(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 104508670 A
(43)申请公布日 2015.04.08

(21)申请号 201380039795.0
(22)申请日 2013.06.21
(30)优先权数据
61/662,658 2012.06.21 US
(85)PCT国际申请进入国家阶段日
2015.01.27
(86)PCT国际申请的申请数据
PCT/EP2013/062984 2013.06.21
(87)PCT国际申请的公布数据
WO2013/190086 EN 2013.12.27
(71)申请人 菲利普莫里斯生产公司
地址 瑞士纳沙泰尔
申请人 向阳 朱丽娅·亨格
(72)发明人 向阳 朱丽娅·亨格
(74)专利代理机构 中国国际贸易促进委员会专利商标事务所 11038
代理人 杜文树
(51)Int.Cl.
G06F 19/24(2006.01)

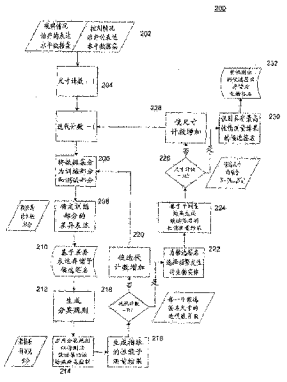
权利要求书2页 说明书12页 附图6页

(54)发明名称

用于生成生物标志物签名的系统和方法

(57)摘要

在此描述的系统和方法包括用于生成既可重现又可解释的基因签名的技术。该技术涉及对数据集进行重新采样并且选择具有高出现频率的基因。具体而言，在此描述的系统和方法包括对数据集的重复采样、基于通过重复采样处理生成的基因签名的发生频率对基因进行排名，以及迭代地选择最佳基因签名。



CN 104508670 A

1. 一种用于识别疾病状态的生物签名的计算机实现的方法，包括：

(a) 接收多个数据集，每一个数据集包括在包括疾病状态和控制状态在内的不同状态下的生物系统中的多个生物实体的表达水平；

(b) 针对多次迭代中的每一次：

(i) 将所述多个数据集分为训练部分和测试部分，

(ii) 使用所述多个数据集中的训练部分来将所述训练部分中具有高于阈值的差异表达的预定数目的生物实体存储作为子候选签名，并且生成将所述数据集中的每一个指派给疾病类和控制类中的一个的分类规则，并且

(iii) 使用所述多个数据集中的测试部分来应用所述分类规则以将每一个数据集指派给所述疾病类和所述控制类中的一个，并且基于所述指派生成性能子测量结果；

(c) 通过从所述子候选签名的聚合中选择频繁识别出的生物实体来生成具有预定数目生物实体的候选签名；

(d) 基于所述性能子测量结果来生成与所述候选签名相关联的性能测量结果；

(e) 针对所述预定数目的多个不同值重复步骤 (b) 至 (d) 以生成多个候选签名和多个相关联的性能测量结果；以及

(f) 将与最高性能测量结果相关联的所述候选签名存储作为所述生物签名。

2. 如权利要求 1 所述的方法，还包括通过比较对应的疾病状态表达水平和对应的控制状态表达水平来使用所述训练部分确定每一个生物实体的差异表达。

3. 如权利要求 1-2 中任一个所述的方法，其中，所述分类规则基于所述数据集内的所述生物实体的表达水平来指派所述数据集中的每一个。

4. 如权利要求 1-3 中任一个所述的方法，其中，通过将每一个数据集的指派与和所述数据集相关联的不同状态相比较来生成所述性能子测量结果。

5. 如权利要求 1-4 中任一个所述的方法，其中，所述多个生物实体包括基因、miRNA、蛋白质或者前述者中的两者或多者的组合中的一个或多个。

6. 如权利要求 1-5 中任一个所述的方法，其中，表达水平包括甲基化数据、基因表达数据、miRNA 表达数据和蛋白质表达数据中的一个或多个。

7. 如权利要求 1-6 中任一个所述的方法，其中，确定差异表达包括微阵列显著性分析 (SAM) 分析和 Limma 分析中的至少一种。

8. 如权利要求 1-7 中任一个所述的方法，其中，生成分类规则包括支持向量机方法。

9. 如权利要求 1-8 中任一个所述的方法，其中，生成所述性能子测量结果包括计算正确指派的数据集的百分比。

10. 如权利要求 1-9 中任一个所述的方法，其中，生成所述性能子测量结果包括计算所指派的数据集的马修相关系数。

11. 如权利要求 1-10 中任一个所述的方法，其中，所述子候选签名的聚合包括所述子候选签名中包括的所有生物实体的并集。

12. 如权利要求 1-11 中任一个所述的方法，其中，生成所述性能测量结果包括求与所述预定数目相关联的子候选签名的所有性能子测量结果的平均数。

13. 如权利要求 1-12 中任一个所述的方法，还包括在显示设备上显示与所述预定数目的多个不同值形成对照的多个性能测量结果的示图，以及可选地显示所述候选签名中包括

的生物实体的列表。

14. 一种包括计算机可读指令的计算机程序产品，所述计算机可读指令当在包括至少一个处理器的计算机化的系统中执行时使所述处理器执行如权利要求 1-13 中任一个所述的方法中的一个或多个步骤。

15. 一种包括配置有非临时计算机可读指令的至少一个处理器的计算机化的系统，所述非临时计算机可读指令当被执行时使所述至少一个处理器执行如权利要求 1-13 中任一个所述的方法。

用于生成生物标志物签名的系统和方法

[0001] 相关申请的引用

[0002] 本发明根据 35U.S.C.§119 要求在 2012 年 6 月 21 日提交的题为“Systems and Methods for Generating Biomarker Signatures”的美国临时专利申请第 61/662,658 号的优先权。

技术领域

[0003] 在生物医学领域，识别指示特定生物状态的物质（即生物标志物 (biomarker)）是重要的。随着基因组和蛋白质组的新技术出现，生物标志物在生物发现、药物开发和卫生保健中正变得愈发重要。生物标志物不仅对许多疾病的诊断和预后有用，而且对理解疗法开发的基础有用。生物标志物的成功和有效识别可以加速新药物开发过程。随着疗法与诊断和预后的结合，生物标志物识别也将增强当前医疗治疗的质量，从而在药物遗传学、药物基因组学和药物蛋白质组学的用途中发挥重要作用。

[0004] 包括高吞吐量筛选在内的基因组和蛋白质组分析提供了关于以细胞表达的蛋白质的数目和形式的大量信息并且提供了针对每一个细胞来识别特定细胞状态的表达蛋白质特性的简档的可能。在某些情况下，该细胞状态可能是以与疾病相关联的异常生理反应为特征。结果，识别并且比较来自具有疾病的患者的细胞状态和来自正常患者的对应细胞的细胞状态可以提供诊断和治疗疾病的机会。

[0005] 这些高吞吐量筛选技术提供了基因表达信息的大数据集。研究者已经尝试开发用于将这些数据集组织为对个体的多样人群是可重现诊断性的模式的方法。一种方法是聚集来自多个源的数据以形成组合数据集然后将该数据集分为发现 / 训练集和测试 / 验证集。然而，转录分析数据 (profiling data) 和蛋白质表达分析数据经常以与样本的可用数目有关的大量变量为特点。

[0006] 来自患者群组或者控制群组的标本的表达谱 (expression profile) 之间的观察差异通常被若干因素遮蔽，这些因素包括疾病或者控制人群内的生物变化性或者未知子表型、由研究方案的差异引起的特定于部位的偏差、标本处理、由仪器条件的差异（例如，芯片批次等）引起的偏差以及由测量误差引起的变化。

[0007] 若干基于计算机的方法已被开发以寻找最好说明疾病和控制样本之间的差异的一组特征（标志物）。一些早期方法包括诸如 LIMMA 之类的统计测试、用于识别与乳腺癌有关的生物标志物的 FDA 批准的 mammaprint 技术、逻辑回归技术以及诸如支持向量机 (SVM) 之类机器学习方法。一般而言，从机器学习的角度，生物标志物的选择通常是分类任务的特征选择问题。然而，这些早期解决方案面临若干缺点。通过这些技术生成的签名不是可重现的，这是因为对象的包括和排除可以导致不同的签名。这些早期解决方案也不是鲁棒性的，这是因为它们对具有小样本尺寸和高维度的数据集进行操作。此外，通过这些技术生成的签名包括许多假阳性并且难以以生物方式解释，这是因为技术和基因签名本身都不阐明底层生物机制。结果，因为它们不是可重现的并且难以解释，因此它们对临床诊断可能不是特别有用。

