



(12) 发明专利

(10) 授权公告号 CN 113129999 B

(45) 授权公告日 2024. 06. 18

(21) 申请号 202011621885.9

(22) 申请日 2020.12.30

(65) 同一申请的已公布的文献号
申请公布号 CN 113129999 A

(43) 申请公布日 2021.07.16

(30) 优先权数据
10-2019-0179469 2019.12.31 KR
10-2020-0177206 2020.12.17 KR

(73) 专利权人 高丽大学校产学协力团
地址 韩国首尔市城北区安岩路145高丽大
学校(安岩洞五街)

(72) 发明人 姜在雨 全敏志 张富禄 朴贞洙
朴星俊 金善奎

(74) 专利代理机构 北京同立钧成知识产权代理
有限公司 11205
专利代理师 杨文娟 臧建明

(51) Int.Cl.
G16B 15/30 (2019.01)
G16C 20/50 (2019.01)
G16H 70/40 (2018.01)

(56) 对比文件
US 2019114390 A1, 2019.04.18
审查员 王思佳

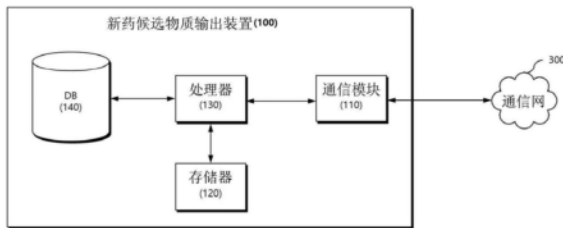
权利要求书2页 说明书8页 附图3页

(54) 发明名称

新药候选物质输出方法及装置、模型构建方法
与记录介质

(57) 摘要

本发明涉及一种新药候选物质的输出方法及装置、药物学习模型构建方法与记录介质。根据本发明一实施例的新药候选物质输出装置包括:通信模块;存储器,存储新药候选物质输出程序;处理器,实行所述新药候选物质输出程序,所述新药候选物质输出程序提供如下药物学习模型、即针对化学合成物的化学结构的嵌入向量与针对由各化学合成物诱导的转录量的变化信息的嵌入向量位于相同的向量空间的药物学习模型,并输出与输入到所述药物学习模型的针对新型物质的化学结构的嵌入向量匹配的转录量变化信息结果,或者输出与成为输入到所述药物学习模型的靶标的转录量变化信息匹配的一个以上的药物的信息。



1. 一种新药候选物质输出装置,在新药候选物质输出装置中,包括:
通信模块;
存储器,存储新药候选物质输出程序;
处理器,实行所述新药候选物质输出程序,

所述新药候选物质输出程序提供如下药物学习模型、即针对化学合成物的化学结构的嵌入向量与针对由各化学合成物诱导的转录量变化信息的嵌入向量位于相同的向量空间的药物学习模型,并输出与输入到所述药物学习模型的针对新型物质的化学结构的嵌入向量匹配的转录量变化信息结果,或者输出与成为输入到所述药物学习模型的靶标的转录量变化信息匹配的一个以上的药物的信息,所述转录量变化信息是指转录量在施用化学合成物前和施用化学合成物后的变化,

其中所述药物学习模型以如下方式构成:通过三重态损失函数将表示转录量变化信息的第一嵌入向量与诱导相应转录量变化的化学合成物的嵌入向量之间的距离最小化,且将所述第一嵌入向量与并不诱导相应转录量的变化的化学合成物的嵌入向量之间的距离最大化。

2. 根据权利要求1所述的新药候选物质输出装置,其中

所述药物学习模型在输入关于所述化学合成物的给药前的转录量的数据时进行重复学习,以将所述药物学习模型输出的变化的转录量的数据与实际化学合成物给药后转录量的数据之间的差异最小化。

3. 一种药物学习模型构建方法,是用于发现新药候选物质的新药候选物质输出装置的药物学习模型构建方法,包括如下步骤:

构建药物学习模型,所述药物学习模型使针对化学合成物的化学结构的嵌入向量与针对由各化学合成物诱导的转录量变化信息的嵌入向量配置在相同的向量空间,所述转录量变化信息是指转录量在施用化学合成物前和施用化学合成物后的变化;以及

在将关于所述化学合成物的给药前的转录量的数据输入到所述药物学习模型时执行重复学习,以将所述药物学习模型推导出的转录量的数据与实际化学合成物给药后转录量的数据之间的差异最小化,

