(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 115280415 A
(43)申请公布日 2022.11.01

(21)申请号 202180019685.2
(22)申请日 2021.01.15
(30)优先权数据
2000649.0 2020.01.16 GB
2013386.4 2020.08.26 GB
2013387.2 2020.08.26 GB
(85)PCT国际申请进入国家阶段日
2022.08.30
(86)PCT国际申请的申请数据
PCT/GB2021/050086 2021.01.15
(87)PCT国际申请的公布数据
WO2021/144578 EN 2021.07.22
(71)申请人 康捷尼科有限公司
地址 英国剑桥郡
(72)发明人 S.莫加内拉 Y.达曼 L.庞廷 E·S·麦凯
(74)专利代理机构 北京市柳沈律师事务所 11105
专利代理师 张贵东
(51)Int.Cl.
G16B 20/20(2006.01)
G16B 40/20(2006.01)

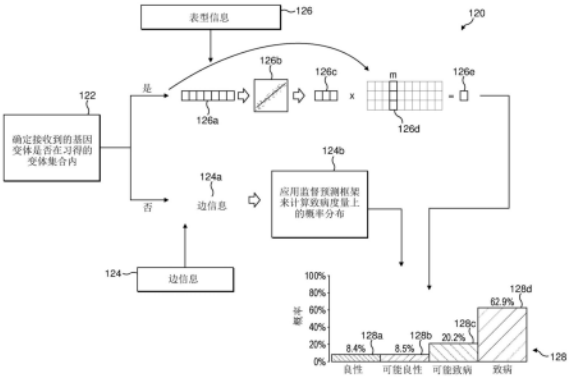
权利要求书4页 说明书18页 附图7页

(54)发明名称

致病性模型的应用和其训练

(57)摘要

提供了一种用于评估变体对患者的致病性的计算机实施的方法。接收变体。基于习得的变体集合，确定所述变体的与致病度量相关的至少一种概率。所述致病度量包括用于确定所述变体的至少一种概率的至少一个遗传病症簇的数据表示。输出所述患者的所述变体的至少一种概率的组合表示。



CN 115280415 A

1.一种用于评估变体对患者的致病性的计算机实施的方法,所述方法包括:

接收变体;

基于习得的变体集合确定所述变体的与致病度量相关的至少一种概率,其中所述致病度量包括用于确定所述变体的所述至少一种概率的至少一个遗传病症簇的数据表示;以及

输出所述患者的所述变体的所述至少一种概率的组合表示。

2.根据权利要求1所述的计算机实施的方法,其中所述至少一个遗传病症簇的所述数据表示是由所述习得的变体集合导出的,并且是关于患者的表型信息集进行加权的。

3.根据权利要求1或2所述的计算机实施的方法,其中所述变体被包含在所述习得的变体集合中,所述方法进一步包括:

接收所述患者的表型信息;

基于所述患者的所述表型信息确定与所述至少一个遗传病症簇中的每个遗传病症簇相关联的贡献;以及

基于根据所述至少一个遗传病症簇的所述数据表示确定的所述贡献来调整所述变体的所述至少一种概率。

4.根据权利要求2或3所述的计算机实施的方法,其进一步包括:

评估所述患者的所述表型信息的可用性;以及

基于所述可用性确定是否调整所述至少一个遗传病症簇以输出所述组合表示。

5.根据权利要求3或4所述的计算机实施的方法,其中基于所述患者的所述表型信息确定与所述至少一个遗传病症簇中的每个遗传病症簇相关联的贡献,进一步包括:

使用一个或多个回归模型对所述至少一个遗传病症簇中的每个遗传病症簇进行分割,其中所述一个或多个回归模型在给定所述患者的所述表型信息的情况下预测对所述至少一个遗传病症簇中的每个遗传病症簇的所述贡献。

6.根据权利要求1或2所述的计算机实施的方法,其中所述变体不包含在所述习得的变体集合中,所述方法进一步包括:

从所述习得的变体集合中标识与所述变体相关的至少一个邻近变体;

接收与所述至少一个邻近变体中的每个邻近变体相对应的边信息集,其中所述边信息集包括一个或多个指标;

基于所述边信息集标识最接近的变体;以及

在确定所述变体的与所述致病度量相关的所述至少一种概率时,应用所述最接近的变体作为所述变体。

7.根据权利要求6所述的计算机实施的方法,其中基于所述边信息集通过应用与所述至少一个邻近变体相关联的相似性度量来标识所述最接近的变体;和/或其中关于所述边信息集对所述相似性度量进行加权。

8.根据权利要求7所述的计算机实施的方法,当所述相似性度量从所述习得的变体集合中标识出至少一个其它变体具有等效相似性评分时,通过对所述至少一个邻近变体中的每个邻近变体求平均来确定所述变体的所述至少一种概率。

9.一种用于生成至少一个遗传病症簇的计算机实施的方法,所述至少一个遗传病症簇用于确定变体的与致病度量相关的至少一种概率,所述方法包括:

接收至少一个患者的与变体集合相关联的带注释的数据,其中所述带注释的数据包括

解释信息以及与所述致病度量相对应的相关观察结果；

确定至少一个患者的所述带注释的数据的数据表示，其中所述数据表示是使用一种或多种生成模型导出的；以及

基于所述数据表示生成所述至少一个遗传病症簇。

10.根据权利要求9所述的计算机实施的方法，其中所述带注释的数据进一步包括患者的表型信息集和/或边信息集。

11.根据权利要求10所述的计算机实施的方法，其中所述表型信息集与和所述至少一个患者相关的所述解释信息相关联；和/或其中所述边信息集与和所述变体集合相关的所述解释信息相关联。

12.根据权利要求10或11所述的计算机实施的方法，其进一步包括：

基于所述表型信息集调整与所述至少一个遗传病症簇相关联的权重集，其中所述权重集与所述至少一个遗传病症簇对所述表型信息集的贡献相对应；以及

基于经调整的权重集将一个或多个回归模型配置成确定与所述致病度量相关的所述贡献。

13.根据权利要求10到12所述的计算机实施的方法，其中所述边信息集包括与所述变体集合相关联的指标的数据表示。

14.根据权利要求10到13所述的计算机实施的方法，其中在所述变体未被包含在所述变体集合中时，应用所述边信息集以从所述变体集合中标识用于确定所述变体的所述至少一种概率的最接近的变体；和/或其中使用提供所述边信息集的监督学习框架确定所述变体的所述至少一种概率。

15.根据权利要求14所述的计算机实施的方法，其中所述变体被包含在所述变体集合中，以通过应用与所述最接近的变体相关联的注释来更新所述最少一个遗传病症簇。

16.根据权利要求9到15所述的计算机实施的方法，其进一步包括：

基于所述带注释的数据确定所述至少一个遗传病症簇的最优集；以及

在预测期间应用所述至少一个遗传病症簇的所述最优集来确定变体的与所述致病度量相关的所述至少一种概率。

17.根据权利要求16所述的计算机实施的方法，其中所述至少一个遗传病症簇的所述最优集被配置成用新的带注释的数据迭代地更新。

18.一种用于使用边信息集评估未知变体对患者的致病性的计算机实施的方法，所述方法包括：

接收所述未知变体，其中所述未知变体未在所述习得的变体集合中标识出；

使用与所述习得的变体集合的每个子集相对应的所述边信息集来训练监督学习框架；以及

基于经训练的监督学习框架来评估所述未知变体的所述致病性。

19.根据权利要求18所述的计算机实施的方法，其进一步包括：比较与所述习得的变体集合的每个子集相对应的所述边信息集，其中与所述习得的变体集合的每个子集相对应的所述边信息集是关于与所述习得的变体集合的所述子集相关联的相似性评分比较的。

20.根据权利要求18或19所述的计算机实施的方法，其进一步包括：关于最接近的变体的所述致病性评估所述未知变体的所述致病性，所述评估进一步包括：

基于习得的变体集合确定所述最接近的变体的与致病度量相关的至少一种概率，其中所述致病度量包括用于计算所述最接近的变体的所述至少一种概率的至少一个遗传病症簇的数据表示；以及

生成所述至少一种概率的组合表示，其中所述组合表示是相对于所述致病度量输出的。

21.根据权利要求20所述的计算机实施的方法，其进一步包括：

响应于所述习得的变体集合的所述子集包括两个或更多个具有等效相似性评分的变体使得无法确定所述最接近的变体而通过对所述习得的变体集合的子集的每个变体的所述至少一种概率求平均来生成所述组合表示；和/或

在给定所述边信息集的情况下基于所述习得的变体集合的子集的每个变体的至少一种概率使用所述监督学习框架来生成所述组合表示，其中所述监督学习框架包括一个或多个监督预测模型。

22.根据前述权利要求1到8和10到17中任一项所述的计算机实施的方法，其中所述表型信息包括与一种或多种疾病相关联的表型本体。

23.根据前述权利要求9到17中任一项所述的计算机实施的方法，其中所述一种或多种生成模型被配置成分解与所述致病度量相关的带注释的数据的所述数据表示。

24.根据前述权利要求9到17、22和23中任一项所述的计算机实施的方法，其中所述一种或多种生成模型包括至少一个基于矩阵因式分解算法的公式。

25.根据前述权利要求1到17和20到24中任一项所述的计算机实施的方法，其中所述致病度量包括指示致病性程度的至少一种分类。

26.根据权利要求25所述的计算机实施的方法，其中所述至少一种分类中的每种分类与所述至少一个遗传病症簇的不同最优集相关联。

27.一种计算机可读介质，其包括存储在其上的计算机可读代码或指令，所述计算机可读代码或指令当在处理器上执行时，使所述处理器实施根据前述权利要求中任一项所述的计算机实施的方法。

28.一种系统，其包括至少一个电路系统，所述至少一个电路系统被配置成执行根据权利要求1到26中任一项所述的计算机实施的方法。

29.一种设备，其包括处理器、存储器和通信接口，所述处理器连接到所述存储器和所述通信接口，其中所述设备适于或被配置成实施根据权利要求1到26中任一项所述的计算机实施的方法。

30.一种用于确定变体对患者的致病性的设备，所述设备包括：

输入组件，所述输入组件被配置成接收所述变体；

处理组件，所述处理组件被配置成确定所述变体是否在习得的变体集合内；

预测组件，响应于确定所述变体存在于所述习得的变体集合中，所述预测组件被配置成生成所述变体的与致病度量相关的至少一种概率，其中所述致病度量包括用于确定所述变体的所述至少一种概率的至少一个遗传病症簇的数据表示；以及

显示组件，所述显示组件被配置成显示所述变体的关于所述致病度量的所述至少一种概率，其中所述至少一种概率被归一化。

31.根据权利要求30所述的设备，其中响应于确定所述变体不存在于所述习得的变体

集合中，所述预测组件被配置成接收边信息集，其中所述边信息用于关于所述变体标识最接近的变体，所述最接近的变体作为所述变体被应用以生成所述至少一种概率。

32.根据权利要求30所述的设备，其中所述输入组件被配置成接收与所述患者相关联的表型信息，其中所述表型信息用于调整所述变体的与所述至少一个遗传病症簇相关的所述至少一种概率。