[0008] 较新的技术涉及将关于正则通路(canonical pathway)和蛋白质-蛋白质交互作用的知识集成到基因选择算法中。另外,若干特征选择技术已被开发,并且这些技术包括过滤方法、包装方法和嵌入方法。过滤方法独立于分类器设计而工作并且通过考虑数据的内在属性来执行特征选择。包装和嵌入方法通过利用特定分类模型来执行特征选择。包装方法在分类模型的预测性能的引导下在可能特征子集的空间中使用搜索策略。嵌入式方法利用分类模型内部参数来执行特征选择。然而,这些技术也面临若干缺点。

[0009] 因此,存在对为了临床诊断、预后或者这两者而识别生物标志物的改进技术的需要。

发明内容

[0010] 如上面提到的,早期的解决方案以及更新的嵌入和包装方法面临若干缺点。具体而言,申请人已经认识到这些方法依赖于所使用的具体类型的分类方法。换言之,如果分类方法不适合用户数据的类型,那么这些方法通常倾向于失败或者不佳地执行。申请人已经进一步认识到多个方法的整体倾向于做得比单独方法更好。在此描述的计算机系统和计算机程序产品实现了包括一个或多个这种整体技术并且包括用于生成可重现且可解释的基因签名的方法。该技术涉及对数据集进行重新采样并且选择具有高出现频率的基因。具体而言,在此描述的计算机实现的方法包括对数据集的重复采样、基于通过重复采样处理生成的基因签名的发生频率对基因进行排名,以及迭代地选择最佳基因签名。

[0011] 在某些方面,在此描述的系统和方法包括用于识别疾病情况的生物签名或者一组生物标志物的装置和方法。这些方法可以包括接收多个数据集,每一个数据集包括生物系统中的多个生物实体中的每一个生物实体的活动或者表达水平数据。生物系统可以处于若干个状态中的一个。例如,生物系统可以处于由暴露于物质而引起的扰动状态。在另一个示例中,生物系统可以处于疾病情况的状态,或者处于控制情况或者正常情况的状态。这些方法还可以包括多次迭代,针对每一次迭代,将多个数据集分为训练部分和测试部分。这多个数据集中的训练部分可以被用来通过比较对应于生物系统的两种不同状态(例如,疾病状态和正常状态)的表达水平来确定每一个生物实体的差异表达。另外,训练部分可以被用来将训练部分中具有高于阈值的差异表达的预定数目的生物实体存储作为子候选签名。训练部分还可以被用来生成基于数据集内的识别出的生物实体的表达水平将数据集中的每一个指派给疾病类和正常或控制类中的一个的分类规则。

[0012] 针对多次迭代中的每一次,这些方法还可以包括使用多个数据集中的测试部分来应用分类规则以将每一个数据集指派给疾病类和正常/控制类中的一个,并且通过将每一个数据集的指派与和该数据集相关联的生物系统的状态相比较来生成子候选签名的性能子测量结果。在某些实施例中,这些方法包括通过从子候选签名的聚合中选择频繁排名高的生物实体来生成具有预定数目生物实体的候选签名,以及基于性能子测量结果来生成与候选签名相关联的性能测量结果。在某些实施例中,这些方法包括针对预定数目的多个不同值重复以上步骤中的一个或多个以生成多个候选签名和多个相关联的性能测量结果。然后,与最高性能测量结果或者超过某一阈值的性能测量结果相关联的候选签名被存储作为生物签名。

[0013] 在上述方法的某些实施例中,多个生物实体包括基因和miRNA中的一个或多个。

表达水平可以包括甲基化数据、基因表达数据、miRNA 表达数据和蛋白质表达数据中的一个或多个。在上述方法的某些实施例中，确定差异表达的步骤包括微阵列显著性分析 (SAM) 分析和 Limma 分析中的至少一种。Limma 较之 SAM 可以是优选的，这是因为 Limma 与更高效率和对计算能力的更低要求相关联。在这些方法的某些实施例中，生成分类规则的步骤可以包括支持向量机方法。一般而言，分类器可以包括基于网络的支持向量机、基于神经网络的分类器、逻辑回归分类器、基于决策树的分类器、使用线性判别分析技术、随机森林分析计数的分类器，或者前述者的组合。

[0014] 在这些方法的某些实施例中，生成性能子测量结果的步骤可以包括计算正确指派的数据集的百分比。在这些方法的某些实施例中，生成性能子测量结果的步骤包括计算所指派的数据集的马修相关系数。在这些方法的某些实施例中，子候选签名的聚合可以包括子候选签名中包括的所有生物实体的并集。在这些方法的某些实施例中，生成性能测量结果的步骤还可以包括求与预定数目相关联的子候选签名的所有性能子测量结果的平均数。在这些方法的某些实施例中，这些方法还包括显示与预定数目的多个不同值形成对照的多个性能测量结果的示图，以及可选地显示候选签名中包括的生物实体的列表。在某些实施例中，这些方法包括在显示设备上显示与预定数目的多个不同值形成对照的多个性能测量结果的示图。这些方法还可以包括在显示设备显示候选签名中包括的生物实体的列表。

[0015] 本发明的计算机系统如上所述包括用于实现方法的各种实施例的装置。例如，计算机程序产品被描述，该产品包括计算机可读指令，这些计算机可读指令当在包含至少一个处理器的计算机化系统中执行时使处理器执行在上面描述的任何方法中的一个或多个步骤。在另一个示例中，计算机化系统被描述，该系统包含配置有非临时计算机可读指令的处理器，这些非临时计算机可读指令当被执行时使处理器执行在上面描述的任何方法。计算机程序产品和在此描述的计算机化的方法可以在具有一个或多个计算设备的计算机化系统中实现，每个计算设备包括一个或多个处理器。一般而言，在此描述的计算机化系统可以包含一个或多个引擎，这一个或多个引擎包括被配置为具有硬件、固件和软件以执行在此描述的一种或多种计算机化的方法的处理器或设备，例如，计算机、微处理器、逻辑器件或者其他器件或处理器。这些引擎中的任何一个或多个可以是与一个或多个其他引擎在物理上可分离的，或者可以包括多个在物理上可分离的组件，例如共同或者不同的电路板上的分离处理器。本发明的计算机系统包含用于实现如上所述的方法及其各种实施例的装置。引擎可以时不时地互连，并且还时不时地与一个或多个数据库连接，这一个或多个数据库包括可测量数据库、实验数据数据库和文献数据库。在此描述的计算机化系统可以包括具有通过网络接口通信的一个或多个处理器和引擎的分布式计算机化系统。这样的实现方式可能适合于经由多种通信系统进行的分布式计算。

附图说明

[0016] 在考虑到结合附图理解的以下具体实施方式之后，本公开的其他特征、其性质和各种优点将会显而易见，在附图中相似的引用符号自始至终指的是相似的部件，并且在附图中：

[0017] 图 1 示出了用于识别一个或多个生物标志物签名的示例性系统；

[0018] 图 2 示出了用于识别一个或多个生物标志物签名的示例性处理；

[0019] 图 3 是示出数据样本的分类和分类规则的确定的示图；

[0020] 图 4 是示出每一个具有不同数目成分的多个生物标志物签名的性能的示图；

[0021] 图 5 是示例性生物标志物签名生成工具的截屏；

[0022] 图 6 示出了由图 1 的系统生成的示例性 420 基因签名生物标志物的热图 (heatmap)；并且

[0023] 图 7 是诸如图 1 的系统中的任何组件和图 5 的截屏的计算设备的框图。

具体实施方式

[0024] 为了提供对在此描述的系统和方法的整体理解，现在将描述某些例示性实施例，包括用于识别基因生物标志物签名的系统和方法。然而，本领域普通技术人员将会明白在此描述的系统和方法可以针对其他合适应用而被适配和修改并且这种其他添加和修改将不脱离其范围。

[0025] 在此描述的系统和方法包括用于生成可重现且可解释的基因签名的技术。这些技术涉及对数据集进行重新采样并且选择具有高出现频率的基因。具体而言，在此描述的系统和方法包括对数据集的重复采样、基于通过重复采样处理生成的基因签名的发生频率对基因进行排名，以及迭代地选择最佳基因签名。一般而言，在此描述的计算机化系统可以包括一个或多个引擎，这一个或多个引擎包括被配置为具有硬件、固件和软件以执行在此描述的一种或多种计算机化的方法的一个或多个处理装置，例如计算机、微处理器、逻辑器件或者其他器件或处理器。

