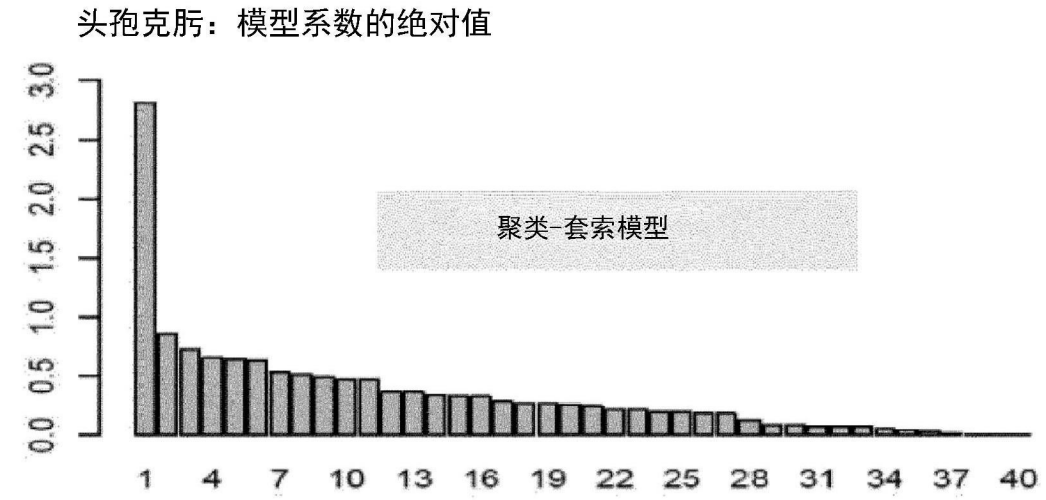

套索模型
AAC3
第一子图
无注释
第二子图
AAC3/TEM
第三子图
聚类-套索模型
AAC3, (Bla)OXA, (Bla)IMP, (Bla)TEM, (Bla)CARB, (Phe)CATB
第一子图
聚类 #{1, 3, 4, 5, 6, 7, 8, 10, 13, 14, 18, 21, 27, 29, 33}

图11


头孢克肟：模型系数的绝对值
套索模型
0.0
0.2
0.4
0.6
0.8
1.0
1 3 5 7 9 12 15 18 21 24 27 30 33 36 39 42 45


头孢克肟：模型系数的绝对值
聚类-套索模型
0.0 0.5 1.0 1.5 2.0 2.5 3.0
1 4 7 10 13 16 19 22 25 28 31 34 37 40