其中构建所述药物学习模型的步骤为:通过三重态损失函数将表示转录量变化信息的第一嵌入向量与诱导相应转录量变化的化学合成物的嵌入向量之间的距离最小化,且将所述第一嵌入向量与并不诱导相应转录量的变化的化学合成物的嵌入向量之间的距离最大化。

4. 一种新药候选物质输出方法,是使用新药候选物质输出装置的新药候选物质输出方法,包括如下步骤:

提供药物学习模型,所述药物学习模型中针对化学合成物的化学结构的嵌入向量与针对由各化学合成物诱导的转录量变化信息的嵌入向量位于相同的向量空间,所述转录量变化信息是指转录量在施用化学合成物前和施用化学合成物后的变化;以及

将针对新型物质的化学结构的嵌入向量或成为靶标的转录量变化信息输入到所述药物学习模型;

所述药物学习模型输出与针对所述新型物质的化学结构的嵌入向量匹配的转录量变化信息结果,或者输出与成为所述靶标的转录量变化信息匹配的一个以上的药物的信息,

其中所述药物学习模型以如下方式构成:通过三重态损失函数将表示转录量变化信息的第一嵌入向量与诱导相应转录量变化的化学合成物的嵌入向量之间的距离最小化,且将所述第一嵌入向量与并不诱导相应转录量的变化的化学合成物的嵌入向量之间的距离最大化。

5.根据权利要求4所述的新药候选物质输出方法,其中

所述药物学习模型在输入关于所述化学合成物的给药前的转录量的数据时进行重复学习,以将所述药物学习模型输出的变化的转录量的数据与实际化学合成物给药后转录量的数据之间的差异最小化。

6.一种计算机可读存储介质,其上存储有程序,所述程序被处理器执行时实现如权利要求3至5中任一项所述的方法的步骤。

新药候选物质输出方法及装置、模型构建方法与记录介质

技术领域

[0001] 本发明涉及一种利用转录组表现型(transcriptome phenotype)以导出新药候选物质的新药候选物质的输出方法及装置、药物学习模型构建方法与记录介质。

背景技术

[0002] 新药通过包括新药发现步骤与新药开发步骤的过程来开发。新药发现步骤包括靶标确认、候选物质设计、功效测定、候选药物选择。在新药开发步骤中包括安全性评估及候选药物的临床试验。尽管在通过新药发现步骤与新药开发步骤将药品商用化时要花费大量的时间与成本,但是已知其成功率并不高。

[0003] 在新药开发管线(pipe line)中,确定适合于疾病的靶蛋白(target protein)并找到与靶标结合的分子非常重要。一旦确定关于疾病的靶标,则可通过高效筛选(screening)发现能够与靶标结合的化合物,且还能够选择与靶标结合的药物的结构类似物作为药物候选物质。

[0004] 如此选定约5000~10000个以上作为药物候选物质,但是由于经过实验与验证直至出售为止的成功率小于0.02%,因此需要大量新药开发的成本与开发时间。

[0005] 如上所述,新药开发过程不仅需要大量时间与成本而且是困难的过程,并且不能保证所开发的新药实际上会成功。此外,制药业的研发成本不断增加,自1950年代以来,按研发成本对新批准医药品的数量的比率计算的生产率每年持续下降。由于新药开发的成功取决于新药候选选择,因此选择成功概率高的候选新药以提高新药开发的生产率是重要的。

[0006] 在新药发现步骤中,用于发现新药候选物质的传统方式是以靶标为中心发现新药候选物质的方法。此要进行如下过程:作为揭示发病相关主要因素的过程的靶蛋白发现、作为找到可与靶蛋白物理结合并抑制机能的化合物的过程的苗头物质发现(Hit discovery)、对前述所找到的苗头物质进行结构上的最优化的先导物质最优化过程(Lead optimization)。在从靶蛋白中发现的药物候选组中,通过开发(Development)步骤最终挑选出对细胞、组织乃至个体有效果的药物。

[0007] 然而,以下方面是以靶标为中心的新药开发过程的限制:1)靶蛋白假设是必不可少的,2)即使找到靶蛋白假设,如果化合物具有不能结合到靶蛋白的结构(无成药性靶标(undruggable target)),则也难以进行药物开发,3)即使是化合物能够结合到靶蛋白的结构,也难以实验验证无数种化合物,且要花费约5.5年以上的相当久的时间以导出候选物质。具体示例如下所示。