[0026] 图 1 示出了用于识别一个或多个生物标志物签名的示例性系统 100。系统 100 包括生物标志物生成器 102 和生物标志物合并器 (consolidator)104。系统 100 还包括用于控制生物标志物生成器 102 和生物标志物合并器 104 的操作的某些方面的中央控制单元 (CCU)101。在操作期间，在生物标志物生成器 102 处接收到诸如基因表达数据之类的数据。生物标志物生成器 102 处理该数据以生成多个候选生物标志物和对应的错误率。生物标志物合并器 104 接收这些候选生物标志物和错误率并且选择具有最佳的性能测量结果和尺寸的合适生物标志物。

[0027] 生物标志物生成器 102 包括用于处理数据和生成一组候选生物标志物和候选错误率的若干组件。具体而言，生物标志物生成器包括用于将数据分为训练数据集和测试数据集的数据预处理引擎 110。生物标志物生成器 102 包括用于接收训练数据集并生成候选生物标志物的生物标志物识别引擎 112，用于接收候选生物标志物并将测试数据分为两类之一（例如，疾病数据和控制数据）的分类器 114。生物标志物生成器 102 包括用于确定候选生物标志物相对于由数据预处理引擎 110 选择的测试数据的性能的分类器性能监视引擎 116。分类器性能监视引擎 116 生成性能测量结果，性能测量结果可以包括一个或多个候选生物标志物的候选错误率。生物标志物生成器 102 还包括用于存储一个或多个候选生物标志物和候选性能测量结果的生物标志物存储装置 118。

[0028] 生物标志物生成器可以受 CCU 101 控制，CCU 101 继而可以被自动控制或是用户操作的。在某些实施例中，生物标志物生成器 102 可以操作来生成多个候选生物标志物，每次将数据随机分为训练数据集和测试数据集。为了生成这样的多个候选生物标志物，生物标志物生成器 102 的操作可以被迭代多次。CCU 101 可以接收包括候选生物标志物的期望

数目的一个或多个系统迭代参数，这一个或多个系统迭代参数继而可以被用来确定生物标志物生成器 102 的操作可以被迭代的次数。CCU 101 还可以接收包括期望生物标志物尺寸的其他系统参数，期望生物标志物尺寸可以代表生物标志物中的组件数目（例如，生物标志物基因签名中的基因数目）。生物标志物尺寸信息可以被生物标志物识别引擎 112 用来根据训练数据生成候选生物标志物。参考图 2-4 更详细地描述了生物标志物生成器 102 及其各个引擎的操作。

[0029] 生物标志物生成器 102 生成一个或多个候选生物标志物和候选错误率，这一个或多个候选生物标志物和候选错误率被生物标志物合并器 104 用来生成鲁棒的生物标志物。生物标志物合并器 104 包括生物标志物合意引擎 128，生物标志物合意引擎 128 接收多个候选生物标志物并且生成具有跨这多个候选生物标志物最频繁发生的基因的新生物标志物签名。生物标志物合并器 104 包括用于确定跨这多个候选生物标志物的总体错误率的错误计算引擎 130。类似于生物标志物生成器 102，生物标志物合并器 104 也可以受 CCU 101 控制，CCU 101 继而可以被自动控制或是用户操作的。CCU 101 可以接收、确定或者接收并确定最小生物标志物尺寸的合适阈值，并且使用该信息来确定用来操作生物标志物生成器 102 和生物标志物合并器 104 两者的迭代数目。在一个实施例中，在每次迭代期间，CCU 101 使生物标志物尺寸减一并且迭代生物标志物生成器 102 和生物标志物合并器 104 两者直到阈值被达到为止。在这样的一个实施例中，生物标志物合意引擎 128 针对每一次迭代输出新生物标志物签名和新总体错误率。生物标志物合意引擎 128 从而输出各自具有从阈值起上至最大生物标志物尺寸的不同尺寸的一组新生物标志物签名。生物标志物合并器 104 还包括生物标志物选择引擎 126，生物标志物选择引擎 126 审核这些新生物标志物签名中的每一个的性能测量结果或者错误率并选择最佳生物标志物以供输出。参考图 2-4 更详细地描述了生物标志物合并器 104 及其各个引擎的操作。

[0030] 图 2 示出了用于使用图 1 中的示例性系统 100 来识别一个或多个生物标志物签名的示例性处理 200。处理 200 以在数据预处理引擎 110 处接收一个或多个数据集开始（步骤 202）。一般而言，数据可以表示样本中的多个不同基因的表达值、诸如任何生物重要分析物的水平之类的各种表型特点，或者这两者。在某些实施例中，数据集可以包括疾病情况治疗的表达水平数据和控制情况治疗的表达水平数据。基因表达水平可以指的是由基因编码的分子数量，例如 RNA 或者多肽。mRNA 分子的表达水平可以包括 mRNA 的数量，mRNA 的数量由将 mRNA 编码的基因的转录活动和 mRNA 的稳定性决定，mRNA 的稳定性继而由 mRNA 的半衰期决定。基因表达水平还可以包括与由基因编码的给定氨基酸序列相对应的多肽的数量。相应地，基因的表达水平可以对应于从基因转录的 mRNA 的数量、由基因编码的多肽的数量，或者这两者。基因的表达水平还可以按照不同形式的基因产品的表达水平来分类。例如，由基因编码的 RNA 分子可以包括差异表达的剪接变异体、具有不同的开始或者停止部位的转录物、其他差异处理形式，或者这两者。由基因编码的多肽可以包含裂开、修改形式的多肽，或者这两者。多肽可以通过磷酸化作用、脂化、异戊烯化、硫酸盐化作用、羟基化、乙酰化作用、核糖基化作用、法呢酰化、碳水化合物的添加等来修改。另外，具有给定形式修改的多个形式的多肽可以存在。例如，多肽可以在多个部位被磷酸化并且表达不同水平的差异磷酸化蛋白质。

[0031] 在某些实施例中，细胞或者组织中的基因表达水平可以由基因表达谱表示。基因

表达谱可以指的是诸如细胞或者组织之类的标本中的基因的表达水平的特点表示。来自个体的标本中的基因表达谱的确定表示个体的基因表达状态。基因表达谱反映了信使 RNA 或者多肽或者其由细胞或者组织中的一个或多个基因编码的形式的表达。基因表达谱一般可以指的是生物细胞（核酸、蛋白质、碳水化合物）的谱，生物细胞的谱显示不同细胞或组织当中的不同表达模式。

[0032] 在某些实施例中，数据集可以包括表示样本中的多个不同基因的基因表达值的元素。在其他实施例中，数据集可以包括表示通过质谱分析法检测到的峰或者峰的高度。一般而言，每一个数据集可以包括至少一个生物状态类的多个形式。例如，生物状态类可以包括但不限于：样本的源（即，从中获得样本的患者）中的疾病的存在 / 不存在；疾病的阶段；疾病的风险；疾病复发的似然性；一个或多个基因位点处的共享基因型（例如，共同的 HLA 单体型；基因突变；基因的修改，诸如甲基化等）；暴露到剂（例如，诸如有毒物质或者潜在的有毒物质、环境污染物、候选药物等）或者情况（温度、pH 等）；人口统计特性（年龄、性别、重量；家庭史；先存情况的历史）；对剂的抵抗、对剂的敏感性（例如，对药物的反应度）等。

[0033] 数据集可以彼此独立以减少最终分类选择中的采集偏差。例如，它们可以使用不同的排除或者包括标准而被从多个源采集并且可以被在不同时间采集和可以从不同地点采集，即当考虑到定义生物状态类的特性之外的特性时数据集可以是相对异构的。对异构性有贡献的因素包括但不限于：由性别、年龄、种族划分引起的生物变化性；由饮食、锻炼、睡眠行为引起的个体变化性；以及由血液处理的临床方案引起的样本处理变化性。然而，生物状态类可以包括一个或多个共同特性（例如，样本源可以表示具有疾病和相同性别或者一个或多个其他共同人口统计特性的个人）。

[0034] 在某些实施例中，来自多个源的数据集是通过在不同时间、在不同条件下或者在不同时间且在不同条件下从相同的患者人群采集样本而生成的。然而，来自多个源的数据集不包括更大数据集的子集，即，来自多个源的数据集是独立采集的（例如，来自不同地点、在不同时间、在不同采集条件下，或者前述者的组合）

[0035] 在某些实施例中，多个数据集是从多个不同临床试验地点获得的并且每一个数据集包括在每一个单独试验地点获得的多个患者样本。样本类型包括但不限于血液、血清、血浆、乳头抽出物、尿液、泪液、唾液、脊髓液、淋巴液、细胞、组织溶解产物、激光微解剖的组织或者细胞样本、嵌入的细胞或者组织（例如，在石蜡块中或者冷冻）；新鲜或者存档的样本（例如，来自验尸），或者前述者的组合。可以例如从试管中的细胞或组织培养中取得样本。可替代地，可以从活的有机体或者从诸如单细胞有机体的一群有机体取得样本。