33.一种用于使用边信息集来确定未知基因变体的致病性概率分布的计算机实施的方法，所述方法包括：

接收患者的所述未知变体，其中所述未知变体未在与多个患者相关联的习得的变体集合中标识出或者对所述习得的变体集合来说是新的；

基于所述边信息集通过使用监督学习框架来评估所述未知基因变体的所述致病性；以及

基于所述评估确定所述致病性概率分布。

34.根据权利要求33所述的计算机实施的方法，其进一步包括：

在给定所述边信息集的情况下，计算所述未知变体的与致病度量集相关联的概率。

35.根据权利要求33或34所述的计算机实施的方法，其进一步包括：

基于习得的变体集合确定所述未知变体的与致病度量相关的至少一种概率；以及

生成所述至少一种概率的组合表示，其中所述组合表示是相对于所述致病度量输出的。

36.根据权利要求33到35所述的计算机实施的方法，其中所述监督学习框架包括一种或多种预测模型。

37.根据权利要求33到35所述的计算机实施的方法，其中所述监督学习框架包括非参数分类器。

38.根据权利要求33到37所述的计算机实施的方法，其中所述边信息集与所述未知基因变体相关联。

39.根据权利要求33到38所述的计算机实施的方法，其中所述方法在关于根据权利要求27到32中任一项所述的计算机可读介质、系统或设备的处理器上实施。

致病性模型的应用和其训练

[0001] 本申请涉及一种用于评估变体对患者的致病性的系统、设备和方法，以及对用于评估所述系统、所述设备和所述方法的模型进行的训练。

背景技术

[0002] 医学和计算技术的进步实现了基于表型属性对生物样品的基因组测序进行分析。用于基于这些属性来预测致病DNA突变的基因组分析一直是研究与开发的热门领域。由于基因组数据的固有复杂性以及存在大量噪声，这些预测仍然存在很多不确定性。例如，尽管测序过程期间存在噪声，但这种复杂性可能归因于范围从单核苷酸变体(SNV)到大且复杂的重排的突变。对这些突变的预测的不确定性对现有技术或计算工具提出了挑战，现有技术或计算工具效率低下且不准确，特别是对于分析特定变体或突变。

[0003] 尽管如此，已经开发出几种计算工具以进行基因组数据分析和解释，以获得对遗传变体的见解。然而，这些工具需要使用大量经标记和/或未经标记的训练数据对其基础模型进行大量训练以运行嵌入式机器学习算法，所述嵌入式机器学习算法具有长度运行时并且由此是资源密集型的。例如，常规机器学习或人工智能模型在将与受试者的先前输入相关的新输入馈送到此类模型中时会经历完整再训练，在诊断测试结果以及与受试者相关的其它信息通常不容易获得，并且通常仅在进行诊断测试并且与患者相关的额外数据可用时才能获得的情况下，这是不期望的。因此，在这种情况下再训练常规模型不仅会在与受试者相关的基因组数据的评估中产生时间滞后，而且还会增加基因组解释的不确定性，有与误释相关联的风险。在上面的实例中，在给定患者的血液样品进行测序与几年后可能发现新的相关科学信息之间可能会存在时间滞后；新的相关科学信息关注在进行表达时特定基因会做什么。由于时间滞后，给定患者的医疗记录可能会被潜在地标记为“未解决”，并且给定患者的记录之后在更多信息变得可用时也不能被再访问。

[0004] 因此，鉴于前述讨论，需要克服前面提到的与用于处理、分析或解释基因组数据的常规方法相关的缺点，以减少噪声的影响并防止过拟合。更具体地，需要一种用于处理拷贝量的固有地复杂的复杂基因组数据的过程，以便根据变体的致病性准确评估患者的生物序列中的变体或突变。

[0005] 下文所描述的实施例不限于解决上文所描述的已知方法的任何或全部缺点的实施方式。

发明内容

[0006] 提供本发明内容的目的是以简化形式介绍下文在具体实施方式中进一步描述的一系列概念。本发明内容不旨在标识要求保护的主题的关键特征或必要特征，也不旨在用于确定要求保护的主题的范围；促进本发明的工作和/或用于实现基本上类似的技术效果的各种变型和替代性特征应被视为落入本文公开的本发明的范围中。

[0007] 本公开提供了一种算法框架，所述算法框架使得能够在给定患者的基因组谱和特定表型属性的情况下标识致病DNA突变。

[0008] 在第一方面，本公开提供了一种用于评估变体对患者的致病性的计算机实施的方法，所述方法包括：接收变体；基于习得的变体集合确定所述变体的与致病度量相关的至少一种概率，其中所述致病度量包括用于确定所述变体的所述至少一种概率的至少一个遗传病症簇的数据表示；以及输出所述患者的所述变体的所述至少一种概率的组合表示。

[0009] 在第二方面，本公开提供了一种用于生成至少一个遗传病症簇的计算机实施的方法，所述至少一个遗传病症簇用于确定变体的与致病度量相关的至少一种概率，所述方法包括：接收至少一个患者的与变体集合相关联的带注释的数据，其中所述带注释的数据包括解释信息以及与所述致病度量相对应的相关观察结果；确定至少一个患者的所述带注释的数据的数据表示，其中所述数据表示是使用一种或多种生成模型导出的；以及基于所述数据表示生成所述至少一个遗传病症簇。

[0010] 在第三方面，本公开提供了一种用于使用边信息集来评估未知变体对患者的致病性的计算机实施的方法，所述方法包括：接收所述未知变体，其中所述未知变体未在习得的变体集合中标识出；使用与所述习得的变体集合的每个子集相对应的所述边信息集来训练监督学习框架；以及基于经训练的监督学习框架来评估所述未知变体的所述致病性。

[0011] 在第四方面，本公开提供了一种用于确定变体对患者的致病性的设备，所述设备包括：输入组件，所述输入组件被配置成接收所述变体；处理组件，所述处理组件被配置成确定所述变体是否在习得的变体集合内；预测组件，响应于确定所述变体存在于所述习得的变体集合中，所述预测组件被配置成生成所述变体的与致病度量相关的至少一种概率，其中所述致病度量包括用于确定所述变体的所述至少一种概率的至少一个遗传病症簇的数据表示；以及显示组件，所述显示组件被配置成显示所述变体的关于所述致病度量的所述至少一种概率，其中所述至少一种概率被归一化。

[0012] 在第五方面，本公开提供了一种用于使用边信息集来确定未知基因变体的致病性概率分布的计算机实施的方法，所述方法包括：接收患者的所述未知变体，其中所述未知变体未在与多个患者相关联的习得的变体集合中标识出或者对所述习得的变体集合来说是新的；基于所述边信息集通过使用监督学习框架来评估所述未知基因变体的所述致病性；以及基于所述评估确定所述致病性概率分布。

[0013] 本文中所描述的方法可以由呈机器可读形式的软件在例如呈计算机程序的形式的有形或非暂时性存储介质上执行，所述计算机程序包括计算机程序代码装置，所述计算机程序代码装置适于在所述程序在计算机上运行时以及在所述计算机程序可以在计算机可读介质上体现的情况下执行本文中所描述的方法中的任何方法的所有步骤。有形(或非暂时性)存储介质的实例包含：磁盘、拇指驱动器、存储卡等，并且不包含传播的信号。所述软件可以适于在并行处理器或串行处理器上执行，使得方法步骤可以以任何合适的顺序或同时执行。

[0014] 本申请承认固件和软件可以是有价值的、可单独交易的商品。旨在涵盖在“哑”或标准硬件上运行或控制“哑”或标准硬件的软件，以实现期望的功能。还旨在涵盖“描述”或定义硬件的配置的软件，如HDL(硬件描述语言)软件，如用于设计硅芯片或用于配置通用可配置芯片，以实现期望的功能。

[0015] 如对技术人员显而易见的，优选特征可以适当地结合，并且可以与本发明的任何方面结合。

附图说明

[0016] 将通过实例的方式并参考附图来描述本发明的实施例,在附图中:

[0017] 图1a是展示了根据本发明的评估变体对患者的致病性的实例的流程图;

[0018] 图1b是展示了根据本发明的其中关于表型信息和边信息评估变体对患者的致病性的实例的示意图;

[0019] 图2a是展示了根据本发明的生成用于确定变体的与致病度量相关的至少一种概率的遗传病症簇的实例的流程图;

[0020] 图2b是根据本发明的用于确定变体的概率的遗传病症簇的实例的示意图;

[0021] 图3是展示了根据本发明的使用边信息集评估未知变体对患者的致病性的实例的流程图;

[0022] 图4是展示了根据本发明的从带注释的数据中提取以在给定致病度量的情况下预测变体的概率的遗传病症簇的实例的示意图。

[0023] 图5是适于实施本发明的实施例的计算机系统的示意图。

[0024] 所有附图均使用相同的附图标记来表示相似的特征。

具体实施方式

[0025] 下文仅通过实例的方式描述本发明的实施例。这些实例表示申请人目前已知的实践本发明的最佳模式,但是这些实例并不是可以实现本发明的唯一方法。描述阐述了实例的功能以及构造和操作实例的步骤的序列。然而,可以通过不同的实例实现相同或等同的功能和序列。

[0026] 发明人提出了一种用于评估或预测特定变体(例如基因变体)对所关注患者的致病性的过程。所述过程利用至少一种预测模型,所述至少一种预测模型使用表型信息和/或解释信息的注释训练数据进行训练,所述注释训练数据被编译为导出潜变量集,以便进行合适的评估或预测。进而,潜变量集可以被感知为(隐藏)遗传病症簇的数据表示。遗传病症簇被改编为基于模型习得的变体集合来确定变体的一组概率。在致病度量方面评估概率,其中每个度量归属于一个确定的概率。所述一组概率的组合表示是通过计算接口或设备输出给用户的。因此,输入变体是否是致病性的(例如,良性的还是致病性的)可能性或其致病性可以根据输出的概率来确定或考虑。

[0027] 此过程可以迭代,并且预测模型可以随着表型信息和/或解释信息的更多输入的流入而继续递增。表型信息和/或解释信息包括与患者相关联的数据点、变体和来自过去的患者解释的体现为多维数据矩阵的对应观察结果。数据点相对于矩阵的大小可以是高度稀疏的,因为数据矩阵的观察结果是大约99.96%不存在。这至少是由于变体池的大小和与每个变体相关联的观察结果有限可用性引起的。然而,本文中描述为方法的过程、系统、介质或设备至少提出了用于通过应用遗传病症簇来克服数据稀疏性困境的解决方案。实际上,遗传病症簇抽象地将变体映射到其潜在致病性,在某种程度上是解决本文所述的其它技术问题之中的数据稀疏的客观问题。

[0028] 本文的致病性是指引起特定疾病的性质。变体的致病性是变体在引起疾病方面的能力。变体的致病性是对所述变体以及所述变体和对疾病的起因的贡献的可能性的定性和定量评估。变体为致病性的可能性可以表示为概率。这些概率与变体相关联,并且提供对变

体在致病性方面的定量评估。

[0029] 变体是遗传(DNA)序列和其转录物(RNA)的突变,其包含基因变体或其它序列突变。具体地,基因变体是指单核苷酸多态性(SNP)、拷贝数变体(CNV)、基因重排和插入缺失(indel)等。一般来说,具有变体的患者可能具有由疾病引起的病状或疾病,在某种程度上患者承继了SNP或基因组DNA中的突变。此患者可以具有一种或多种变体,所述一种或多种变体包含但不限于:例如,拷贝数变体(CNV)、插入缺失、单核苷酸变体(SNV)和负责遗传疾病的其它突变。因此,在基因筛查的背景下,变体是在基因组DNA方面健康个体与患者之间的任何差异。

[0030] 例如,基因"X"可以具有两种变体:"A"和"B"。"A"变体和"B"变体两者位于基因"X"的不同基因座处,并且负责疾病"D"。假设某种DNA突变(例如在预期的"A"核苷酸被"C"核苷酸替代的情况下)在存在于基因的特定编码区中时使得此基因为潜在地致病性的,则在变体"A"的基因座处存在这段DNA可以容易地将变体"A"与新患者的疾病"D"相关联,这与变体"B"相反,所述变体B未展示出相同的DNA序列。与基因"X"和其与疾病"D"的对应关系相关联的变体可以根据以下部分中所述的模型改编,并且被改编为本文所述的方法、系统、介质或设备的习得的变体。