[0008] 例如,在现有的以靶标为中心的药物开发方式中,首先设定靶蛋白,并探索与相应蛋白质结合的化合物。但是,如果由于对相应疾病的了解度不高而无法鉴定靶蛋白,则无法开始可对相应疾病进行治疗的新药开发过程。假设在阿尔茨海默氏病(Alzheimer's disease)的情况下,由于无法揭示出作为相应疾病的重要原因发挥作用的明确的靶蛋白,因此很难探索出新药候选物质。

[0009] 另外,即使已知对相应疾病治疗发挥重要作用的靶蛋白,但实际上由于不能制备与所述蛋白质结合的物质情况,因此药物开发仍困难重重。作为肺癌或结肠癌的靶蛋白而被熟知的KRAS由于没有药物可结合的结合位点(binding site),因此目前尚无KRAS抑制剂(inhibitor)。这意味着无论对相应疾病的了解度如何都不可能通过现有的以靶标为中心的方式来开发针对此疾病的新药。

[0010] [现有技术文献]

[0011] [专利文献]

[0012] [专利文献1]韩国公开专利公报第10-2018-0058648号(发明的名称:以非结构-结构转变部位为目标的新药候选物质发现方法及新药候选物质发现装置)

发明内容

[0013] [发明所要解决的问题]

[0014] 本发明为了解决所述的问题,目的在于提供一种新药候选物质输出装置及方法,所述新药候选物质输出装置根据本发明的一实施例可通过转录量变化数据及输入各化学合成物数据进行学习的药物学习模型来输出新药候选物质。

[0015] 但是,本实施例欲实现的技术课题并不限于如上所述的技术课题,可存在又一些技术课题。

[0016] [解决问题的技术手段]

[0017] 作为用于达成所述的技术课题的技术手段,根据本发明一实施例的新药候选物质输出装置包括:通信模块;存储器,存储新药候选物质输出程序;处理器,实行所述新药候选物质输出程序,所述新药候选物质输出程序提供如下药物学习模型、即针对化学合成物的化学结构的嵌入向量(embedding vector)与针对由各化学合成物诱导的转录量的变化信息的嵌入向量位于相同的向量空间的药物学习模型,并输出与输入到所述药物学习模型的针对新型物质的化学结构的嵌入向量匹配的转录量变化信息结果,或者输出与成为输入到所述药物学习模型的靶标的转录量变化信息匹配的一个以上的药物的信息。

[0018] 另外,根据本发明的另一实施例的用于发现新药候选物质的学习模型构建方法包括如下步骤:构建药物学习模型,所述药物学习模型使针对化学合成物的化学结构的嵌入向量与针对由各化学合成物诱导的转录量的变化信息的嵌入向量配置在相同的向量空间;以及在将关于所述化学合成物的给药前的转录量的数据输入到所述药物学习模型时执行重复学习,以将所述药物学习模型推导出的转录量的数据与实际化学合成物给药后转录量的数据之间的差异最小化。

[0019] 另外,根据本发明的另一实施例的新药候选物质输出方法包括如下步骤:提供药物学习模型,所述药物学习模型中针对化学合成物的化学结构的嵌入向量与针对由各化学合成物诱导的转录量的变化信息的嵌入向量位于相同的向量空间;以及将针对新型物质的化学结构的嵌入向量或成为靶标的转录量变化信息输入到所述药物学习模型;所述药物学习模型输出与针对所述新型物质的化学结构的嵌入向量匹配的转录量变化信息结果,或者输出与成为所述靶标的转录量变化信息匹配的一个以上的药物的信息。

[0020] [发明的效果]

[0021] 根据所述的本发明的课题解决手段,可大大缩短发现新药候选物质的筛选步骤所

需的时间与成本。

[0022] 另外,即使在不存在靶蛋白假设或已知靶蛋白但实际上不能制造与此蛋白质结合的物质,的情况下也可开发新药。

附图说明

[0023] 图1是示出根据本发明一实施例的新药候选物质输出装置的构成的图。

[0024] 图2是用于说明根据本发明一实施例的新药候选物质输出装置的操作的例示图。

[0025] 图3是用于说明根据本发明一实施例的新药候选物质输出方法的操作的顺序图。

[0026] [符号的说明]