[0036] 在一个示例中，当识别特定癌症的生物标志物时，可能从由两个不同测试地点处的独立群组选择的对象中采集血液样本，从而提供将根据其形成独立数据集的样本。

[0037] 返回图 2，在某些实施例中，可能期望使用生物标志物来在疾病情况治疗和控制情况治疗之间进行分类。在这种实施例中，数据可以包括例如疾病情况治疗的表达水平数据集和控制情况治疗的表达水平数据集。CCU 101 可以设置包括每一次迭代的计数的大小、迭代次数和初始迭代计数在内的系统参数（步骤 204）。在一个示例中，大小和迭代计数被设置为 1。

[0038] 数据预处理引擎 110 接收数据并且将数据分为训练数据集和测试数据集（步骤

206)。在某些实施例中,数据预处理引擎 110 随机地将数据分割或者划分为这两组。随机地划分数据对于预测类别和生成鲁棒基因签名可能是期望的。在其他实施例中,数据预处理引擎 110 基于数据的类型或者标签将数据分为两个或者更多组。一般而言,在不脱离本公开范围的情况下可以按照期望的任何合适方式将数据分为训练数据集和测试数据集。训练数据集和测试数据集可以具有任何合适尺寸并且可以是相同尺寸或者不同尺寸的。在某些实施例中,数据预处理引擎 110 在将数据分为训练数据集和测试数据集之前可以丢弃一条或多条数据。在某些实施例中,数据预处理引擎 110 在任何进一步处理之前可以丢弃来自训练数据集、测试数据集或者这两者的一条或多条数据。

[0039] 数据预处理引擎 110 将训练数据集传递给识别候选网络的生物标志物识别引擎 112(步骤 208)。生物标志物识别引擎 112 还接收生物标志物尺寸。在某些实施例中,生物标志物尺寸可以被选择为可允许的最大生物标志物尺寸,其中系统 100 进行迭代并且倒计数至最小生物标志物尺寸。在某些实施例中,生物标志物识别引擎 112 使用合适的统计技术来确定训练数据的差异表达。例如,每一个训练数据可以包括多个训练数据集,其中每一个训练数据集包括多个基因的探测集。对于这多个基因中的每一个,数据集包括对应于控制的已知值和治疗的另一个值。在某些实施例中,生物标志物识别引擎 112 跨多个训练数据集针对每一个基因来确定控制值与治疗值之间的距离。该距离可以通过 t 统计值——诸如通过 SAM 或者 Limma 计算的温和 t 统计值——来测量。Limma 是因对基因表达微阵列数据的分析——尤其是用于分析差异表达的线性模型的用途 (Smyth 2004, Statistical Applications in Genetics and Molecular Biology,第 3 卷,第 1 号,第 3 条,其通过引用而整体结合于此)——而众所周知的软件方法包。Limma 由于其效率和比 Sam 更低的对计算能力的要求而是优选的。生物标志物识别引擎 112 然后可以按照基因的 t 统计值来给基因进行排名。在某些实施例中,高排名可以表示该基因在控制和治疗之间被高度差异表达,并且低排名可以表示对于该基因在控制和治疗之间几乎没有差异。生物标志物识别引擎 112 可以选择基因的排名列表的一部分,例如基因列表的上半部。生物标志物识别引擎 112 所选择的基因的数目可以基于由 CCU 101 输入的生物标志物尺寸。在一个示例中,一个或多个转录因素即主调控基因可以被选择。所选择的基因然后可以是代表性的或者可以构成候选生物标志物。生物标志物识别引擎 112 可以将候选生物标志物输出给分类器 114、分类器性能监视引擎 116 和生物标志物存储装置 118(步骤 210)。

[0040] 分类器 114 可以接收来自生物标志物识别引擎 112 的一个或多个候选生物标志物。分类器 114 还可以接收来自数据预处理引擎 110 的一组或多组测试数据。在某些实施例中,分类器 114 使用候选生物标志物来生成分类规则(步骤 212)。图 3 以图形方式示出了这样的一个分类规则 300。分类器 114 可以应用分类规则以将测试数据集指派给两个类中的任一个。例如,分类器 114 可以应用分类以将测试数据集指派给疾病或者控制(步骤 214)。在某些实施例中,分类器 114 可以包括支持向量机 (SVM) 分类器。在其他实施例中,分类器 114 可以包括基于网络的 SVM、基于神经网络的分类器、逻辑回归分类器、基于决策树的分类器、使用线性判别分析技术、随机森林分析技术的分类器,或者前述者的组合。

[0041] 分类器性能监视引擎 116 可以使用合适的性能度量来分析分类器 114 的性能(步骤 216)。具体而言,当分析分类器 114 的性能时,分类器性能监视引擎 116 可能正在分析一个或多个候选生物标志物的鲁棒性或者性能。在某些实施例中,性能度量可以包括错误率。

性能度量还可以包括被除以尝试的总预测的正确预测的数目。性能度量可以是不脱离本公开范围的任何合适度量。候选生物标志物和对应的性能度量可以被存储在生物标志物存储装置 118 中。

[0042] 在某些实施例中，从步骤 206 到步骤 216 的处理可以被重复任何次数以生成多个候选生物标志物以及对应的性能度量。在每一次重复期间，数据可以被随机地划分为训练集和测试数据集。CCU 101 可以控制生物标志物生成器 102 的操作来执行这种重复分析。在某些实施例中，CCU 101 可以提供固定的迭代计数 R（步骤 218）。在这种实施例中，可以通过重复来生成 R 个候选生物标志物，每次增加迭代编号（步骤 220）。一旦迭代已经完成，CCU 101、生物标志物生成器 102 或者这两者可以计算所有候选生物标志物的复合性能分数。复合性能分数可以是候选生物标志物的性能度量的平均值。在某些实施例中，数据集可以是不平衡的（即，例如治疗和控制的不同状态的不等数目）。在这种实施例中，可以使用马修相关系数 (MCC) 来确定性能分数。

[0043]

$$MCC = \frac{TP*TN - FP*FN}{\sqrt{(TP+FP)*(TP+FN)*(TN+FP)*(TN+FN)}}$$

[0044] 其中 TP：真阳性；FP：假阳性；TN：真阴性；FN：假阴性。

[0045] 如早先提到，CCU 101 还可以控制生物标志物合并器 104 的操作以基于在生物标志物生成器 102 中生成并存储的候选生物标志物来生成合适且鲁棒的生物标志物。生物标志物合并器 104 包括生物标志物合意引擎 128，生物标志物合意引擎 128 接收来自生物标志物存储装置 118 的一个或多个候选生物标志物。生物标志物合意引擎 128 可以为新的生物标志物签名选择一个或多个候选生物标志物内频繁发生的基因（步骤 222）。新的生物标志物签名可以包括 N 个基因，其中 N 是生物标志物的期望尺寸、生物标志物的最大允许尺寸、生物标志物的最小允许尺寸或者最大和最小尺寸之间的尺寸。在某些实施例中，数目 N 可以是用户可选择的并且可以是按需可调节的。

[0046] 在某些实施例中，生物标志物合意引擎 128 基于每一个基因跨所有候选生物标志物的出现来计算每一个基因的频率。在数学上，生物标志物合意引擎 128 可以取候选网络中的基因的并集然后计算每一个基因的发生频率如下：

[0047]

$$r_{j,N} = \frac{\sum_{iter=1}^{R} f(j, iter, N) \times P(N, iter)}{R},$$

[0048]

$$f(j, iter, N) = 1, j \in GS(N, iter); 0, j \notin GS(N, iter)$$

[0049] 其中 $r_{j,N}$ 是当我们选择头 N 个基因时基因 j 的总体加权频率；GS(N, iter) 是头 N 个基因针对迭代 iter 的子基因签名；P(N, iter) 是测试数据中的 GS(N, iter) 的预测性能。生物标志物合意引擎 128 可以生成按照基因跨候选生物标志物的发生频率来排名的基因的列表。

[0050] 生物标志物合意引擎 128 可以选择该列表的子集来形成期望长度的新生物标志物签名。错误计算引擎 130 确定所有候选生物标志物的总体性能度量（步骤 224）。该总体