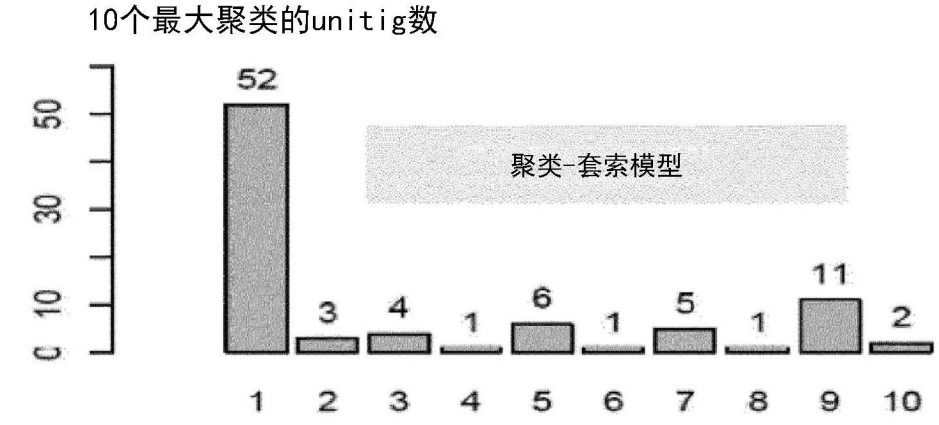

10个最大聚类的unitig数
聚类-套索模型
0 10 30 50
52 3 4 1 6 1 5 1 11 2
1 2 3 4 5 6 7 8 9 10

图12

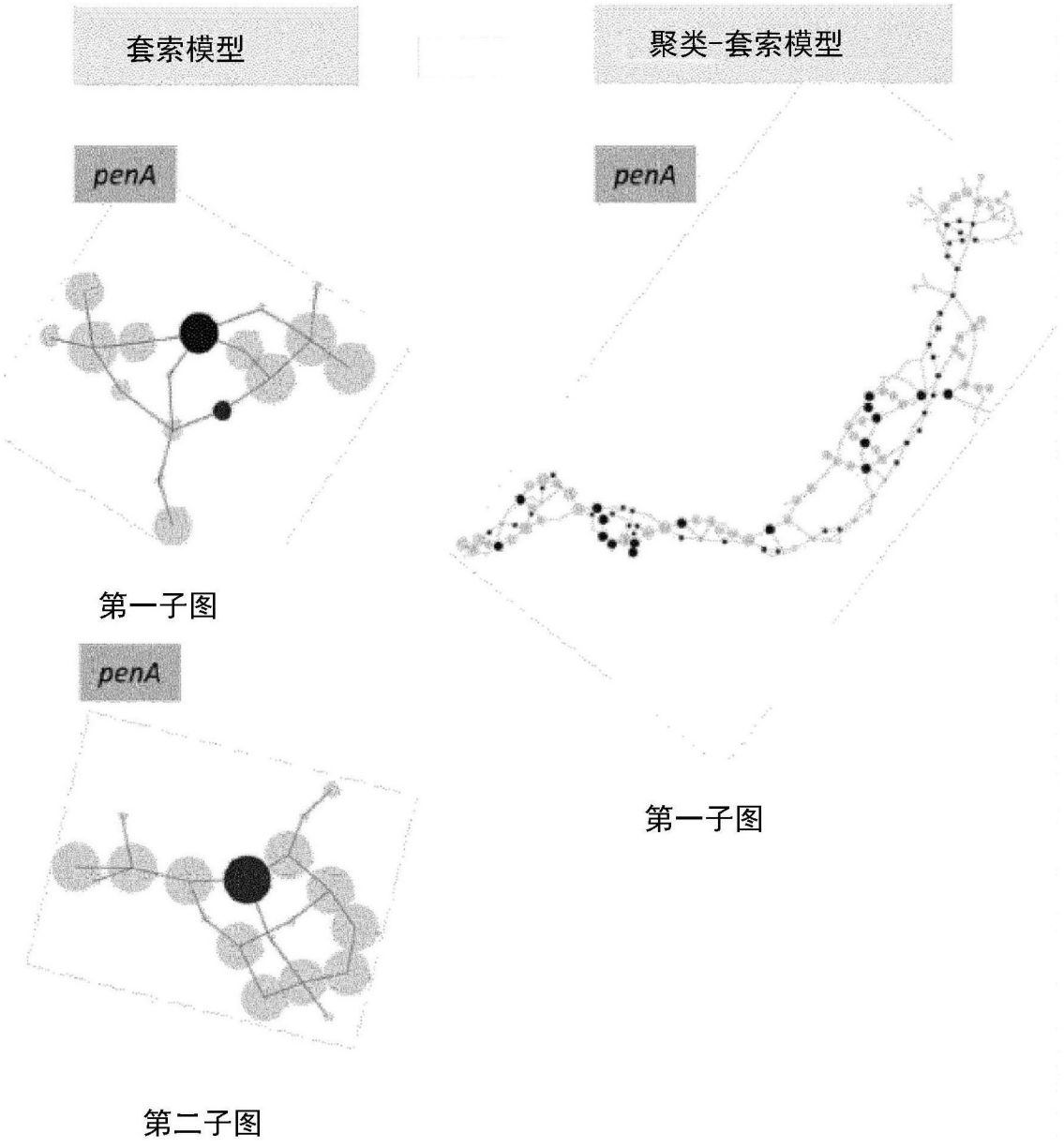


图13

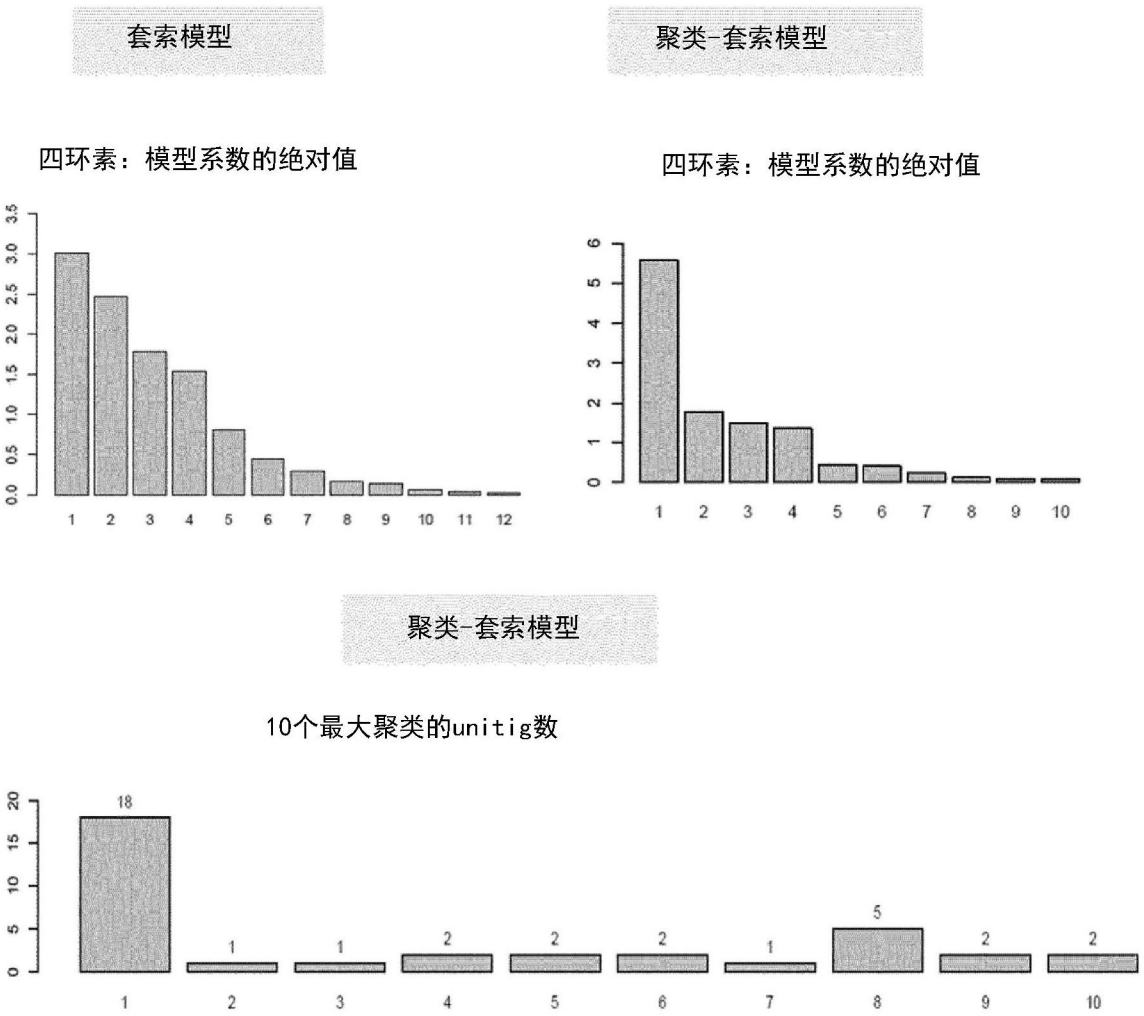

套索模型
聚类-套索模型
四环素：模型系数的绝对值
四环素：模型系数的绝对值
聚类-套索模型
10个最大聚类的unitig数
0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5
1 2 3 4 5 6 7 8 9 10 11 12
0 1 2 3 4 5 6
1 2 3 4 5 6 7 8 9 10
0 5 10 15 20
18 1 1 2 2 2 1 5 2 2
1 2 3 4 5 6 7 8 9 10