[0031] 进一步地,发现基因的某个示例段(例如"AAAAATAAAAAT")在作为变体存在于基因的特定编码区(例如"AA"到"CC")处时使得所述基因为潜在地致病性的(换句话说,重复元素"AACCAT"可能使患者表现出疾病。因此,如果基因"X"的任何其它接近变体(即除了变体"A"和"B"之外)具有同一段基因(例如,AAAAATAAAAAT),则所述变体可以容易地与新患者的疾病"D"相关联。与基因"X"相关联的变体可以是本文所述的方法、系统、介质或设备的习得的变体之一。

[0032] 变体的其它实例可以包含但不限于:转录消融、剪接供体变体、剪接受体变体、终止密码子获得(stop gained)、移码变体、起始密码子丢失(start lost)、起始密码子变体、转录扩增、框内插入、框内缺失、错义变体、改变蛋白质的变体、剪接区变体、不完全末端密码子变体、同义变体、编码序列变体、成熟miRNA变体、5主要UTR变体、3主要UTR变体、非编码转录变体、内含子变体、上游变体、下游变体、转录因子(TF)结合位点变体、调节区消融、转录因子结合位点(TFBS)消融等。

[0033] 习得的变体或其集合是指通过计算模型感知到的或习得的变体。换言之,习得的变体集合包括模型已看作或视为已知或模型已对其进行训练的变体或变体序列。因此,具有带注释的变体或带注释的数据的经训练的模型包含构成每个变体的解释信息(所述解释信息被量化并且用于基于患者和变体的注释来做出致病性的决定)的基础的习得的变体的数据表示,其中注释表明与每个变体相关联的特定观察结果,所述观察结果用于评估变体是表型致病性的(即引起给定病状/疾病)还是良性的(即,无害的)或在致病度量集的背景下致病性的程度。更具体地,注释提供了用于在给定模型的情况下评估变体为致病性的可能性的基础。所述可能性可以通过与表现出的表型相关的概率或概率分布来呈现。

[0034] 由此,以上所述的计算模型被配置成基于致病度量集来评估任何变体,其中致病度量由此是通过已知的或之后作为习得的变体集合的带注释的变体来进行训练。致病度量提供了一种分类方案,其中可以关于致病性的程度对变体进行表型分类。这些类别的实例包含但不限于:B(良性)、LB(可能良性)、LP(可能致病)和P(致病)等。类别中的每种类别都

有确定指示概率的可能性。这样,计算模型可以是被配置成学习训练集的数据分布,以便于通过在输出概率方面的一些变化来生成另外的数据点或预测的生成模型。

[0035] 已知变体或任何变体序列可以从各种数据源获得,所述各种数据来源包含但不限于,例如,基因组数据库、公共科学数据库、研究组织的数据库(例如基因组变体数据库(DGV))、在线人类孟德尔遗传(Online Mendelian Inheritance in Man,OMIM)、MORBID、DECIPHER、研究文献(例如PubMed文献)和其它支持信息等。

[0036] 例如,在OMIM的情况下,向变体指定了基因名称(例如“BICD2”基因)和OMIM标识符(ID)(例如“609797”)。OMIM可以包含关于约15,000个基因的已知孟德尔病的可公开获得的信息,所述信息是定期更新的并且含有表型与基因型之间的关系。也可以指定“MORBID ID”(例如615290)。“MORBID ID”表明疾病以及疾病与其相关联的基因的染色体位置的图表或图示。OMIM知识库中提供了病态地图,其列出了染色体和映射到那些染色体上的特定位点的基因。进一步地,还可以注释与基因(例如,BICD2)基因相关联的已知病状(例如症状:具有常染色体显性遗传的近端脊髓性肌萎缩)。这些对变体的注释用作用于训练模型的基础。

[0037] 在模型的训练中,可以使用带注释的变体以导出或生成本文中创造为遗传病症簇的潜参数。这些遗传病症簇捕获致病类别的抽象概念,其中可以基于致病度量确定对所关注基因的评估。更具体地,遗传病症簇提供了抽象映射,其中特定变体可以与以下表型类别中的每种类别相关:致病度量的B(良性)、LB(可能良性)、LP(可能致病)和P(致病)。总之,遗传病症簇允许对给定变体的致病性的某种概率进行预测。

[0038] 可以使用各种计算技术来导出这些遗传病症簇。这些计算技术可以包含如本文所述的一种或多种机器学习(ML)技术。这些技术还可以包含一种或多种矩阵因式分解算法,所述矩阵分解算法可以应用于协同过滤和推荐系统应用,其中目的是通过使用潜参数对关系数据进行建模。这些合适的方法的实例包含但不限于:潜在狄利克雷分配(Latent Dirichlet Allocation)、非负矩阵因式分解、贝叶斯和非贝叶斯概率矩阵因式分解(Bayesian and non-Bayesian Probabilistic Matrix Factorization)、主成分分析、神经网络矩阵因式分解等。

[0039] 在应用遗传病症簇时,可以评估表型类别(即良性)的证据或度量,以生成与特定类别相关联的概率。所述模型可以输出与患者的所关注变体的表型类别相关联的概率中的每种概率的组合表示。这种组合表示可以呈如图1b所示的直方图的形式或适于显示组合中的模型的所产生的概率的其它图形表示的形式。

[0040] 遗传病症簇通过表型信息集加权以通过调整对相关联的表型的某种贡献来对模型进行微调,而与患者相关联的表型信息的额外输入可以基于表型信息集返回更准确的预测。具体地,表型信息集可以是包括患者队列的表型数据的矩阵,例如,人类表型本体(HPO)术语或来自可用数据源的表型的其它编码。分配了表型数据,这提供了用于表示人类疾病中遇到的表型异常的标准化方式。在HPO术语的情况下,如果基因序列(例如BICD2)先前被报道为致病性的并且是习得的变体集合的一部分,则可以自动检索所述术语。例如,HPO术语包含:“HP:0000347‘小颌畸形’、HP:0001561‘羊水过多’、HP:0001989‘胎儿运动不能序列’、HP:0001790‘非免疫性胎儿水肿′、HP:0002803‘先天性挛缩’”。这些HPO术语在基于致病度量的预测期间与遗传病症簇组合使用。更具体地,HPO术语或更普遍地表型数据被用于训练与遗传病症簇中的每个遗传病症簇相关联的权重。训练是使用本文描述的一种或多种

ML技术或通过曲线拟合算法完成的,所述曲线拟合算法包含但不限于使用具有不同惩罚术语的线性回归(即LASSO、岭(RIDGE)、弹性网络(Elastic Net)等)。

[0041] 除了表型信息之外,可以引入边信息集以表征未知基因变体,即不是习得的变体集合的一部分的变体的致病性。边信息集或边信息可以指与本文所述的一个或多个基因变体相关联的指标。

[0042] 具体来说,边信息集与模型习得的一个或多个已知变体有关。边信息的实例包含各种表型指标和基因型指标。这些指标包含但不限于GERP评分(定义多物种序列比对中取代的数量相比于中性预期的数量的减少)、SIFT评分(预测氨基酸取代是否影响蛋白质功能)、变体效应预测器(VEP)结果(变体的坐标和与其效应相关联的核苷酸变化),MVP评分(通过深度学习ML模型预测错义变体的致病性)。可替代地,也可以使用HI评分和ADA评分。例如,可以将HI评分(例如0.176)指定给具有对接合性的指示的基因的变体以及为已知变体注释的VEP结果。

[0043] 对未知基因变体的致病性的预测可以通过使用监督学习框架来进行。在给定未知基因变体和其边信息的情况下,作为框架的基础的预测模型被配置成生成每个致病度量(例如,良性、可能良性、可能致病和致病等)的概率。也就是说,至少一个模型(M)在给定其边信息(SI)的情况下或者在M=P(Vm|SI)时计算变体与那些致病度量(Vm)中的每个致病度量相关联的概率。

[0044] 可以通过使用作为独立变量的边信息和致病度量(例如,良性、可能良性、可能致病和致病等)来训练监督学习框架或基础预测模型中的任何基础预测模型。监督学习框架可以包含非参数分类器。所述框架还可以包含但不限于线性回归、逻辑回归、神经网络、支持向量机(SVM)等。这些模型将生成针对不同边信息的可以用于解释预测的不同权重(例如,GERP评分可以具有相比于SIFT评分更高的权重,并且这在计算致病性时,将使GERP评分具有相比于SIFT评分更显著的影响)。

[0045] 可以使用机器学习(ML)技术来生成经训练的模型,如但不限于,例如,基于被称为与表型信息和解释信息相关联的训练数据的输入数据的一个或多个生成式ML模型或分类器。所述输入数据还可以包含本文所述的边信息。利用如生物信息学等领域中的正确注释的训练数据集,可以使用技术来生成另外经训练的ML模型、分类器和/或生成模型,以用于下游过程,如举例来说但不限于药物发现、标识和优化以及信息学和/或生物信息学领域中的其它相关生物医学产物、治疗、分析和/或建模等。

[0046] 用于生成如本文所述的本发明可以使用的经训练的模型的示例ML技术可以包含或可以基于举例来说但不限于以下中的一项或多项:可以用于生成经训练的模型的任何ML技术或算法/方法;一种或多种监督ML技术;半监督ML技术;无监督ML技术;线性和/或非线性ML技术;与分类相关联的ML技术;与回归相关联的ML技术等和/或其组合。ML技术/模型结构的一些实例可以包含或可以基于举例来说但不限于以下中的一项或多项:主动学习、多任务学习、迁移学习、神经消息解析、一次性学习、降维、决策树、相关联规则学习、相似性学习、数据挖掘算法/方法、人工神经网络(NN)、自动编码器/解码器结构、深度NN、深度学习、深度学习ANN、归纳逻辑编程、支持向量机(SVM)、稀疏字典学习、簇、贝叶斯网络、强化学习、表示学习、相似性和度量学习、稀疏字典学习、遗传算法、基于规则的机器学习、学习分类器系统和/或其一个或多个组合等。

[0047] 训练数据或带注释的数据的类型包含但不限于与患者ID、患者表型、变体ID、致病度量和边信息相关联的数据集。患者ID可以是每个患者的唯一标识符，并且在图2b的矩阵222a和222b中显示为行ID。患者表型是观察到的患者的表型，并且可以呈现为人类表型本体(HPO)术语。HPO术语的一个实例为:HP:0000729，用于具有自闭症行为表型的患者；并且另一实例为HP:000986，用于具有肢体发育不良表型的患者。在图2b的二进制矩阵222a中，HPO术语显示为列ID。每个变体的变体ID可以是唯一的。变体ID可以呈现通过下划线连接和分隔的特征。例如，变体ID 2_1765342_C_T_NM_00193456唯一地标识染色体2上在碱基对定位1765342处开始涉及转录物NM_00193456上的突变C＞T的变体。在此，变体ID 2_1765342_C_T_NM_00193456标识了染色体、起始、参考等位基因、替代性等位基因和转录物ID等。变体ID在图2b的矩阵222b和222c中被示出为列ID。致病度量可以由如美国医学遗传学学院(American College of Medical Genetics)指定的变体致病性的水平来表示。例如，可以存在良性的致病度量B、可能良性的LB、可能致病的LP，致病的P，以及意义不确定的VUS。这些指标可以是，例如，被改编为矩阵因式分解算法和图2b的矩阵222b中示出的条目的替代性训练标签。边信息可以呈现为余弦相似性中使用的变体的注释，或者可以以监督学习框架中使用的任何合适的格式来组织。其在图2b的矩阵222c中被示出为列ID。

[0048] 训练数据或带注释的数据用于训练致病性模型，以评估和计算基因变体的概率分布，以评估变体对患者的致病性。具体地，训练数据或带注释的数据可以以计算机可读格式组织，所述计算机可读格式包含但不限于适用于用本文所述的一种或多种模型、框架、算法、技术和方法进行处理的实数、二进制、分类、标识符、列表和字符串格式。