性能度量可以与如上所述由生物标志物生成器 102 确定的复合分数相同。

[0051] 研究者在识别生物标志物时面临的一个挑战是确定其尺寸。每一种疾病情形可以证实不同尺寸的生物标志物并且因而研究者可能难以确信地决定生物标志物应当是多长。本发明人已经认识到该问题的解决方案是迭代经过各种尺寸的生物标志物并且达到最好地预测测试数据并为其分类的一个。在某些实施例中，用户可以选择最大生物标志物签名尺寸和最小生物标志物签名尺寸。系统 100 可以迭代经过最大和最小生物标志物签名之间的每一个尺寸。在每一次迭代期间，生物标志物合意引擎 128 可以生成新的生物标志物签名并且错误计算引擎 130 可以生成该新生物标志物签名的对应性能分数。在某些实施例中，系统 100 可以从最大尺寸开始并且倒计数到最小尺寸。在其他实施例中，系统 100 可以从最小尺寸开始并且向上迭代至最大尺寸。系统 100 可以选择跳过某些尺寸或者可以重复某些尺寸，而不脱离本公开的范围。生物标志物选择引擎 126 然后可以从具有最高性能测量结果的一组生物标志物签名中选择合适的新生物标志物签名（步骤 230）。图 4 以图形方式示出了生物标志物选择引擎 126 的操作。具体而言，图 4 示出了由生物标志物合意引擎 128 生成的新生物标志物签名和由错误计算引擎 130 生成的对应性能测量结果的示图。N* 长度的生物标志物签名因为具有最高性能测量值而被选择。

[0052] 图 5 是用于识别和生成生物标志物签名的工具的截屏 500。该工具可以被实现在计算机上，由此后端是系统 100 并且前端显示在截屏 500 中示出的图形用户界面 (GUI)。GUI 可以被用来允许用户与系统 100 交互并且从而提供数据集并且接收关于潜在生物标志物签名的信息。例如，GUI 可以包括标识屏幕或者程序的标签 502、输入区域 504 和输出区域 506。输入区域 504 包括一个或多个文本框、标签、下拉菜单、单选按钮、命令按钮或者前述者的组合，以允许用户输入系统 100 的一个或多个变量、参数或者度量。例如，输入区域 504 可以包括用来输入在完成处理之前生物标志物生成器 102、生物标志物合并器 104 或者这两者应当迭代经过的次数的组件。输入区域 504 还可以允许用户输入最大签名尺寸、最小签名尺寸或者任何合适的签名尺寸。输入区域 504 还允许用户通过从本地磁盘或者远程磁盘上传来提供一个或多个数据集。GUI 还可以包括输出区域 506，输出区域 506 可以包括一个或多个候选生物标志物、新生物标志物签名、最终生物标志物签名或者两者的显示。输出区域 506 还可以包括一个或多个包括在图 3 和图 4 中示出的示图的示图。一般而言，GUI 可以包括来自系统 100 中的任何组件的任何输入、输出或者输入和输出。GUI 还可以允许包括电源管理、通信、显示、存储和数据管理在内的任何其他计算操作。

[0053] 示例

[0054] 在一个示例中，包括系统 100 在内的在此描述的系统和方法被用来生成并识别一基因签名，该基因签名帮助把烟草产品的前吸烟者与烟草产品的当前吸烟者区分开来。在这样一个示例中，提供给数据预处理引擎 110 的数据包括来自德克萨斯大学的 M.D. 安德森癌症中心的可公开获得的数据。这种数据被描述于“Impact of smoking cessation on global gene expression in the bronchial epithelium of chronic smokers”，Zhang L 等人，Cancer Prev. Res. 1:112-118，2008，其通过引用而被整体结合于此。该数据是通过对 13 个健康吸烟者 (HS) 和 8 个健康前吸烟者 (HExs)——即在采样被执行前多于 12 个月前戒烟的那些——的气道进行采样来生成的。吸烟者和前吸烟者的采样集是 78％白人和 61％男性。为了获取数据，来自气道的 RNA 提取被混杂生成 (hybridize) 为

Affymetrix GeneChip® Human Genome U133Plus 2.0 阵列。

[0055] 系统 100 被建立以分析该数据并生成将帮助将吸烟者与前吸烟者区分开来的基因签名。在该示例中，包括签名的最大尺寸在内的生物标志物尺寸被设置为 500，并被输入到 CCU 101。包括重新采样的最大数目的在内的系统迭代参数被设置为 300。数据预处理引擎 110 将数据随机分为包括大约 10％数据的测试数据集和包括剩余的大约 90％数据的训练数据集。在该示例中，分类器 114 被选择为诸如在下文中描述的分类器的 SVM 分类器：“Support vector networks. Machine Learning”，Cortes, C. 和 V. Vapnik, 1995-20(3):p. 273-297，该文通过引用而被整体结合于此。为了对基因进行排名，系统 100 包括合适的 SAM 引擎，诸如在下文中描述的 SAM：“Significance analysis of microarrays applied to the ionizing radiation response,"Tusher, V. G., R. Tibshirani 和 G. Chu, Proc Natl Acad Sci U S A, 2001. 98(9):p. 5116-21，该文通过引用而被整体结合于此。

[0056] 根据本发明的方法的系统 100 生成将前吸烟者与当前吸烟者区分开来的稳定 420- 基因签名。所生成的签名是具有小于或者等于 500 的尺寸的一组候选签名中表现最好的签名。图 6 示出了 420- 基因签名 600 的热图。热图的颜色可能未被用灰度清楚示出，但是图 6 的数据示出了其中氧化应激和异生物质代谢被富集 (enrich) 的 194 个基因在健康前吸烟者 (HExS) 的气道中被下调；其中细胞形态发生被富集的 226 个基因在 HExS 的气道中被上调。图 6 中示出的热图可以在用户界面 500 中被显示。

[0057] 本主题的实现方式可以包括，但不限于，包含如在这里描述的一个或多个特征以及包含可操作来使一个或多个机器（例如，计算机、机器人）产生在这里描述的操作的机器可读介质的物品的系统方法和计算机程序产品。在这里描述的方法可以通过位于单个计算系统或者多个计算系统中的一个或多个处理器或者引擎来实现。这样的多个计算系统可以被连接并且可以经由一个或多个连接来交换数据和 / 或命令或者其他指令等，这一个或多个连接包括但不限于通过网络（例如，因特网、无线广域网、局域网、广域网、有线网络等）的连接、经由多个计算系统中的一个或多个之间的直接连接。

[0058] 图 7 是计算设备的框图，计算设备例如是图 1 的系统 100 中的任何组件和包括用于执行参考图 2-4 描述的处理的电路的图 5 的 GUI500。系统 100 的每一个组件都可以在一个或多个计算装置 650 上实现。在某些方面，多个上述组件和数据库可以包括于一个计算设备 650 内。在某些实现方式中，组件和数据库可以跨若干个计算设备 650 而实现。

[0059] 计算设备 650 包括至少一个通信接口单元、输入 / 输出控制器 610、系统存储器和一个或多个数据存储设备。系统存储器包括至少一个随机存取存储器 (RAM 602) 和至少一个只读存储器 (ROM 604)。所有这些元件都与中央处理单元 (CPU 606) 通信以促进计算设备 650 的操作。计算设备 650 可以按照许多不同的方式来配置。例如，计算设备 650 可以是常规的独立式计算机，或者作为替代，计算设备 650 的功能可以跨多个计算机系统和体系架构分布。计算设备 650 可以被配置用于执行数据分割、差分、分类、评分、排名和存储操作中的一些或全部。在图 7 中，计算设备 650 经由网络或本地网络链接至其他服务器或系统。

[0060] 计算设备 650 可以按照分布式体系架构来配置，其中数据库和处理器被安放于分离的单元或位置。某些此类单元执行初级处理功能并且最低程度地含有通用控制器或处

理器和系统存储器。在这方面，这些单元中的每一个都经由通信接口单元 608 连结至用作与其他服务器、客户端或用户计算机及其他相关设备间的初级通信链路的通信集线器或端口（未示出）。通信集线器或端口自身可以具有最小限度的处理能力，主要用作通信路由器。多种通信协议可以作为系统的一部分，包括，但不限于：以太网 (Ethernet)、SAP、SAS™、ATP、BLUETOOTH™、GSM 和 TCP/IP。