图14

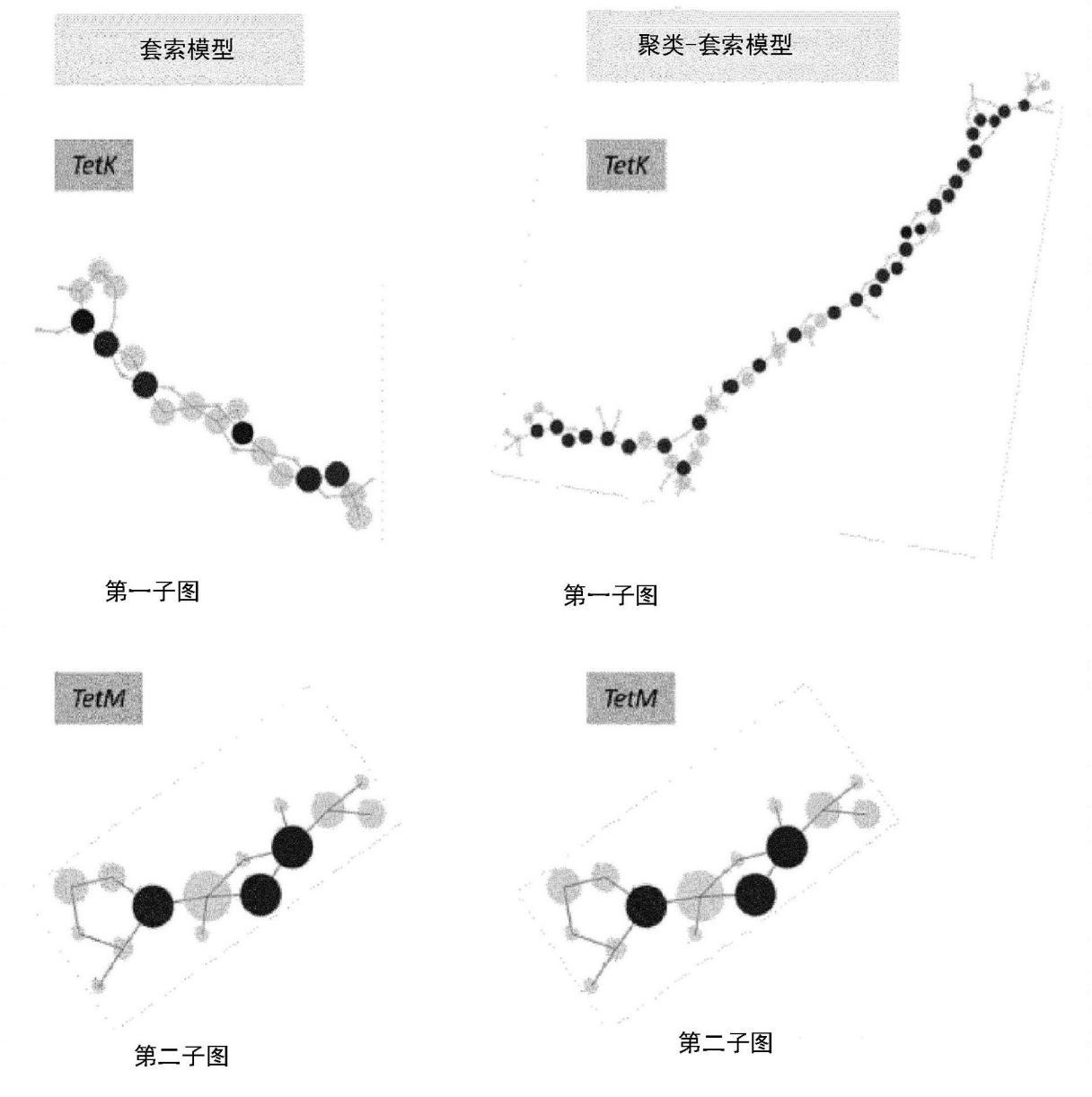


图15

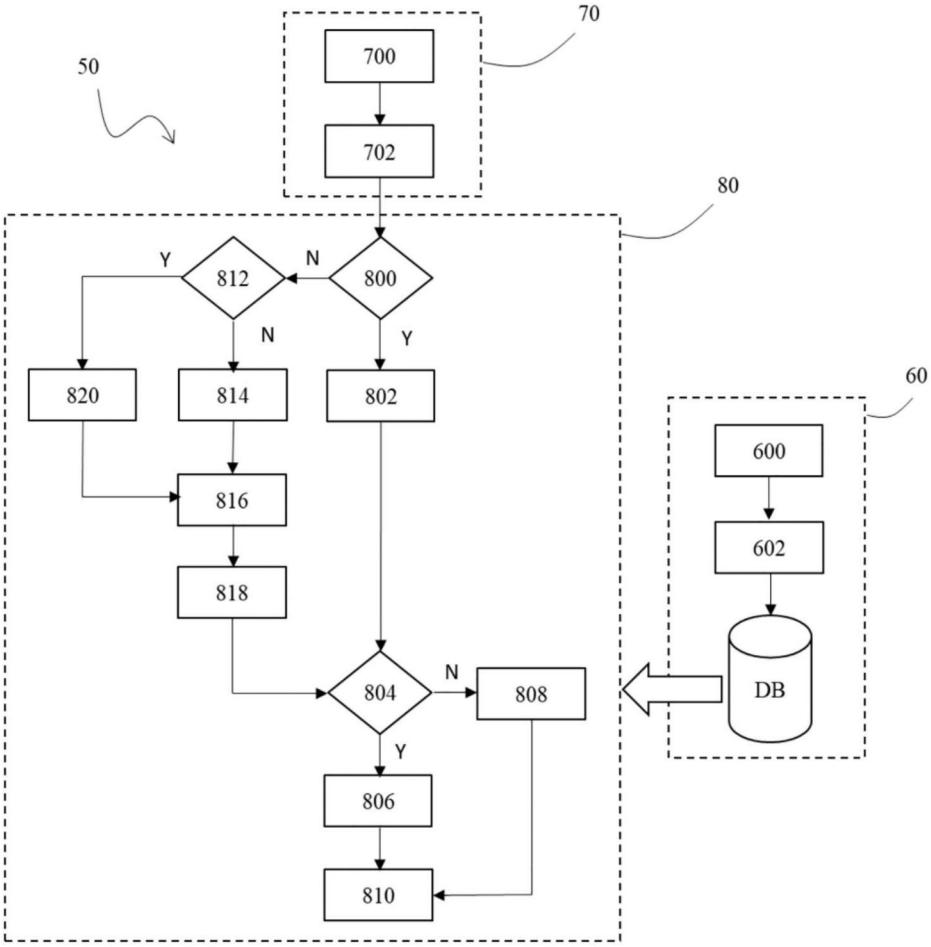

50
70
700
702
80
800
812
820