[0049] 训练数据或与训练数据的类型相关的带注释的数据的实际实例在以下表1中示出。所述表还示出了与给定变体的边信息相关联的特征。例如，一个特征可以是患者的最大等位基因频率；另一个特征可以是同一患者的功能性蛋白质结构域中的非同义氨基酸变化。每个特征(特征1到11的每个特征)在表中呈现为与患者ID、患者表型、变体ID和致病度量相关。所述特征还可以与上述表型指标和基因型指标相对应，所述指标包含但不限于GERP评分、SIFT评分、变体效应预测器(VEP)结果、MVP评分等。训练数据的其它呈现包含表1中的实例但不限于此实例。训练数据可以被呈现和组织为与所应用的模型、框架、算法、技术或方法相关。训练数据可以被呈现为随着输入来调节以如本文所述训练致病性模型。

[0050] 表1

[0051]

患者ID	患者表型	变体ID	致病度量	特征1	特征2	特征3	特征4	特征5	特征6	特征7	特征8	特征9	特征10	特征11
1	HP:000164	7_1506460	B	0	3.95			姨妈_变体			0.697			0
1	HP:000164	11_768348	LB	0.005277	-0.163			错义_	0.002	0.64	0.208	5		0
1	HP:000164	16_579939	P	0.000124	-1.5	0.03	0.001013	剪接_区_变体			0.68			1
2	HP:000047	12_485164	VUS	0.218986	4.38	0.036	0.004091	内含子_变体			0.21			1
3	HP:000070	8_1007791	B	0.008287	-2.49			同义_变体			0.277		可能良性	0
3	HP:000070	8_5553922	LP	0	4.2			姨妈_变体			0.298			0
3	HP:000070	10_897208	P	0	4.39			终止密码子_获得					致病	0

[0052]

4	HP:000124	9_1194602	B	0	4.43	0.67	0.12	同义_变体		0.192			0
5	HP:000047	3_3865144	B	0.006742	0.209	0.001	0.23	同义_变体		0.242		可能良性	0
5	HP:000047	6_4268955	P	6.06E-05	5.78			错义_	0.203	0.04	0.346	43	0
6	HP:000048	5_8999044	VUS	0.003192	5.81			错义_	0.018		0.066	29	可能良性	0
6	HP:000048	5_7094599	VUS	0.00015	3.84	0.45	0.98	错义_	0.037	0.05	0.032	43	0
7	HP:000058	2_1795474	LB	0.01105	-3.98			同义_变体			0.352		可能良性	0
7	HP:000058	18_485934	P	1.00E-04	5.49	0.34	0.109	错义_	0.912	0.04	1	32	不确定	0
8	HP:000194	9_1171857	VUS	0.009235	4.41			错义_	0.88		0.248	98	可能良性	0
8	HP:000194	11_663347	B	0.000539	-1	0.001	0.876	同义_变体			0.109			0
8	HP:000194	X_4907497	LB	0	4.73			终止密码子_获得			0.231			0
9	HP:000194	3_1506583	VUS	0.001079	0.649	0.762	0.999956	剪接_受体_变体			0.166		不确定	1
9	HP:000194	6_1372193	LP	0	5.96			错义_	0.905	0.13	0.096	22		0
9	HP:000194	10_735581	B	0.005642	4.63			同义_变体			0.274		可能良性	0
9	HP:000194	17_364935	LP	0.005394	3.1			错义_	0.052	0.13	0.07	43	不确定	0
10	HP:000194	10_735376	B	0.000458	-11			错义_变体			0.274	23		0
11	HP:000150	4_3634519	LB	0	2.58	0.987	0.567	错义_	0.026	0.46		145		0
11	HP:000150	15_784016	P	0.0032	-7.53	0.26	0.02	同义_变体			0.313			0
12	HP:000047	11_119212	VUS	0.008287	-6.19	0.4	0.6	同义_变体			0.158		可能良性	0
13	HP:000070	2_2024980	B	0.006272	1.46	0.6	0.24	同义_变体			0.073		可能良性	0

[0053] 图1a是展示了根据本发明的评估变体对患者的致病性的示例过程100的流程图。可以通过使用带注释的数据训练的至少一种预测模型来评估致病性的水平。通过过程100评估变体的致病性的步骤如下：

[0054] 在步骤102中，接收与患者相关联的变体。变体可以是模型已知的变体或未知的变体。另外或可替代地，与变体一起，患者的表型信息还可以用于评估致病性。

[0055] 在步骤104中，确定变体的与预测模型的致病度量相关的至少一种概率。对预测模型进行训练以保留模型习得的变体集合或变体的数据表示。习得的变体集合包括在确定变体本身的至少一种概率时至少一个遗传病症簇的数据表示。另外或可替代地，至少一个遗传病症簇的数据表示是由习得的变体集合导出的，并且是关于患者的表型信息集进行加权的。可以考虑在某种程度上在不存在患者的表型信息的情况下调整为至少一个遗传病症簇以输出组合表示的情况下评估和确定的患者的表型信息的可用性。作为选项，可以将组合表示，即对表型指标中的每个表型指标生成的概率，相对于相应概率归一化为100％或1。

[0056] 在步骤106中，输出患者的变体的至少一种概率。输出可以是生成的概率的组合表示。在一个实例中，输出可以是界面的一部分，其中用户可以将潜在概率视为具有准备用户的解释以供审查的自动化助理。更具体地，与概率的组合表示一起，界面可以提示至少一个输出，所述至少一个输出包含但不限于与致病性的水平、对表型的贡献、报告类别等相对应的指定标签。可以将另外的解释性信息呈现为组合输出的一部分。

[0057] 另外或可替代地，一旦接收到患者的表型信息，条件是在某种程度上变体被视为

是至少一种预测模型已知时所述变体被包含在习得的变体集合中，则可以基于患者的表型信息来确定与至少一个遗传病症簇中的每个遗传病症簇相关联的贡献。在此确定的情况下，作为选项，使用至少一种预测模型的一个或多个回归模型对至少一个遗传病症簇中的每个遗传病症簇进行划分。一个或多个回归模型在给定患者的表型信息的情况下预测对至少一个遗传病症簇中的每个遗传病症簇的贡献。因此，基于与所述至少一个遗传病症簇的数据表示相关的贡献来调整变体的至少一种概率。实际上，贡献提供了与所提供的表型信息一致的提高的准确性。

[0058] 在将未知变体呈现给至少一种预测模型，使得变体不包含在习得的变体集合中的情况下，使用监督学习框架以在给定未知变体的边信息集的情况下计算致病度量上的概率分布，所述边信息可以包括一种或多种表型指标和/或基因组指标。实际上，预测模型未知或未看到的任何变体都可以基于已知或习得的变体的储库或集合来相应地评估。

[0059] 图1b是展示了根据本发明的基于参考图1a描述的示例过程100的其中关于表型信息126和边信息124评估变体对患者的致病性的示例过程120的示意图。确定122接收到的变体是否在习得的变体集合内。如果“是”，则接收到的变体是预测模型已知的，应用患者的表型信息以确定对潜变量或遗传病症簇的贡献。如通过一种或多种生成模型或ML模型导出的遗传病症簇或应用本文描述的ML技术进而提供了基于致病度量对致病性的经验评估。

[0060] 在一个实例中，可以根据线性回归模型126b使用患者的HPO术语126a来确定潜变量中的每个潜变量的贡献126c。使用LDA导出潜变量，在所述LDA中进行矩阵分解。因此，可以使用患者的额外表型信息和/或在接收到的变体的情况下直接通过应用潜变量或隐藏遗传病症簇来确定输入的变体是良性的还是另一种致病度量的证据或概率。可以基于致病度量，例如，良性、可能良性、可能致病和致病来确定相似性概率。也就是说，致病度量可以包括指示致病性的程度或水平的至少一个分类。至少一个分类可以与至少一个遗传病症簇的不同最优集合相关联，使得可以呈现和输出这些度量的具有良性128a、可能良性128b、可能致病128c和致病128d的潜在概率的组合表示128。

[0061] 在“否”的情况下，则接收到的变体对于预测模型是未知的，可以使用与监督学习框架相关的归属于一个或多个表型和/或基因组指标的另外的边信息124。可以应用监督学习框架以基于接收到的边信息124a来计算致病度量124b的概率分布。边信息用于评估指示致病性的程度的与致病度量相关联的所产生的概率。实际上，边信息的应用克服了将未知变体呈现给预测模型的困境。

[0062] 图2a是展示了根据本发明的生成用于确定变体的与致病度量相关的至少一种概率的遗传病症簇的示例过程200的流程图。在此实例中，使用带注释的数据来训练预测模型。具体地，带注释的数据用于导出与至少一种生成模型或ML模型相关联或者应用本文所述的一种或多种ML技术的隐藏遗传病症簇。在此实例中，生成遗传簇的过程200可以包含以下步骤：

[0063] 在步骤202中，接收至少一个患者的与变体集合相关联的带注释的数据。接收到的带注释的数据可以包括解释信息和与致病度量相对应的观察结果。解释信息在性质上可以是基因型的。另外或可替代地，带注释的数据可以进一步包括患者的与和至少一个患者相关的解释信息和/或边信息集相关联，与和变体集合相关的解释信息相关联的表型信息集，在某种程度上，边信息集可以包含与变体集合相关联的指标的数据表示。

[0064] 具体地,边信息集可以在变体未被包含在变体集合中或未作为带注释的数据的一部分被接收时用于通过使用监督学习框架来计算在致病度量上的概率分布。

[0065] 作为选项,可以基于表型信息集来调整与至少一个遗传病症簇相关联的权重集。所述权重集可以与至少一个遗传病症簇对表型信息集的贡献相对应。可以基于经调整的权重集将一个或多个回归模型配置成确定与致病度量相关的贡献。另外或可替代地,还可以应用一种或多种ML模型或技术来获得对遗传病症簇的贡献。

[0066] 在步骤204中,可以使用本文描述的一种或多种生成模型或对应ML模型或ML技术来确定和导出至少一个患者的接收到的带注释的数据的数据表示。一种或多种生成模型被配置成分解与致病度量相关的带注释的数据的数据表示。例如,可以应用如LDA等矩阵因式分解算法。

[0067] 在此实例中,LDA的隐藏遗传病症簇为抽象参数,其是使用患者、变体和对应观察结果的多维数据矩阵的分解导出的。导出的遗传病症簇使得能够汇编可以用于评估给定变体的致病性的概率。在对多维数据矩阵进行分解或因式分解之后,可以例如,通过使用期望最大化来确定遗传病症簇的最优数量。因此,遗传病症簇的数量可能会在预测模型随着更多数据递增时变化。如k折(k-fold)交叉验证(例如,k=5)的替代技术也可以适用,因为遗传病症簇的最优数量可以使用困惑度的概念作为评估评分来确定和评分——最优解决方案是使困惑度最小化的解决方案。在这种情况下,应该对与表型度量相关联的每个二进制矩阵执行不同的分解,使得每个分解可以具有不同最优数量的遗传病症簇或潜变量。

[0068] 在步骤206中,基于数据表示生成至少一个遗传病症簇。数据表示可以是抽象参数或者可替代地本文所述的一种或多种ML模型的ML特征。还可以使用一种或多种ML模型或技术以基于带注释的数据加上本申请的实例中的任何实例中描述的技术或与所述技术结合来确定至少一个遗传病症簇的最优集。进而,可以使用至少一个遗传病症簇的最优集以预测变体的与致病度量相关的至少一种概率。另外或可替代地,至少一个遗传病症簇的最优集可以被配置成使用新的或额外的带注释的数据迭代地更新。