[0061] CPU 606 包括处理器，例如，一个或多个常规的微处理器和一个或多个辅助协处理器，例如，用于转移 CPU 606 的工作负载的数学协处理器。CPU 606 与通信接口单元 1008 和输入 / 输出控制器 610 通信，由此 CPU 606 与诸如其他服务器、用户终端或设备之类的其他设备通信。通信接口单元 608 和输入 / 输出控制器 610 可以包括用于与例如其他处理器、服务器或客户端终端同步通信的多种通信通道。相互通信的设备不需要持续地相互发送信号。相反地，这样的设备只需要在必要时彼此发送信号，可以实际上大部分时间实际都避免交换数据，并且可以需要执行几个步骤来建立装置之间的通信链路。

[0062] CPU 606 同样与数据存储设备通信。数据存储设备可以包括磁存储器、光存储器或半导体存储器的适当组合，并且可以包括例如 RAM602、ROM 604、闪存驱动器、光盘（例如，紧凑盘）或者硬盘或硬盘驱动器。例如，CPU 606 和数据存储设备各自都可以完全位于单个计算机或其他计算设备之内；或者通过通信介质（例如，USB 端口、串口线、同轴线、以太网型线、电话线、射频收发器或者其他类似的无线或有线介质，或者前述者的组合）彼此连接。例如，CPU 606 可以经由通信接口单元 608 与数据存储设备连接。CPU 606 可以被配置用于执行一个或多个特定的处理功能。

[0063] 数据存储设备可以存储例如 (i) 计算设备 650 的操作系统 1012；(ii) 适用于根据这里描述的系统和方法并且尤其是根据针对 CPU606 详细描述的过程来引导 CPU 606 的一个或多个应用 614（例如，计算机程序代码或计算机程序产品）；或者 (iii) 适用于存储可以用来存储程序所需的信息的信息的一个或多个数据库 616。在某些方面，一个或多个数据库包括存储实验数据以及发表的文献模型的数据库。

[0064] 操作系统 612 和应用 614 可以按照例如压缩的、不压缩的和加密的格式来存储，并且可以包括计算机程序代码。程序的指令可以从数据存储设备以外的计算机可读介质——例如从 ROM 604 或 RAM 602——读入处理器的主存储器内。虽然程序中的指令序列的执行促使 CPU 606 执行在此描述的过程步骤，但是也可以使用硬连线电路来代替用于实现本发明的过程的软件指令或者与其结合。因而，所描述的系统和方法并不限定于硬件和软件的任何具体结合。

[0065] 合适的计算机程序代码可以被提供用于执行与在此描述的建模、评分和聚合相关的一个或多个功能。程序同样可以包括程序元素，例如，操作系统 612、数据库管理系统以及允许处理器经由输入 / 输出控制器 610 与计算机外围设备（例如，视频显示器、键盘、计算机鼠标等）接口连接的“设备驱动程序”。

[0066] 包含计算机可读指令的计算机程序产品也被提供。计算机可读指令当在计算机系统上加载和执行时使计算机系统根据上面描述的方法或其一个或多个步骤来操作。在这里使用的术语“计算机可读介质”指的是用于给计算设备 650 的处理器（或者在此描述的设备的任何其他处理器）提供或参与提供用于执行的指令的任何非临时性介质。这样的介质可能要采取许多形式，包括但不限于非易失性介质和易失性介质。非易失性介质包括

例如光盘、磁盘或光磁盘，或者集成电路的存储器，例如，闪存存储器。易失性介质包括通常构成主存储器的动态随机存取存储器 (DRAM))。计算机可读介质的共同形式包括，例如，软盘、柔性盘、硬盘、磁带、任何其他磁介质、CD-ROM、DVD、任何其他光介质、打孔卡、纸带、具有孔图形的任何其他物理介质、RAM、PROM、EPROM 或 EEPROM(电可擦可编程只读存储器)、FLASH-EEPROM、任何其他存储器芯片或盒，或者计算机能够读取的任何其他非临时性介质。

[0067] 各种形式的计算机可读介质可以涉及将一个或多个指令的一个或多个序列传递给 CPU 606(或者在此描述的设备的任何其他处理器)，以用于执行。例如，指令最初可以位于远程计算机 (未示出) 的磁盘上。远程计算机能够将指令加载到其动态存储器内，以及经由以太网连接、电线或甚至是使用调制解调器的电话线来发送指令。位于计算设备 650(例如，服务器) 本地的通信装置能够接收在各自的通信线路上的数据，并且将数据放置于处理器的系统总线上。系统总线将数据输送到主存储器，处理器从该主存储器中检索出指令并执行。由主存储器接收到的指令可以在由处理器执行之前或之后可选地存储于存储器内。另外，指令可以经由通信端口作为电信号、电磁信号或光信号来接收，这些信号是携带各种类型的信息的无线通信或数据流的示例性形式。

[0068] 虽然已经参考具体示例具体示出并描述了本发明的实现方式，但是本领域技术人员应当明白在不脱离本公开的精神和范围的情况下可以在其中做出形式和细节的各种改变。

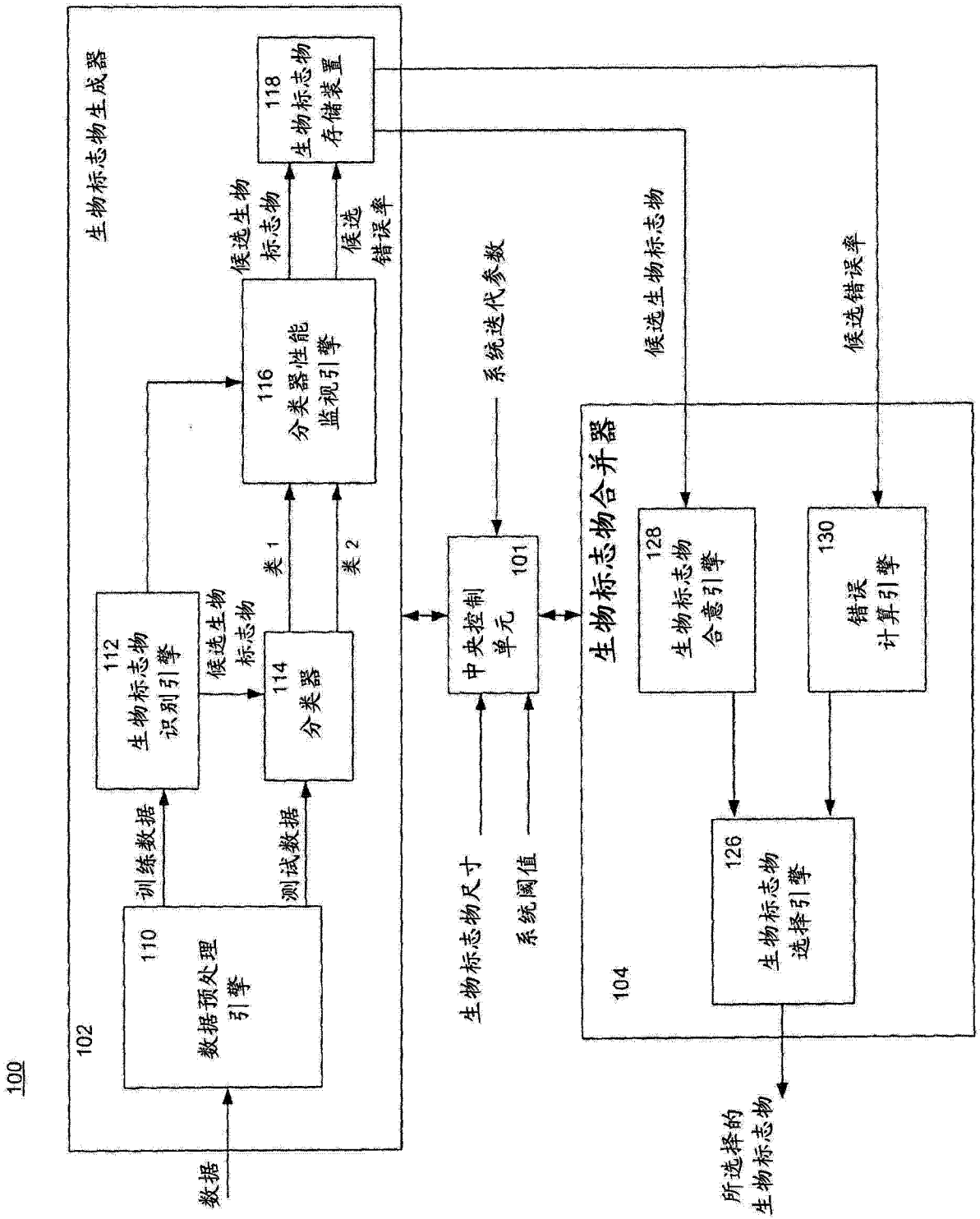

100
生物标志物生成器
102
数据
数据预处理引擎
110
训练数据
测试数据
生物标志物识别引擎
112
候选生物标志物
分类器
114
类 1
类 2
分类器性能监视引擎
116
候选生物标志物
候选错误率
生物标志物存储装置
118
中央控制单元
101
系统迭代参数
生物标志物尺寸
系统阈值
生物标志物合并器
104
生物标志物合意引擎
128
错误计算引擎
130
生物标志物选择引擎
126
候选生物标志物
候选错误率
所选择的生物标志物