814
802
816
818
804
808
806
810
60
600
602
DB
Y
N
N
Y
N
Y

图16

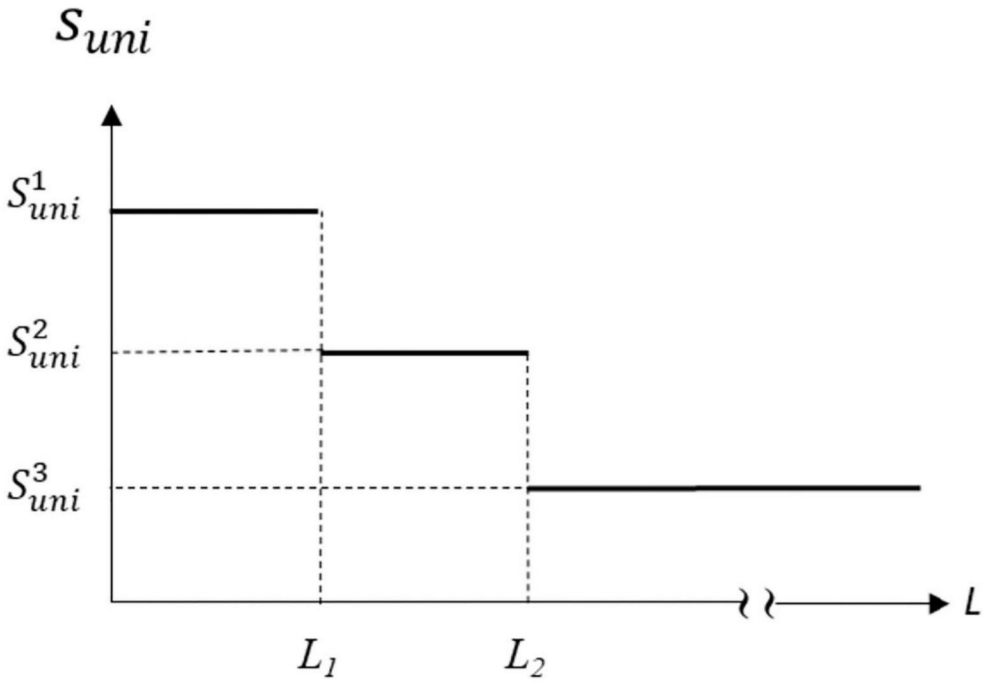


图18