[0027] 100:新药候选物质输出装置

[0028] 110:通信模块

[0029] 120:存储器

[0030] 130:处理器

[0031] 140:数据库

具体实施方式

[0032] 在下文中,将参考附图详细地对本申请的实施例进行说明,使得本申请所属技术领域内技术人员可容易地实施。然而,本申请可以各种不同的形态来实现,且不限于在此说明的实施例。并且在附图中,为了明确地说明本申请,省略了与说明无关的部分,且通过整个说明书对相似的部分赋予相似的附图符号。

[0033] 在本申请的整个说明书中,在记载为某一部分与另一部分“连接”时,这不仅包括“直接连接”的情况,而且还包括在其等中间隔以其他元件而“电连接”的情况。

[0034] 在本申请的整个说明书中,在记载为某一部件位于另一部件“上”时,这不仅包括某一部件与另一部件接触的情况,还包括两个部件之间存在又一部件的情况。

[0035] 以下,将参照附图详细地对本发明的一实施例进行说明。

[0036] 图1是示出根据本发明一实施例的新药候选物质输出装置的构成的图。

[0037] 如图所示,新药候选物质输出装置100可包括通信模块110、存储器120、处理器130及数据库140。新药候选物质输出装置100基于计算(computing)装置来构成,且更包括未图示的电源部、各种输入装置及输出装置等。

[0038] 通信模块110可通过通信网300与外部的计算装置发送/接收关于化学合成物的化学结构的数据、或者与之匹配的关于由化学合成物诱导的转录量的变化信息的数据。通信模块110可为如下装置:包括用于与另一网络装置通过有线/无线连接来发送/接收如控制信号或数据信号等信号所需的硬件及软件。

[0039] 在存储器120中存储新药候选物质输出程序。新药候选物质输出程序提供一种药物学习模型,其中针对化学合成物的化学结构的嵌入向量与针对由各化学合成物诱导的转录量的变化信息的嵌入向量位于相同的向量空间,并基于通过输入装置或通信模块110等输入的查询(query),输出新药候选物质。此时,作为输入的查询可为有关新型物质的化学结构的数据或是成为靶标的转录量变化信息。

[0040] 另外,新药候选物质输出程序为构建药物学习模型的逻辑,但可另外执行对所构

建的药物学习模型执行新的学习过程的学习模型更新过程等。

[0041] 在此种存储器120中存储在用于驱动新药候选物质输出装置100的操作系统(operating system)或丢失数据预测程序的实行过程中产生的各种种类的数据。

[0042] 此时,存储器120是即使不供应电源也继续保持所存储的信息的非易失性存储装置及需要电力维持所存储的信息的易失性存储装置的通称。

[0043] 另外,存储器120可执行临时或永久存储由处理器130处理的数据的功能。此处,存储器120除了需要电力来维持所存储的信息的易失性存储装置之外还可包括磁存储介质(magnetic storage media)或快闪存储介质(flash storage media),但本发明的范围并不限于此。

[0044] 处理器130实行存储在存储器120中的程序,并控制实行新药候选物质输出程序的整个过程。针对处理器130执行的各个操作稍后将更详细地进行阐述。

[0045] 此种处理器130可包括能够处理数据的所有种类的装置。例如,可能意指内置于硬件中的数据处理装置,所述数据处理装置具有为执行由包含于程序内的代码(code)或命令表达的功能而在物理上结构化的电路。作为内置于如上所述硬件中的数据处理装置的一例,可包括微处理器(microprocessor)、中央处理器(central processing unit,CPU)、处理器核(processor core)、多处理器(multiprocessor)、应用专用集成电路(application-specific integrated circuit,ASIC)、现场可编程门阵列(field programmable gate array,FPGA)等处理装置,但是本发明的范围并不限于此。

[0046] 数据库140根据处理器130的控制存储或提供新药候选物质输出装置所需的数据。此种数据库140可与存储器120作为单独的构成要素而包括在内,或也可构建在存储器120的一部分区域中。

[0047] 图2是用于说明根据本发明一实施例的新药候选物质输出装置的操作的例示图,图3是用于说明根据本发明一实施例的新药候选物质输出方法的操作的顺序图。

[0048] 首先,将阐述构建新药候选物质输出装置100的药物学习模型的方法(S310)。

[0049] 新药候选物质输出装置100的药物学习模型可包括:学习转录量变化信息的多阶层的人工神经网络;以及学习化学合成物的化学结构信息的多阶层的人工神经网络。

[0050] 各个阶层可由基本的感知器(perceptron)阶层(完全连接阶层(Fully-connected layer))组成,且为了生成以化学合成物的化学结构为中心的嵌入向量,可使用图神经网络(Graph Neural Network,GNN)等。例如,学习转录量变化信息的多阶层的人工神经网络与学习化学合成物的化学结构信息的多阶层的人工神经网络均具有4个阶层,且可以通过输出相同维度的嵌入向量而生成的完全连接的人工神经网络的形态实现。