[0069] 图2b是根据本发明的基于参照图2a描述的示例过程200的用于确定变体的概率的遗传病症簇的示例过程220的示意图。为了生成遗传病症簇228,多维数据矩阵222的数据表示可以用作用于确定簇的输入224。具体地,数据矩阵222并入了患者的信息、变体和对应观察结果(来自过去的患者解释的“带标记的数据”)。通常情况是,相对于矩阵的大小,矩阵中的观察结果是高度稀疏的,观察结果“单元格”的大约99.96%是空的,因为存在很多可能变体。

[0070] 更具体地,多维数据矩阵222可以关于与患者、变体和对应观察结果相关联的数据在表型信息矩阵222a、解释信息矩阵222b和边信息矩阵222c方面呈现。具体地,可以对解释信息矩阵222b进行分解,以生成遗传病症簇。表型信息的实例可以包含HPO术语(患者1到4中存在的HPO 1到3),并且解释信息可以包含变体或其集合(其中例如,患者1具有两个标记为致病性的变体,并且患者3不具有致病性变体)。另一方面,边信息矩阵与如GREP评分、SIFT评分、VEP结果、MVP评分、HI评分、ADA评分等表型指标和基因型指标相对应。例如,边信息矩阵222c可以包括含有实数(即,最大等位基因频率)的列以及含有分类变量(即,VEP结果)的列。分类变量可以通过使用虚拟编码方案转换为整数(二进制)表示。因此,每个患者都具有将患者的表型(或体征/症状)描述为HPO术语或应用其它表型编码模式(例如OMIM、

IDC10等)的边信息(或二进制向量)。可以使用含有数据集中的所有患者的HPO或其定量值的矩阵来训练,例如,回归模型以确定遗传病症簇。

[0071] 进一步地在图2b中,与致病性度量(例如,B、LB、P、LP)相关的解释信息矩阵被分解(即,分解成H 226b和W 226c,两者一起相乘回去得到V 226a)。解释信息矩阵的分解生成一定数量的二进制矩阵,所述数量等于致病性度量的数量。在此,使用矩阵W 226c以表示训练数据集中的每个遗传病症簇228在每个患者内部的比例。矩阵H 226b含有每个变体与每个遗传病症簇228相关联的次数。因此,遗传病症簇仅仅是矩阵分解的一个维度。进而,可以应用如通过期望最大化进行的LDA等矩阵因式分解算法以优化遗传病症簇的有限集。遗传病症簇的有限集可以通过使用验证技术(例如k折)来确定。遗传病症簇228的有限集的最优数量(例如,5、6、7...25)可以被存储,并且可以在不同数量的遗传病症簇在验证技术期间变得或被确定为最优时继续进行更新。实际上,在给定与四个致病水平相对应的四种分解的情况下,可以确定对习得的变体集合中含有的任何变体的预测。

[0072] 图3是展示了根据本发明的使用边信息集评估未知变体对患者的致病性的示例过程300的流程图。任何未知变体都是未包含在预测模型已习得的习得的变体集合中的变体。基于未知变体的边信息,通过使用监督预测模型在致病度量上的概率分布。

[0073] 在步骤302中,接收到未在习得的变体集合中标识出的未知变体。接收到的未知变体可以是患者的预测模型尚未发现或遗传病症簇未具体分类的任何变体。

[0074] 在步骤304中,可以评估未知变体的致病性。此评估通过使用监督学习框架进行,所述监督学习框架包含一个或多个监督预测模型,所述一个或多个监督预测模型在给定变体的边信息的情况下生成每个致病度量的概率。例如,输出可以以直方图的形式呈现展示每个度量的归一化概率。

[0075] 作为不同选项,将与习得的变体集合的每个子集相对应的边信息集进行比较,以确定最接近的变体。作为另一种选项,关于相似性评分将与习得的变体集合的所述子集中的每个子集相对应的边信息集进行比较。例如,相似性评分可以为余弦相似性评分或适于评估习得的变体集合的子集以确定最接近的变体的其它合适的评分方法。

[0076] 作为另一种选项,可以关于最接近的变体的致病性评估未知变体的致病性。具体地,可以基于习得的变体集合确定最接近的变体的至少一种概率。此确定是关于包括至少一个遗传病症簇的数据表示的致病度量确定的。也就是说,可以应用最后一个遗传病症簇来计算最接近的变体的至少一种概率。所计算的至少一种概率可以被编译为引入组合表示,在这种情况下组合表示是相对于致病度量输出的。所述输出例如可以以直方图的形式显示每个度量的归一化概率。另外或可替代地,可以响应于所述习得的变体集合的所述子集包括两个或更多个具有等效相似性评分的变体使得无法确定所述最接近的变体而通过对习得的变体集合的子集的每个变体的至少一种概率求平均来生成组合表示。

[0077] 作为另一种选项,本文所述的实例中的任何实例的致病度量可以包括指示致病性的程度的至少一个分类。至少一个分类中的每个分类可以进一步与至少一个遗传病症簇的不同最优集相关联。遗传病症的最优集可以在结合期望最大化应用例如LDA时或者可替代地通过本文所述的一种或多种ML模型或技术来确定。具体地,合适的验证技术也可以适用于确定最优集中的遗传病症簇的数量,例如通过将困惑度最小化,使得每个分解可以具有不同的最优数量的遗传病症簇。对于与表型度量相关联的每个二元矩阵,遗传病症的不同

最优数量可以通过使用本文所述的用于确定遗传病症簇的最优数量的任何技术导出。

[0078] 作为另一种选项,可以使用加权相似性度量来标识或确定最佳最接近的变体或在加权相似性度量方面与未知变体最相似的变体。加权相似性度量可以针对不同的边信息而保留不同或相似的权重。具体地,边信息的一个评分可以具有比另一个评分的权重更高的权重,并且在计算最接近的变体时,评分越高,影响将越大。使用加权相似性度量的目的是考虑每个边信息特定的预测能力,并且增强标识最佳最接近的习得的变体的过程。这些权重可以通过使用与本文所述的一种或多种ML技术相关联的线性模型和非线性模型两者来推断。

[0079] 图4是展示了根据本发明的参考图1a到3的从带注释的数据中提取以在给定致病度量的情况下预测变体的概率的遗传病症簇的示例过程400的示意图。在实例中,可以从用作模型的训练数据集的带注释的数据中提取作为预测模型的基础的潜在或隐藏的遗传簇或潜变量。数据集可以呈多维数据矩阵的形式,包括与患者、变体以及在矩阵中以数字方式呈现的对应观察结果相关联的数据点。所提取的遗传病症簇可以是在分解程序时生成的矩阵的单个维度(向量)。每个分解都与致病度量(B、LP、P和LP)相关联,如图所示。除了所示的度量外,具有不同程度的致病性的替代性致病度量也可以适用。在推导出四个分解的情况下,可以进行对驻留在带注释的数据中的任何变体的致病性的预测。在图中,在存在对致病度量中的每个致病度量的所产生的分解的情况下,分解是通过对矩阵进行LDA实现的。分解程序可以可替代地使用许多其它技术来完成,所述其它技术包含为了降低数据的维数而描述的一种或多种ML技术。因此,遗传病症簇的合成向量有效地体现了带注释的数据。

[0080] 进一步地,在此实例中,遗传病症簇可以关于表型信息402b进行加权。遗传病症簇的加权解决了预测被证明对具有不同表型的患者相同的情况。因此,预测模型的准确性由于患者的表型可以作为模型的框架的一部分被包含在内而增加中,并且所产生的预测可能与每个患者的特定特征相关。如图所示,使用线性回归模型作为实例,目的是在给定如患者的HPO术语等表型信息的情况下,预测或计算每个遗传病症簇的贡献408。HPO术语的这些实例可以用于通过将权重与每个遗传病症簇相关联来调整生成的谱的总体概率。作为一种选项,在没有提供HPO术语作为输入的情况下,则不对遗传病症簇应用加权。可以将为每个患者和特定变体生成的谱显示为基于致病性度量410的归一化概率。

[0081] 另外或可替代地,可以在患者的输入变体不存在于带注释的数据中或不是与遗传病症簇相关联的习得的变体的一部分的情况下,使用边信息402a。换言之,在将新的或未知变体呈现给预测模型时,监督预测模型406可以使用边信息402a来确定未知变体的致病度量的概率分布,而无须关于已知解释再训练预测模型。

[0082] 例如,可以使用监督学习框架以通过使用本文所述的边信息402a来计算致病性。因此,预测模型高于预测已知变体和未知变体两者,而无需在遇到未知变体并且增强模型的可持续性时对所需的精度再训练。

[0083] 作为不同选项,可以在患者的输入变体不存在于带注释的数据中或不是与遗传病症簇相关联的习得的变体的一部分的情况下,使用边信息。换言之,在向预测模型呈现新的或未知变体时,使用边信息来确定最接近的变体,而不必关于已知解释再训练预测模型(以及生成/更新新的遗传病症簇)。

[0084] 在不同的选项中,可以使用余弦相似性以在多维图表上绘制变体。使用本文所述

的边信息中的一个或多个边信息，可以将最近的或与习得的变体集合的距离小(基于余弦相似性评分)的变体预测为预测变体。具体地，从多维图表中标识具有最相似余弦评分或有效地具有相似变体边信息的变体。预测变体将替代输入的变体，以便生成每个患者和输入的变体的表型。也就是说，然后使用矩阵H中的最邻近的邻居的条目作为未知变体的代理项，并且以与变体已知的方式相同的方式生成概率预测。如果两个或更多个变体具有相同的(argmax)余弦相似性评分，则通过对跨所有选定变体的结果求平均来计算最终概率。因此，预测模型高于在不需要在遇到未知变体时对所需的精度再训练的情况下预测已知变体和未知变体两者并且增强模型的可持续性。

[0085] 图5是展示了示例计算设备/系统500的示意图，所述示例计算设备/系统可以用于实施预测模型、设备、方法和/或其过程组合、其修改形式和/或如参考图1a到4所述的和/或如本文所述的一个或多个方面。计算设备/系统500包含一个或多个处理器单元502、输入/输出单元504、通信单元/接口506、存储单元508，其中一个或多个处理器单元502连接到输入/输出单元504、通信单元/接口506和存储单元508。在一些实施例中，计算设备/系统500可以是服务器，或者联网在一起的一个或多个服务器。在一些实施例中，计算设备/系统500可以是适于处理或执行用于致病性评估系统、设备、方法和/或其过程组合、其修改形式和/或如参考图1a到4所述的和/或如本文所述的预测模型的一个或多个方面的计算机或超级计算机/处理设施或硬件/软件。通信接口506可以通过通信网络将计算设备/系统500与一个或多个服务、设备、服务器系统、基于云的平台、用于实施主题数据库的系统和/或用于实施本文所述的发明的知识图连接。存储器单元508可以存储一个或多个程序指令、代码或组成部分，如但不限于：与参照图1a到图4描述的变体过程/方法的评估相关联的操作系统和/或代码/组成部分、额外数据、应用、应用固件/软件和/或另外的程序指令、与实施功能和/或一个或多个功能或与承载致病性评估过程/方法/系统、设备、机制和/或用于实施如本文所述的发明的系统/平台/架构的预测模型的装置、服务和/或服务器的方法和/或过程中的一种或多种相关联的功能相关联的代码和/或组成部分、其组合、其修改形式和/或如参考图1a到4中的至少一个描述的。

[0086] 在实施例中，如上所述的本发明的实例，如用于致病性评估过程、方法、系统和/或设备的预测模型可以在一个或多个云平台、一个或多个服务器或计算系统或装置上实施和/或可以包括一个或多个云平台、一个或多个服务器或计算系统或装置。服务器可以包括单个服务器或服务器网络，云平台可以包含多个服务器或服务器网络。在一些实例中，服务器和/或云平台的功能可以由跨地理区域分布的服务器网络，如服务器全球分布式网络来提供，并且用户可以基于用户位置等连接到服务器网络中的合适的一个。