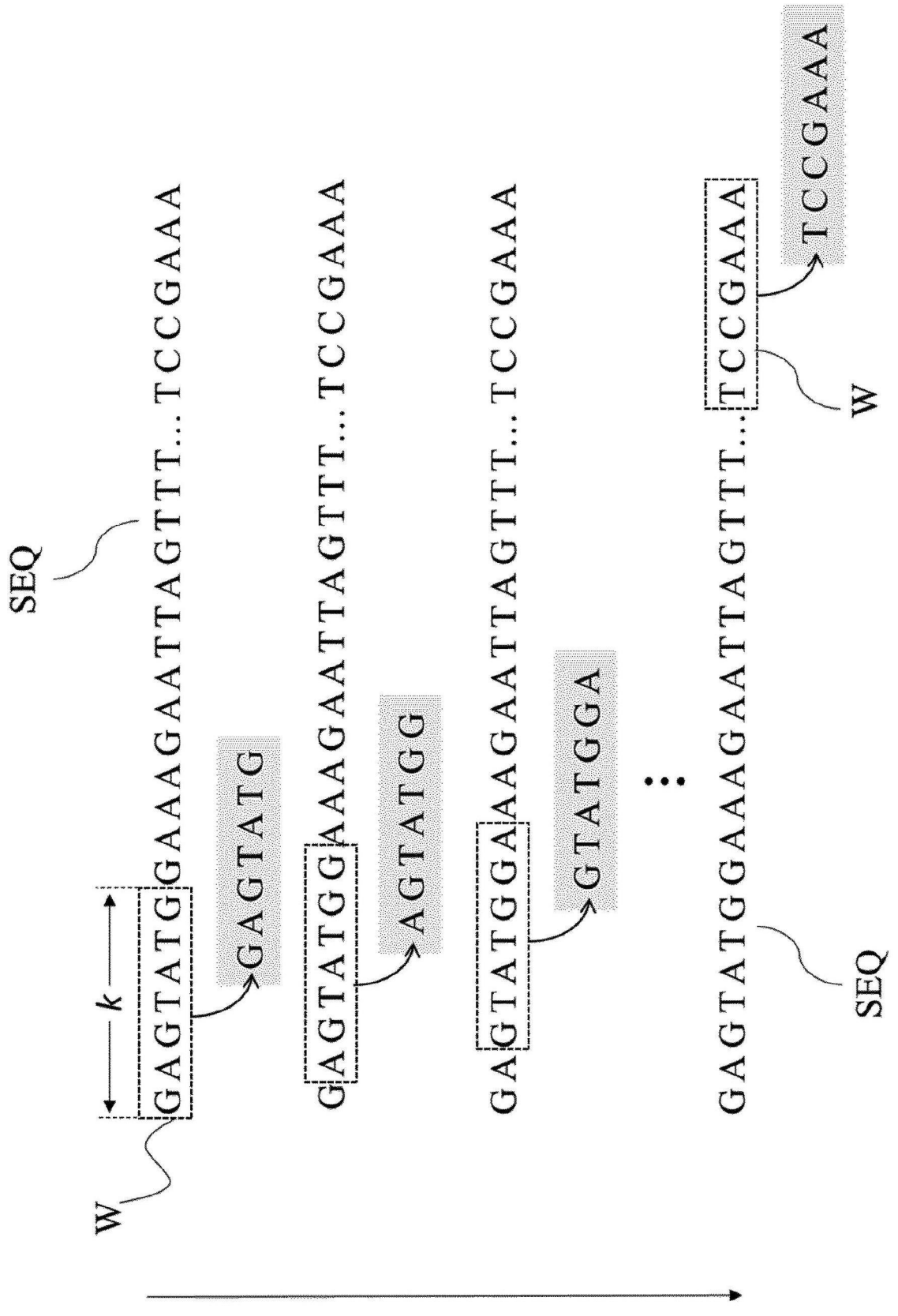

SEQ
GAGTATGGAAAGAATTAGTTTT...TCCGAAA
k
W
GAGTATG
GAGTATGGAAAGAATTAGTTTT...TCCGAAA
AGTATGG
GAGTATGGAAAGAATTAGTTTT...TCCGAAA
GTATGGA
GAGTATGGAAAGAATTAGTTTT...TCCGAAA
W
TCCGAAA
GAGTATGGAAAGAATTAGTTTT...TCCGAAA
SEQ