[0051] 新药候选物质输出装置100的药物学习模型使针对化学合成物的化学结构的嵌入向量与针对由各化学合成物诱导的转录量的变化信息的嵌入向量配置在相同的向量空间。

[0052] 转录量变化信息是指药物给药前后的细胞内的多个基因的表达度(转录量)的变化。在药物给药前后的转录量变化中包括关于如下的信息:药物对细胞带来的影响、即药物与细胞内的所有蛋白质之间以及蛋白质与蛋白质之间的复杂相互作用。此时,转录量的变化程度可定义为从由给药前各基因的转录量(基因表达(gene-expression))组成的高斯(Gaussian)分布的平均值到给药后各基因的转录量的距离。即,变化程度与距离值成比例地增加。此时,可通过如下所述的高斯函数获得两个转录量之间的距离。

$$[0053] \quad K(x_g, m_g) = \exp\left(-\frac{\|x_g - m_g\|^2}{2\sigma_g^2}\right)$$

[0054] m_g 是指给药前基因g的转录量值的平均值, σ_g 是指给药前基因g的转录量值的样品平均值, x_g 是指给药后基因g的转录量值。

[0055] 例如,可考虑包含有关向癌细胞株给药时转录量的变化信息的学习数据,并假定其中包含978个代表基因。此时,由特定药物的给药产生的转录量的变化信息包括从总共978个基因类别的各基因的转录组的高斯分布平均值到给药后各基因的转录量的距离值。

[0056] 由此,可解决使用现有的以靶标为中心的新药开发方式无法进行新药开发的问题。例如,在不存在疾病相关的靶蛋白假设的情况下,可发现可将患者组的基因表达模式诱导为正常组的基因表达模式的药物候选物质。另外,即使已知疾病相关的靶蛋白假设而由于靶蛋白不能结合而使药物开发困难的情况,也可发现能够诱导由于靶基因敲低(Knock down)引起的基因表达度变化的候选物质。

[0057] 如图所示,将用于化学合成物的化学结构的嵌入向量与用于将相应化学合成物作为药物注射时发生的转录量的变化信息的嵌入向量配置在相同的向量空间(例如,单位超球(unit hypersphere)的表面),从而构建药物学习模型(210)。

[0058] 此时,在药物给药前后的转录量变化信息中包括与包含脱靶(off-target)影响的药物反应有关的所有信息,因此通过以使转录量变化信息与诱导此变化的药物位于同一向量空间的相同位置处的方式进行嵌入,可捕获有关药物效果的抽象特性。另外,由于是基于药物效果的嵌入模块,因此即使具有新化学结构的药物也可基于药物特性被发现为新药候选物质。另外,由于本发明的嵌入模块使用药物的化学结构作为输入,因此不仅对现有已知结构的化学合成物,而且对将来可知的所有可合成化合物而言都可通过基于药物效果的嵌入来表达,而不产生单独的过程与成本。

[0059] 另外,本发明的药物学习模型在相同向量空间构建针对化学合成物的化学结构的嵌入向量与关于在将相应化学合成物作为药物注射时发生的转录量的变化信息的嵌入向量,作为关于此的损失函数,使用三重态损失(triplet loss)函数。在本损失函数中,使用表示根据特定化学合成物的给药而发生的转录量的变化信息的嵌入向量(f^a)、表示诱导相应转录量变化的化学合成物的嵌入向量(f^p)、表示不会诱导相应转录量变化的化学合成物的嵌入向量(f^n)。Triplet_loss(锚点(Anchor)、正向(Positive)、负向(Negative))

$$[0060] \quad Loss = \sum_{i=1}^N \max(Dist(f_i^a, f_i^p) - Dist(f_i^a, f_i^n) + \alpha, 0)$$

[0061] 即,三重态损失函数是基于如下值来计算:所述值是从表示转录量的变化信息的嵌入向量(f^a)与表示诱导相应转录量变化的化学合成物的嵌入向量(f^p)之间的距离减去表示转录量的变化信息的嵌入向量(f^a)与表示不诱导相应转录量变化的化学合成物的嵌入向量(f^n)之间的距离而获得。