[0087] 在与图1a至图4相关的一方面，一种用于评估变体对患者的致病性的计算机实施的方法，所述方法包括：接收变体；基于习得的变体集合确定所述变体的与致病度量相关的至少一种概率，其中所述致病度量包括用于确定所述变体的所述至少一种概率的至少一个遗传病症簇的数据表示；以及输出所述患者的所述变体的所述至少一种概率的组合表示。

[0088] 另一方面，一种用于生成至少一个遗传病症簇的计算机实施的方法，所述至少一个遗传病症簇用于确定变体的与致病度量相关的至少一种概率，所述方法包括：接收至少一个患者的与变体集合相关联的带注释的数据，其中所述带注释的数据包括解释信息以及与所述致病度量相对应的相关观察结果；确定至少一个患者的所述带注释的数据的数据表

示，其中所述数据表示是使用一种或多种生成模型导出的；以及基于所述数据表示生成所述至少一个遗传病症簇。

[0089] 在又另一方面，一种用于使用边信息集来评估未知变体对患者的致病性的计算机实施的方法，所述方法包括：接收所述未知变体，其中所述未知变体未在习得的变体集合中标识出；使用与所述习得的变体集合的每个子集相对应的所述边信息集来训练监督学习框架；以及基于所述监督学习框架来评估所述未知变体的所述致病性。

[0090] 在又另一方面，一种计算机可读介质，其包括存储在其上的计算机可读代码或指令，所述计算机可读代码或指令当在处理器上执行时，使所述处理器根据以下任选地描述的任何步骤来实施所述计算机实施的方法。

[0091] 在又另一方面，一种系统，其包括至少一个电路系统，所述至少一个电路系统被配置成根据以下任选地描述的任何步骤来执行所述计算机实施的方法。

[0092] 在又另一方面，一种设备，其包括处理器、存储器和通信接口，所述处理器连接到所述存储器和所述通信接口，其中所述设备适于或被配置成实施根据以下任选地描述的步骤。

[0093] 在又另一方面，一种用于确定变体对患者的致病性的设备，所述设备包括：输入组件，所述输入组件被配置成接收所述变体；处理组件，所述处理组件被配置成确定所述变体是否在习得的变体集合内；预测组件，响应于确定所述变体存在于所述习得的变体集合中，所述预测组件被配置成生成所述变体的与致病度量相关的至少一种概率，其中所述致病度量包括用于确定所述变体的所述至少一种概率的至少一个遗传病症簇的数据表示；以及显示组件，所述显示组件被配置成显示所述变体的关于所述致病度量的所述至少一种概率，其中所述至少一种概率被归一化。

[0094] 在又另一方面为一种用于使用边信息集来确定未知基因变体的致病性概率分布的计算机实施的方法，所述方法包括：接收患者的所述未知变体，其中所述未知变体未在与多个患者相关联的习得的变体集合中标识出或者对所述习得的变体集合来说是新的；基于所述边信息集通过使用监督学习框架来评估所述未知基因变体的所述致病性；以及基于所述评估确定所述致病性概率分布。

[0095] 在适当的情况下，以下任选步骤涉及以上方面中的任何一个或多个方面。

[0096] 任选地，响应于确定所述变体不存在于所述习得的变体集合中，所述预测组件被配置成接收边信息集，其中所述边信息用于关于所述变体标识最接近的变体，所述最接近的变体作为所述变体被应用以生成所述至少一种概率。

[0097] 任选地，所述输入组件被配置成接收与所述患者相关联的表型信息，其中所述表型信息用于调整所述变体的与所述至少一个遗传病症簇相关的所述至少一种概率。

[0098] 任选地，所述至少一个遗传病症簇的所述数据表示是由所述习得的变体集合导出的，并且是关于患者的表型信息集进行加权的。

[0099] 任选地，所述变体被包含在所述习得的变体集合中，所述方法进一步包括：接收所述患者的表型信息；基于所述患者的所述表型信息确定与所述至少一个遗传病症簇中的每个遗传病症簇相关联的贡献；以及基于根据所述至少一个遗传病症簇的所述数据表示确定的所述贡献来调整所述变体的所述至少一种概率。

[0100] 任选地，所述计算机实施的方法进一步包括：评估所述患者的所述表型信息的可

用性;以及基于所述可用性确定是否调整所述至少一个遗传病症簇以输出所述组合表示。

[0101] 任选地,基于所述患者的所述表型信息确定与所述至少一个遗传病症簇中的每个遗传病症簇相关联的贡献,进一步包括:使用一个或多个回归模型对所述至少一个遗传病症簇中的每个遗传病症簇进行分割,其中所述一个或多个回归模型在给定所述患者的所述表型信息的情况下预测对所述至少一个遗传病症簇中的每个遗传病症簇的贡献。

[0102] 任选地,变体未被包含在所述习得的变体集合中,所述方法进一步包括:从所述习得的变体集合中识别与所述变体相关的至少一个邻近变体;接收与所述至少一个邻近变体中的每个邻近变体相对应的边信息集,其中所述边信息集包括一个或多个指标;基于所述边信息集标识最接近的变体;以及在确定所述变体的与所述致病度量相关的所述至少一种概率时,应用所述最接近的变体作为所述变体。

[0103] 任选地,基于所述边信息集通过应用与所述至少一个邻近变体相关联的相似性度量来标识所述最接近的变体。

[0104] 任选地,关于所述边信息集对所述相似性度量进行加权。

[0105] 任选地,当所述相似性度量从所述习得的变体集合中标识出至少一个其它变体具有等效相似性评分时,通过对所述至少一个邻近变体中的每个邻近变体求平均来确定所述变体的所述至少一种概率。

[0106] 任选地,所述带注释的数据进一步包括患者的表型信息集和/或边信息集。

[0107] 任选地,所述表型信息集与和所述至少一个患者相关的所述解释信息相关联;和/或其中所述边信息集与和所述变体集合相关的所述解释信息相关联。

[0108] 任选地,所述计算机实施的方法进一步包括:基于所述表型信息集调整与所述至少一个遗传病症簇相关联的权重集,其中所述权重集与所述至少一个遗传病症簇对所述表型信息集的贡献相对应;以及基于经调整的权重集将一个或多个回归模型配置成确定与所述致病度量相关的所述贡献。

[0109] 任选地,所述边信息集包括与所述变体集合相关联的指标的数据表示。

[0110] 任选地,在所述变体未被包含在所述变体集合中时,应用所述边信息集以从所述变体集合中标识用于确定所述变体的所述至少一种概率的最接近的变体。

[0111] 任选地,所述变体被包含在所述变体集合中,以通过应用与所述最接近的变体相关联的注释来更新所述最少一个遗传病症簇。

[0112] 任选地,所述计算机实施的方法进一步包括:基于所述带注释的数据确定所述至少一个遗传病症簇的最优集;以及在预测期间应用所述至少一个遗传病症簇的所述最优集来确定变体的与所述致病度量相关的所述至少一种概率。

[0113] 任选地,所述至少一个遗传病症簇的所述最优集被配置成用新的带注释的数据迭代地更新。

[0114] 任选地,关于与所述习得的变体集合的每个子集相关联的相似性评分,比较与所述习得的变体集合的所述子集相对应的所述边信息集。

[0115] 任选地,关于所述最接近的变体的所述致病性评估所述未知变体的所述致病性,所述评估进一步包括:基于习得的变体集合确定所述最接近的变体的与致病度量相关的至少一种概率,其中所述致病度量包括用于计算所述最接近的变体的所述至少一种概率的至少一个遗传病症簇的数据表示;以及生成所述至少一种概率的组合表示,其中所述组合表

示是相对于所述致病度量输出的。

[0116] 任选地，所述计算机实施的方法进一步包括：响应于所述习得的变体集合的所述子集包括两个或更多个具有等效相似性评分的变体使得无法确定所述最接近的变体而通过对所述习得的变体集合的子集的每个变体的所述至少一种概率求平均来生成所述组合表示。

[0117] 任选地，所述表型信息包括与一种或多种疾病相关联的表型本体。

[0118] 任选地，一种或多种生成模型被配置成分解与致病度量相关的带注释的数据的数据表示。

[0119] 任选地，所述一种或多种生成模型包括至少一个基于矩阵因式分解算法的公式。

[0120] 任选地，所述致病度量指示致病性的程度或水平的至少一个分类。

[0121] 任选地，所述至少一个分类中的每个分类与所述至少一个遗传病症簇的不同最优集相关联。

[0122] 任选地，在给定所述边信息集的情况下，计算所述未知变体的与致病度量集相关联的概率。

[0123] 任选地，基于习得的变体集合进一步确定所述未知变体与致病度量相关的至少一种概率；以及生成所述至少一种概率的组合表示，其中所述组合表示是相对于所述致病度量输出的。

[0124] 任选地，所述致病度量包括用于计算最接近的变体的至少一种概率的至少一个遗传病症簇的数据表示。

[0125] 任选地，所述监督学习框架包括一种或多种预测模型。

[0126] 任选地，所述监督学习框架包括非参数分类器。

[0127] 任选地，所述边信息集与所述未知基因变体相关联。

[0128] 为清楚起见，上文描述参考单个用户讨论了本发明的实施例。应当理解，实际上，系统可以由多个用户共享并且可能同时由非常大量的用户共享。

[0129] 上述实施例可以被配置成半自动和/或被配置成全自动的。在一些实例中，致病性评估系统/过程/方法的预测模型的用户或操作者可以手动地指示要执行的过程/方法的一些步骤。

[0130] 本发明所述的实施例，根据本发明和/或如本发明所述的用于致病性评估系统、过程、方法和/或设备等的预测模型可以作为任何形式的计算装置和/或电子装置实施。此类装置可以包括一个或多个处理器，所述一个或多个处理器可以是用于对计算机可执行指令进行处理以控制所述装置的操作以采集并记录路由信息的微处理器、控制器或任何其它合适类型的处理器。在一些实例中，例如在使用片上系统架构的情况下，处理器可以包含一个或多个固定功能块(也称为加速器)，所述一个或多个固定功能块以硬件(而不是软件或固件)的形式实施过程/方法的一部分。包括操作系统的平台软件或任何其它合适的平台软件可以设置在基于计算的装置处以使应用软件能够在所述装置上执行。

[0131] 本文所描述的各种功能可以以硬件、软件或其任何组合实施。如果以软件实施，则可以将功能作为一或多个指令或代码存储在计算机可读介质上或者通过计算机可读介质进行传输。计算机可读介质可以包含例如计算机可读存储介质。计算机可读存储介质可以包含在任何方法或技术中实施的用于存储如计算机可读指令、数据结构、程序模块、或其它

数据等信息的易失性或非易失性介质、可移除或不可移除介质。计算机可读存储介质可以是可以被计算机访问的任何可用存储介质。通过举例而非限制，此种计算机可读存储介质可以包含RAM、ROM、EEPROM、闪速存储器或其它存储器装置、CD-ROM或其它光盘存储装置、磁盘存储装置或其它磁性存储装置或可以用于承载或存储采用指令或数据结构形式的期望程序代码并且可以被计算机访问的任何其它介质。如本所使用的，盘和碟包含压缩光碟(CD)、激光碟、光碟、数字通用光碟(DVD)、软盘和蓝光盘(BD)。进一步地，传播的信号不包含在计算机可读存储介质的范围内。计算机可读介质还包含通信介质，所述通信介质包含促进将计算机程序从一处传送到另一处的任何介质。例如，连接或者偶联可以是通信介质。例如，如果使用同轴电缆、光纤电缆、双绞线、DSL或如红外、无线电和微波等无线技术从网站、服务器或其它远程源传输软件，则同轴电缆、光纤电缆、双绞线、DSL或如红外、无线电和微波等无线技术包含在通信介质的定义中。上述内容的组合也应包含在计算机可读介质的范围内。