图 1

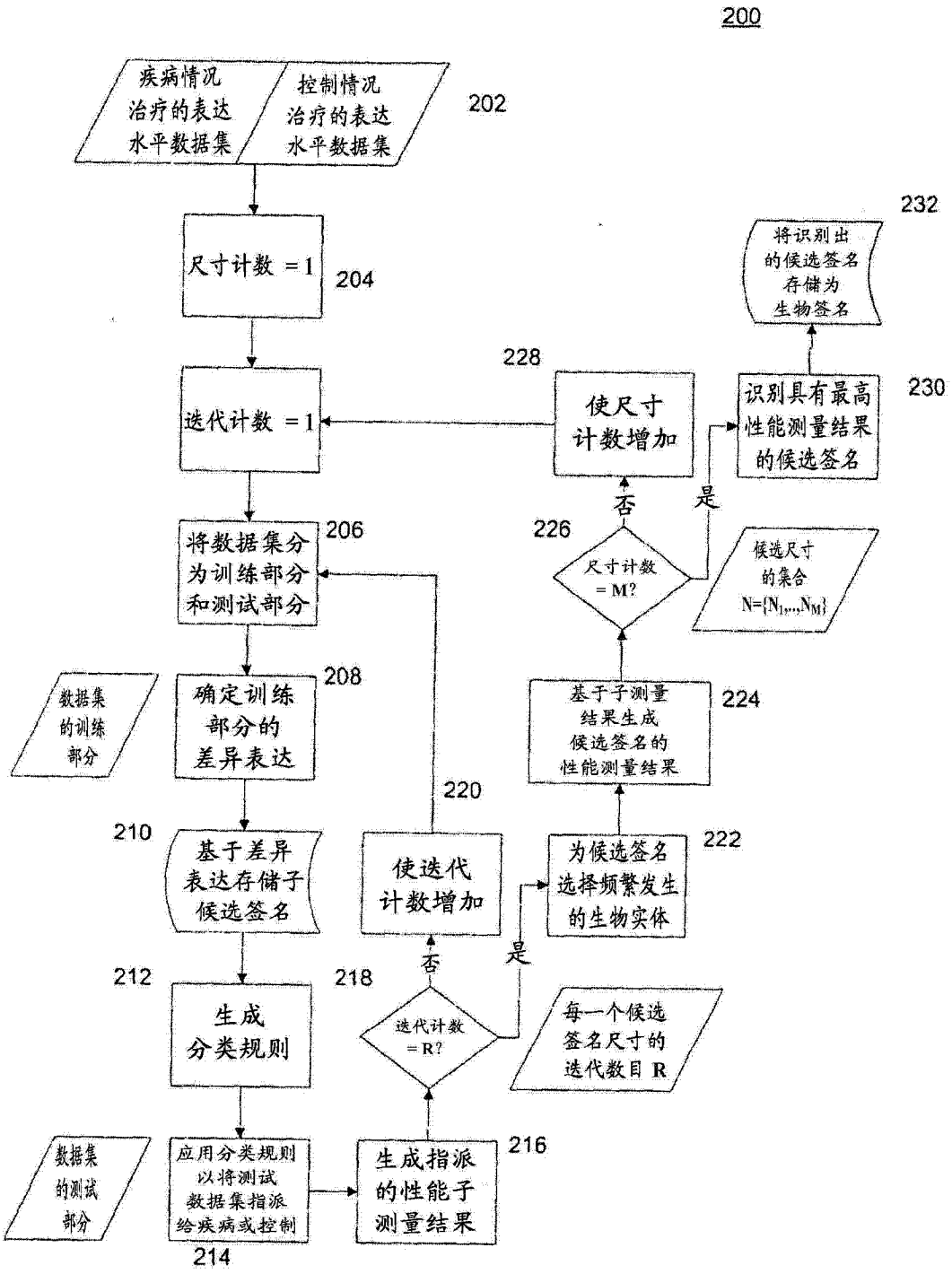

200
疾病情况治疗的表达水平数据集
控制情况治疗的表达水平数据集
202
尺寸计数 = 1
204
迭代计数 = 1
将数据集分为训练部分和测试部分
206
确定训练部分的差异表达
208
数据集的训练部分
基于差异表达存储子候选签名
210
生成分类规则
212
数据集的测试部分
应用分类规则以将测试数据集指派给疾病或控制
214
生成指派的性能子测量结果
216
迭代计数 = R?
218
否
是
使迭代计数增加
220
为候选签名选择频繁发生的生物实体
222
每一个候选签名尺寸的迭代数目 R
基于子测量结果生成候选签名的性能测量结果
224
尺寸计数 = M?
226
否
是
使尺寸计数增加
228
候选尺寸的集合 N={N1,..,NM}
识别具有最高性能测量结果的候选签名
230
将识别出的候选签名存储为生物签名
232