图17

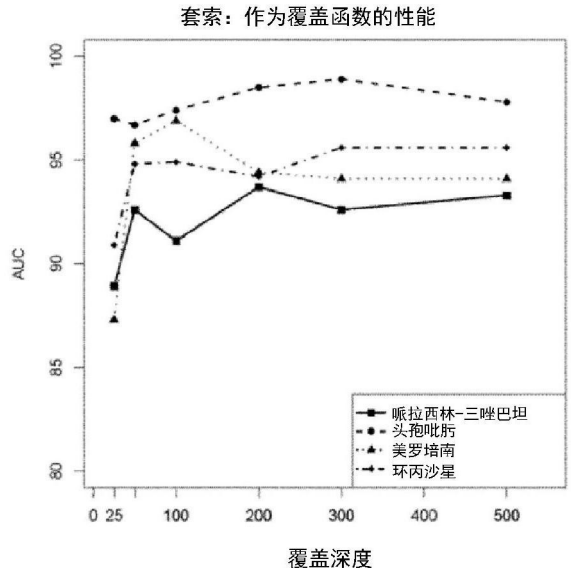

套索：作为覆盖函数的性能
AUC
覆盖深度
哌拉西林-三唑巴坦
头孢吡肟
美罗培南
环丙沙星

图19A

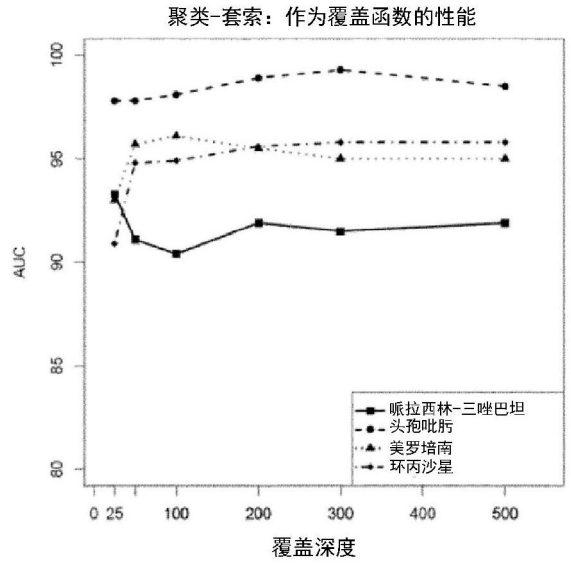

聚类-套索：作为覆盖函数的性能
AUC
覆盖深度
哌拉西林-三唑巴坦
头孢吡肟
美罗培南
环丙沙星

图19B

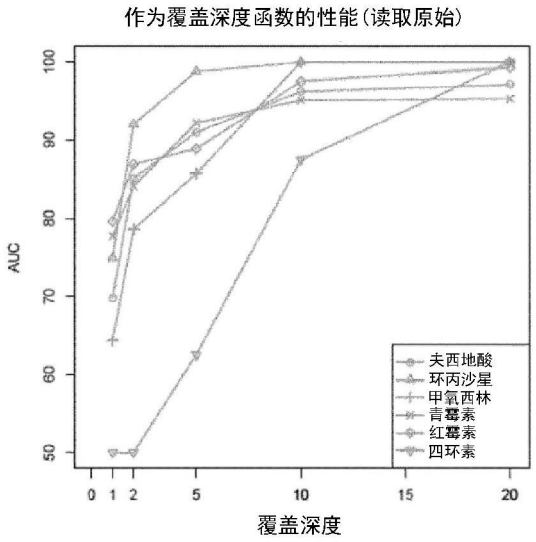

作为覆盖深度函数的性能(读取原始)
AUC
覆盖深度
夫西地酸
环丙沙星
甲氧西林
青霉素
红霉素
四环素

图20

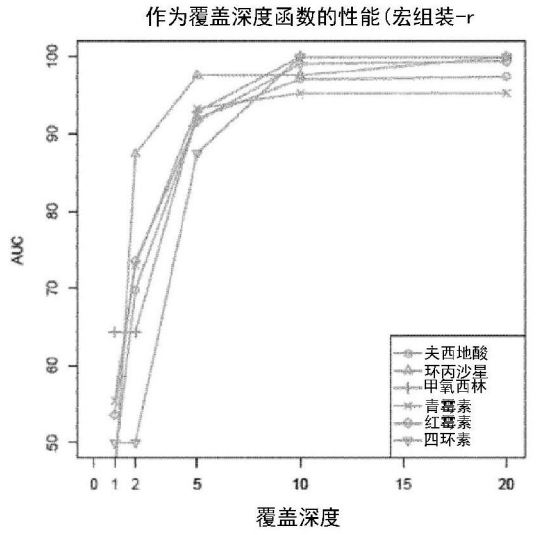

作为覆盖深度函数的性能(宏组装-r
AUC
覆盖深度
夫西地酸
环丙沙星
甲氧西林
青霉素
红霉素
四环素

图21