[0132] 可替代地或另外地，本文所描述的功能性可以至少部分地由一个或多个硬件逻辑组件执行。例如而非限制，可以使用的硬件逻辑组件可以包含现场可编程门阵列(FPGA)、专用集成电路(ASIC)、专用标准产品(ASSP)、片上系统(SOC)、复杂可编程逻辑器件(CPLD)等。

[0133] 尽管被示为单个系统，但是应当理解，计算装置可以是分布式系统。因此，举例来说，几个装置可以通过网络连接进行通信并且可以共同执行被描述为由计算装置执行的任务。

[0134] 尽管展示为本地装置，应当了解，计算装置可以定位在远端并且通过网络或其它通信链路(例如使用通信接口)被访问。

[0135] 术语“计算机”在本文中用于指代具有处理能力使得其可以执行指令的任何装置。本领域技术人员将认识到，此类处理能力并入到许多不同的装置中，并且因此术语“计算机”包含PC、服务器、物联网(IoT)装置、移动电话、个人数字助理和许多其它装置。

[0136] 本领域技术人员将认识到，用于存储程序指令的存储设备可以跨网络分布。例如，远程计算机可以存储被描述为软件的处理实例。本地或终端计算机可以访问远程计算机并且下载软件的一部分或全部以运行程序。可替代地，本地计算机可以按需下载一个软件或执行本地终端处的一些软件指令和远程计算机(或计算机网络)处的一些软件指令。本领域的技术人员还将认识到，通过利用本领域的技术人员已知的常规技术，软件指令的全部或部分可以由专用电路(如DSP、可编程逻辑阵列等)执行。

[0137] 应当理解，上文所描述的益处和优点可以涉及一个实施例或可以涉及若干实施例。所述实施例不限于解决任何或全部所陈述的问题的那些实施例或具有任何或全部所陈述的益处和优点的那些实施例。变体应被认为包含在本发明的范围内。

[0138] 对“一个”项的任何提及是指那些项中的一个或多个项。术语“包含”在本文中用于意指包括经鉴定的方法步骤或要素，但是所述此类步骤或要素不包含排他性列表并且方法或设备可以含有另外的步骤或要素。

[0139] 如本文所使用的，术语“组件”和“系统”旨在涵盖配置有使某些功能性可以在由处理器执行时被执行的计算机可执行指令的计算机可读数据存储装置。计算机可执行指令可以包括例程、函数等。还应当理解，组件或系统可以本地化在单个装置上或跨几个装置分布。进一步地，本文所使用的术语“示例性”旨在意指“充当某些的说明或实例”。进一步地，

对于在具体实施方式或权利要求书中使用了术语“包含”而言，此种术语旨在以与术语“包括”相似的方式是开放式的，这是由于“包括”在被使用时被解释成权利要求项中的过渡词。

[0140] 附图展示了示范性方法。虽然方法被示出和描述为按特定序列执行的一系列动作，但是应当理解和了解，所述方法不受序列的顺序的限制。例如，一些动作可以与本文所描述的顺序不同的顺序发生。另外，一个动作可以与另一个动作同时发生。此外，在一些情况下，可能不需要完成所有动作来实现本文所述的方法。

[0141] 此外，本文所描述的动作可以包括可以由一个或多个处理器实施的和/或存储在一个或多个计算机可读介质上的计算机可执行指令。计算机可执行指令可以包含例程、子例程、程序、执行线程等。仍进一步地，方法的动作的结果可以存储在计算可读介质中、在显示装置上显示和/或等等。

[0142] 本文所描述的方法的步骤的顺序是示例性的，但是这些步骤可以在适当的情况下按任何适合的顺序或同时执行。另外地，在不脱离本文所描述的主题的范围的情况下，可以添加或替换步骤或者可以从任何方法中删除单独的步骤。上文所描述的任何实例的各方面可以与所描述的任何其它实例的各方面结合以在不损失效应的情况下形成另外的实例。

[0143] 应当理解，优选实施例的以上描述仅通过实例的方式给出并且本领域的技术人员可以做出各种修改。

[0144] 上文已描述的内容包含一个或多个实施例的实例。当然，不可能出于描述上文提及的方面的目的而描述出对上述装置或方法的每一种可以想到的修改和改变，但是本领域普通技术人员可以认识到，各方面的许多另外的修改和排列是可能的。因此，所描述的各方面旨在涵盖落入所附权利要求书的范围内的所有此种改变、修改以及变化。

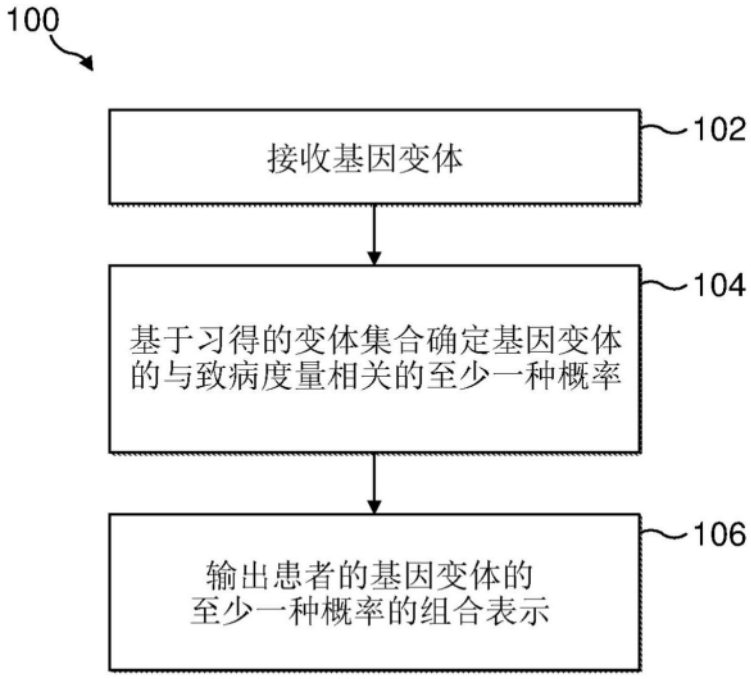

100
接收基因变体
102
基于习得的变体集合确定基因变体的与致病度量相关的至少一种概率
104
输出患者的基因变体的至少一种概率的组合表示
106

图1a

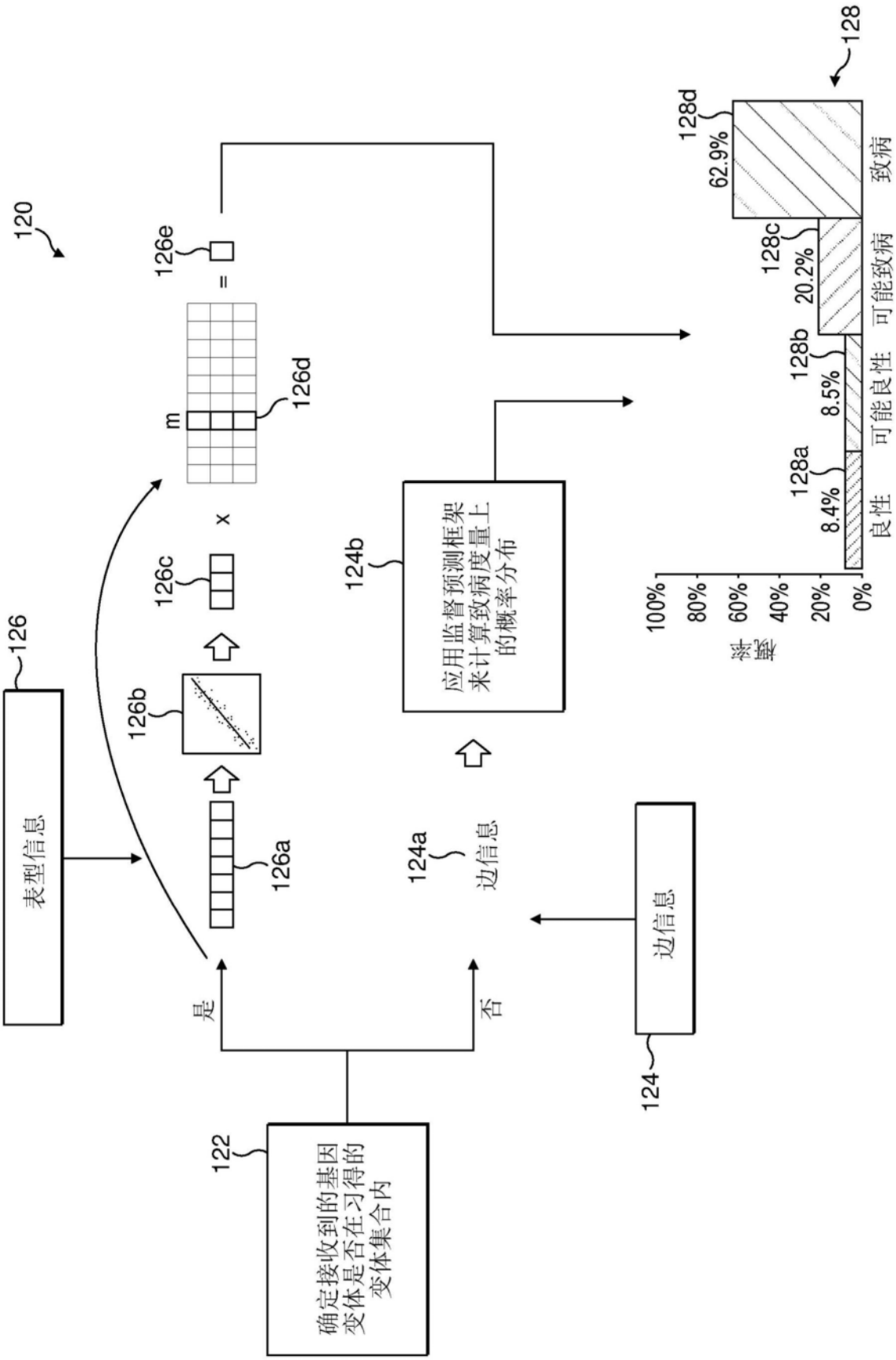

120
126
表型信息
126a
126b
126c
x
126d
m
126e
=
是
否
122
确定接收到的基因变体是否在习得的变体集合内
124
边信息
124a
边信息
124b
应用监督预测框架来计算致病度量上的概率分布
128
128d
62.9%
致病
128c
20.2%
可能致病
128b
8.5%
可能良性
128a
8.4%
良性
100%
80%
60%
40%
20%
0%
概率