图 2

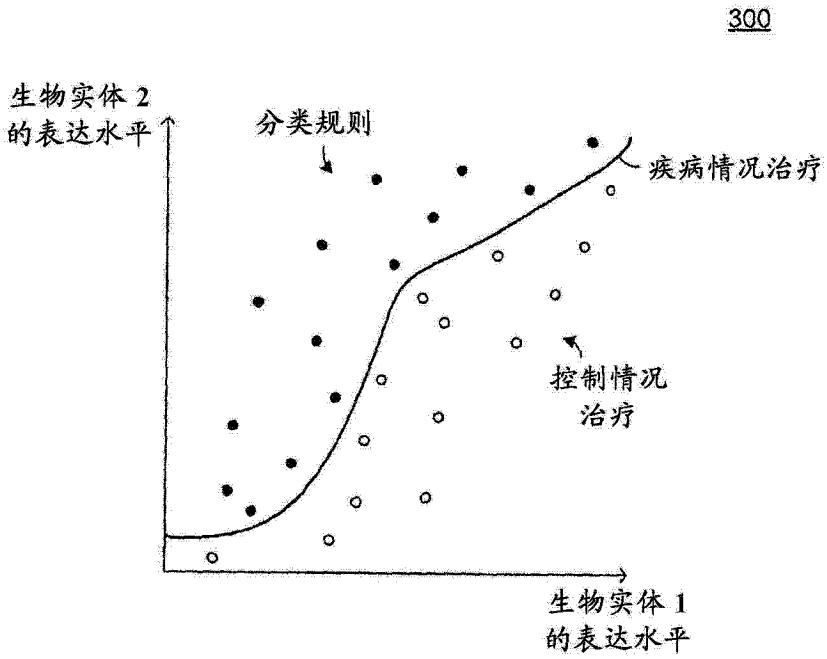

300
生物实体 2
的表达水平
分类规则
疾病情况治疗
控制情况
治疗
生物实体 1
的表达水平

图 3

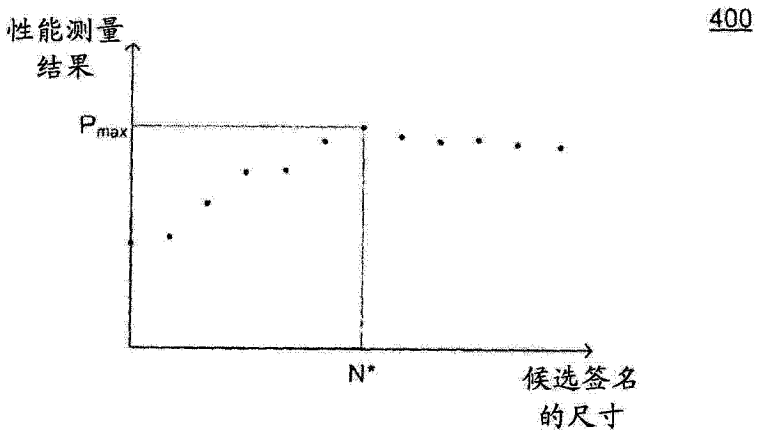

400
性能测量
结果
Pmax
N*
候选签名
的尺寸

图 4

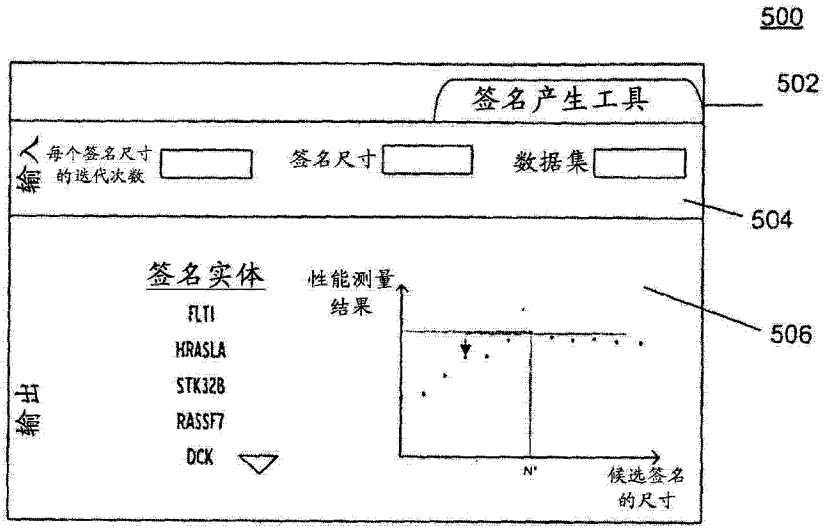

500
502
签名产生工具
输入
每个签名尺寸的迭代次数
签名尺寸
数据集
504
签名实体
FLT1
HRASLA
STK32B
RASSF7
DCK
输出
性能测量结果
506
N'
候选签名的尺寸

图 5

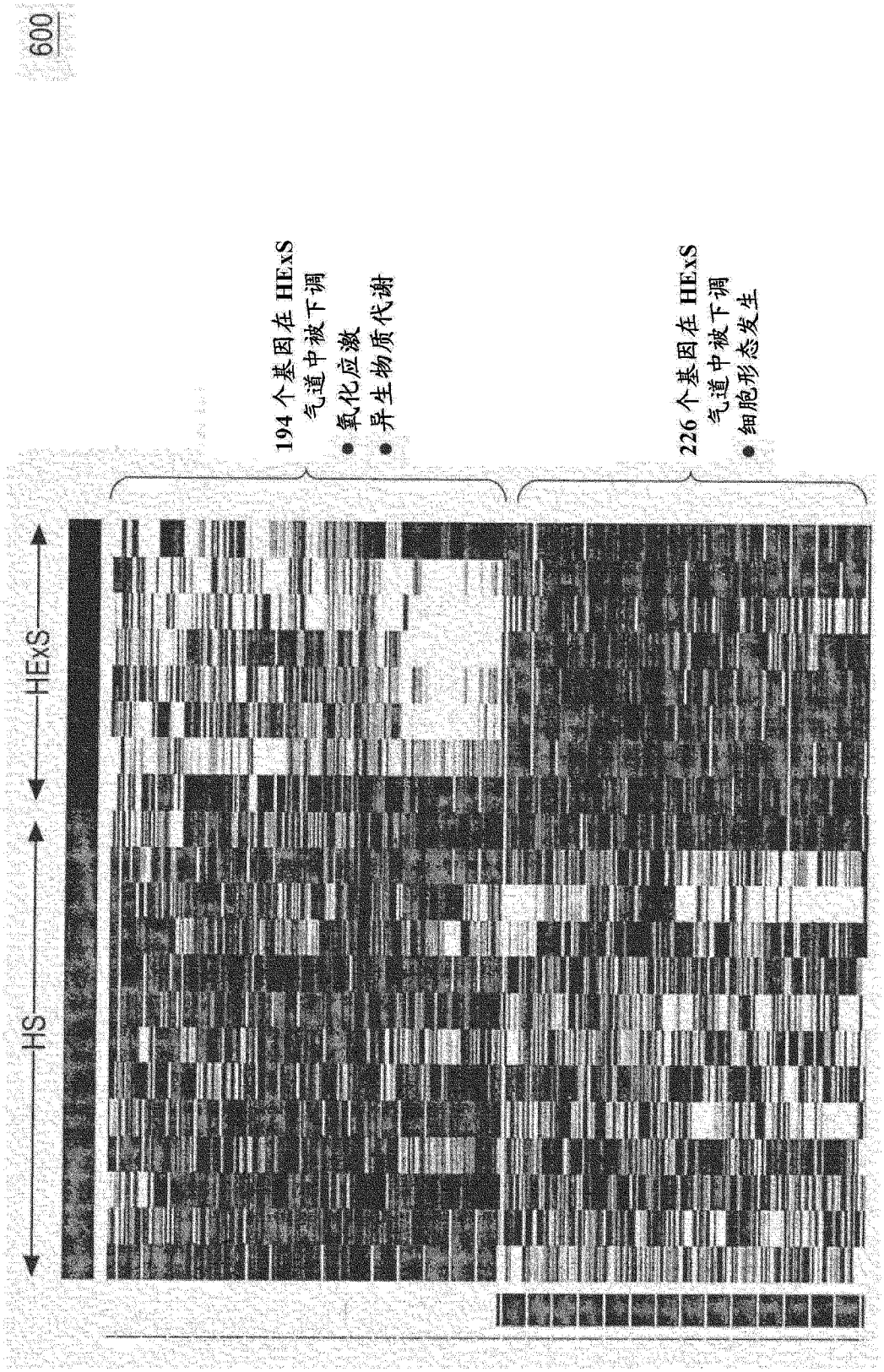

600
194 个基因在 HExS
气道中被下调
● 氧化应激
● 异生物质代谢
226 个基因在 HExS
气道中被下调
● 细胞形态发生
HExS
HS

图 6

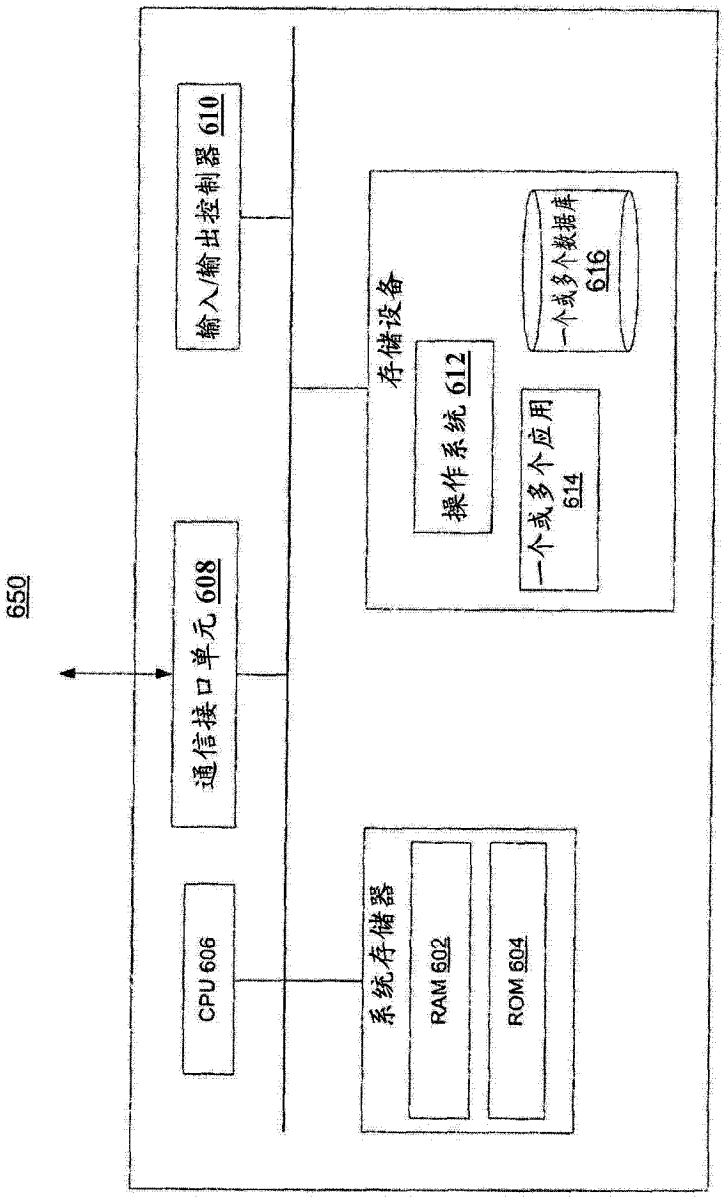

650
输入/输出控制器 610
通信接口单元 608
CPU 606
系统存储器
RAM 602
ROM 604
存储设备
操作系统 612
一个或多个应用 614
一个或多个数据库 616

图 7