图1b

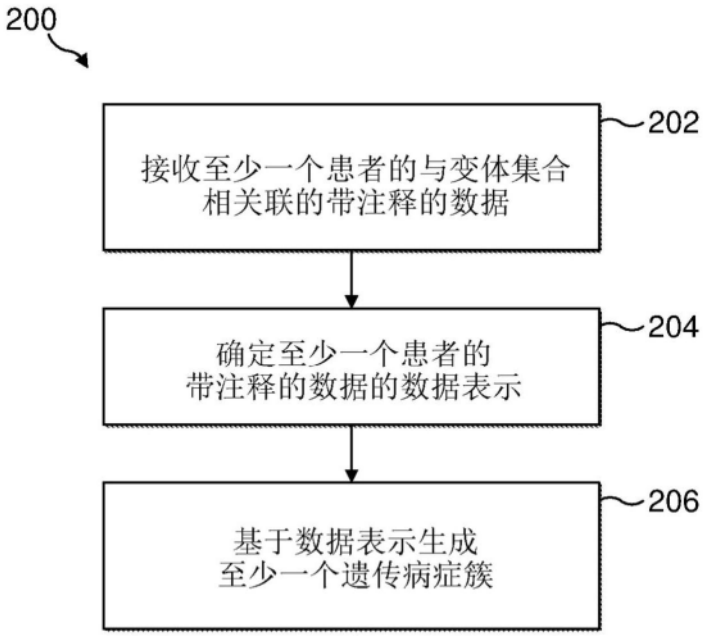

200
接收至少一个患者的与变体集合相关联的带注释的数据
202
确定至少一个患者的带注释的数据的数据表示
204
基于数据表示生成至少一个遗传病症簇
206

图2a

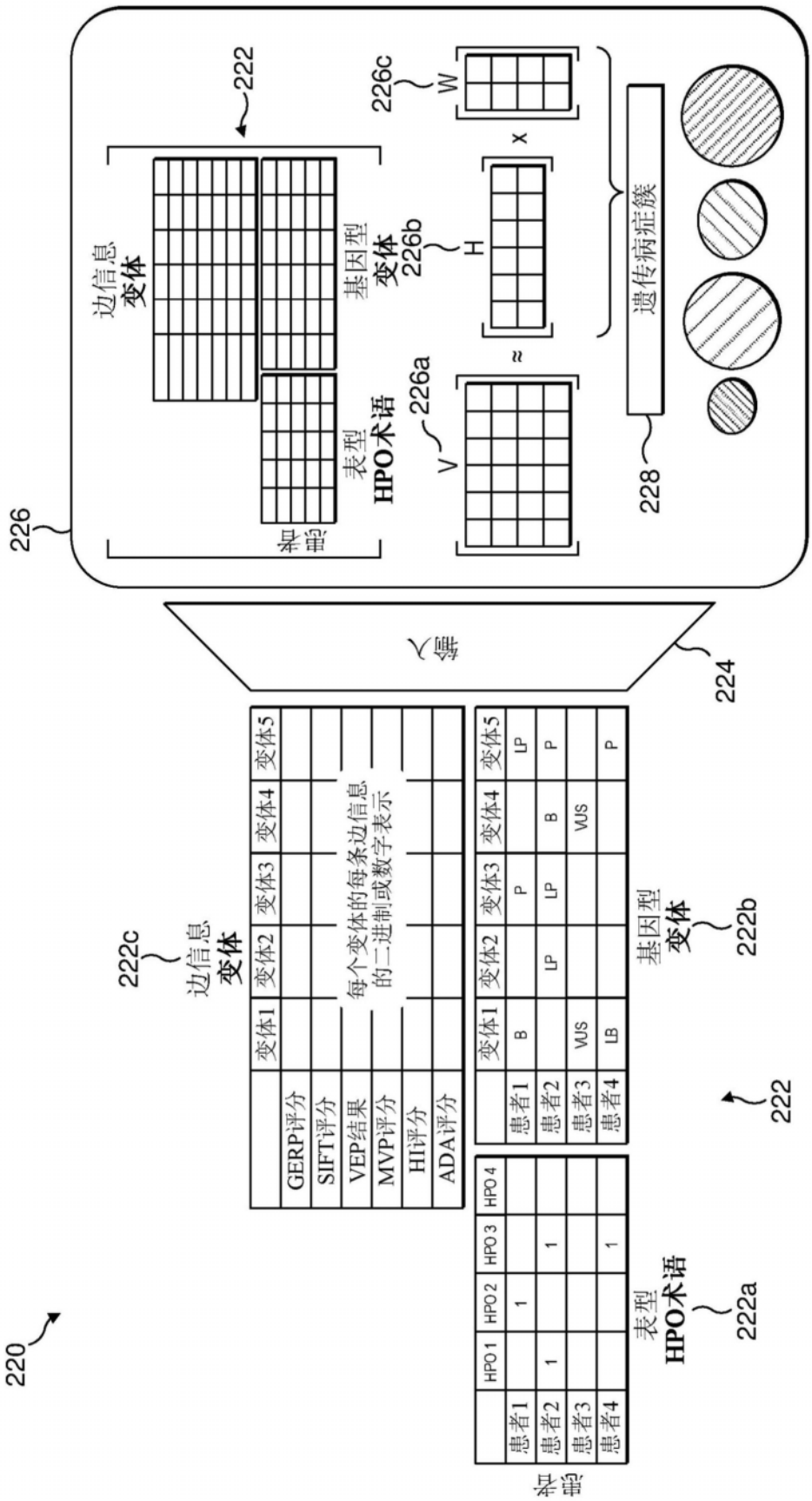

220
222
222a
HPO术语
表型
222b
基因型
变体
222c
边信息
变体
224
226
226a
226b
226c
228
遗传病症簇
每个变体的每条边信息的二进制或数字表示
GERP评分
SIFT评分
VEP结果
MVP评分
H1评分
ADA评分
患者1
患者2
患者3
患者4
HPO1
HPO2
HPO3
HPO4
变体1
变体2
变体3
变体4
变体5

图2b

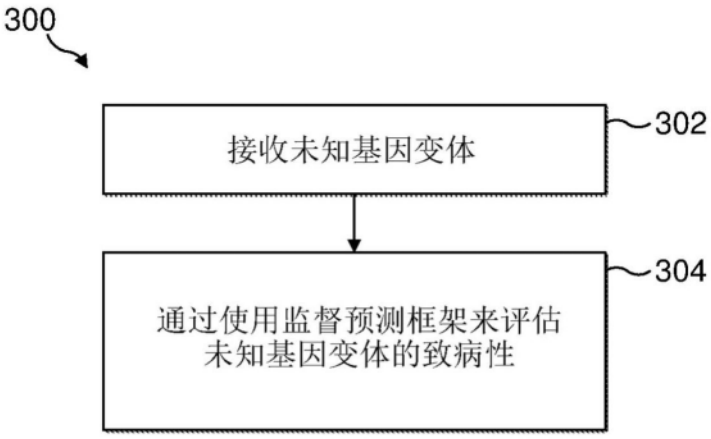


图3

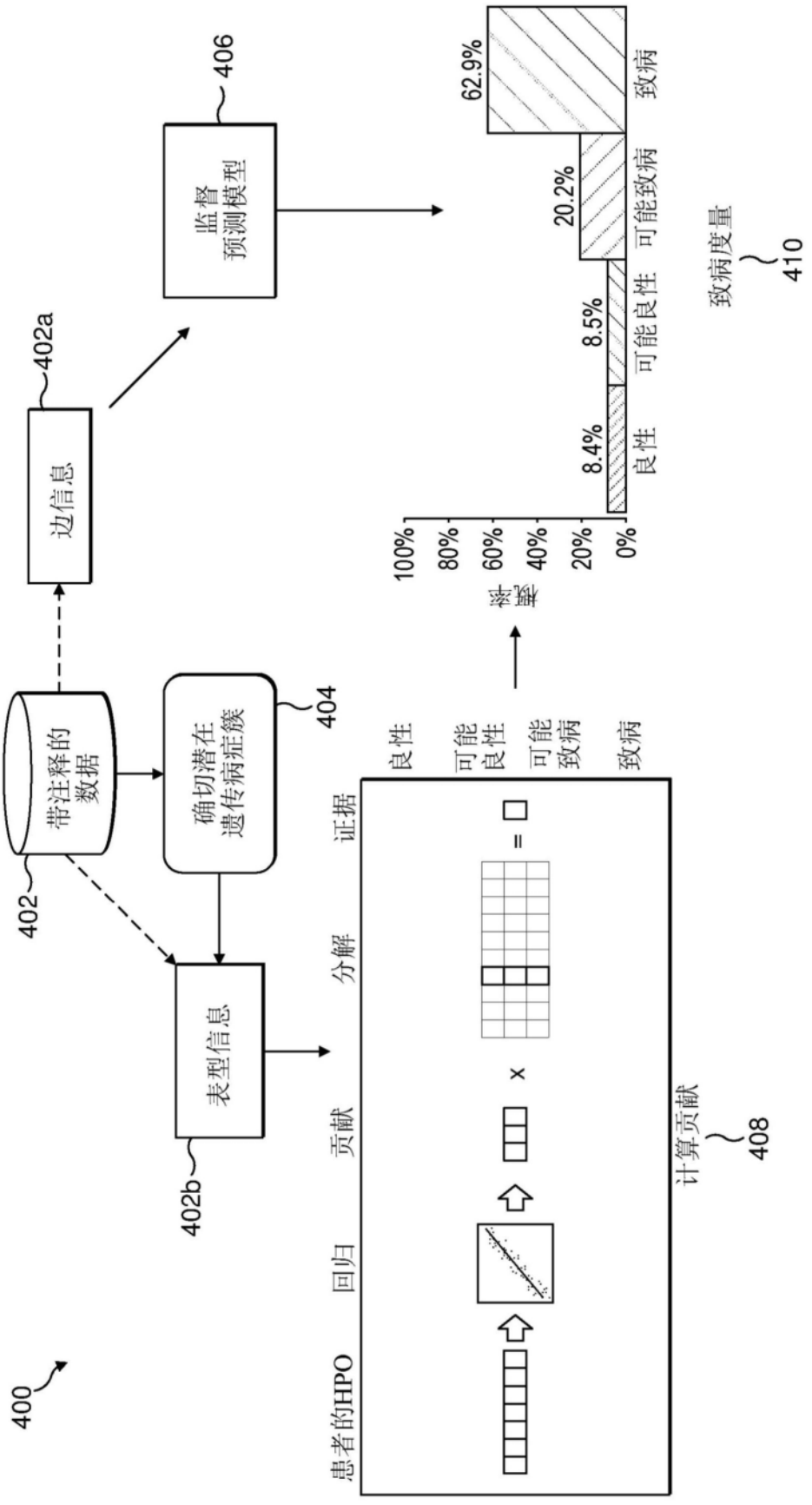

400
患者的HPO
回归
贡献
分解
证据
良性
可能良性
可能致病
致病
计算贡献
408
表型信息
402b
确切潜在遗传病症簇
404
带注释的数据
402
边信息
402a
监督预测模型
406
100%
80%
60%
40%
20%
0%
概率
8.4%
8.5%
20.2%
62.9%
良性
可能良性
可能致病
致病
致病度量
410

图4

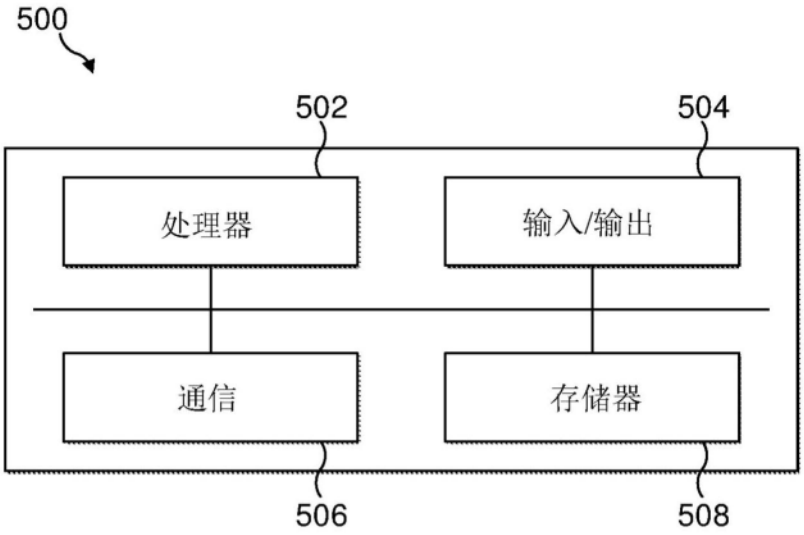


图5