[0062] 此时, α 表示由模型设计者设定的边际值,i表示学习次数或样品的识别编号,N表示诱导各转录量变化的各化学合成物与转录量变化信息的对的个数。例如,如果学习数据包括针对总共21,220种药物及82个细胞系(cell line)进行实验的310,114个转录量变化

信息,则N可为与整体转录量变化信息的个数对应的310,114。

[0063] 关于任意的两个嵌入向量A,B的距离Dist(A,B)使用负余弦相似度(Negative cosine similarity),如下所述般定义。

$$[0064] \quad \text{Dist}(A, B) = 1 - \frac{A \cdot B}{\|A\| \times \|B\|}$$

[0065] 三重态损失函数是以如下方式构成:将表示转录量的变化信息的嵌入向量与诱导相应转录量变化的化学合成物的嵌入向量之间的距离最小化,并将表示转录量的变化信息的嵌入向量与不诱导相应转录量的变化的化学合成物的嵌入向量之间的距离最大化。

[0066] 为了使此种本发明的损失函数的性能进一步极大化,可如下所示般使用双重三重态损失(Double triplet loss)函数。

$$[0067] \quad \text{Loss} = \sum_{i=1}^N \max(\text{Dist}(f_i^a, f_i^p) - \text{Dist}(f_i^a, f_i^n) + \alpha, 0) + \max(\text{Dist}(f_i^p, f_i^a) - \text{Dist}(f_i^p, f_i^n) + \alpha, 0)$$

[0068] 与前文所说明的三重态损失函数相比,后面进一步追加从表示诱导转录量变化的化学合成物的嵌入向量(f^p)与表示转录量的变化信息的嵌入向量(f^a)之间的距离减去表示诱导转录量变化的化学合成物的嵌入向量(f^p)与不诱导转录量变化的化学合成物的嵌入向量(f^n)之间的距离的项。

[0069] 使用前文说明的三重态损失函数,在表示转录量的变化信息的嵌入向量(f^a)与表示不诱导相应转录量变化的化学合成物的嵌入向量(f^n)之间的距离非常长的情况下,损失函数值计算为0。在此种情况下,可能发生如下情况:表示转录量的变化信息的嵌入向量(f^a)与诱导相应转录量变化的化学合成物的嵌入向量(f^p)之间的距离没有变窄。

[0070] 使用双三重态损失函数,即使位于前面的项变为0,通过在后面追加的项也可使表示诱导转录量变化的化学合成物的嵌入向量(f^p)与表示转录量的变化信息的嵌入向量(f^a)之间的距离变窄。

[0071] 另一方面,为了进一步提高如上所述般构建的学习模型,可追加执行重复学习步骤(220)。

[0072] 即,在将关于化学合成物给药前的转录量的数据输入到药物学习模型中时,执行重复学习,以使药物学习模型推导出的转录量的数据与实际化学合成物给药后的转录量的数据之间的差异最小化。为此,可通过使用规定的损失函数来执行更新学习模型的加权值的操作,以使得从学习模型输出的转录量数据与给药后的转录量数据之间的差最小。

[0073] 接下来,对如上所述般构建的学习模型,输入针对新型物质的化学结构的嵌入向量或成为靶标的转录量变化信息作为查询(S320)。

[0074] 在本发明中,对现有已知的所有化合物求得嵌入向量,并预先准备索引(index)以便可快速搜索到它们并将其存储在数据库(DB)等中。(230)基于Zinc15 DB中注册的化合物,现有已知化合物约为13亿种。实现为包括这些化合物在内,对将来追加的化合物也可追加索引。使用如上所述构建的相似向量搜索系统,可在表示为密集向量的几种化合物中找到与查询向量最相似的化合物。此时,查询向量可为针对新型物质的化学结构的嵌入向量,或者可为在药物给药后关于转录量变化信息的嵌入向量。

[0075] 例如,为了开发如药物再利用(药物重新定位(drug repositioning))等正在开发或正在销售中的药物的新医学用途,可将相应药物的嵌入向量用作查询向量,且为了如药物筛选般执行所要求的功能的新药候选物质,可将转录量变化信息用作查询。此时,对于约13亿种化合物的相似性搜索所需时间可在数秒钟很短的时间内选定药物候选组。

[0076] 接下来,药物学习模型是查询向量的输出,输出与针对新型物质的化学结构的嵌入向量匹配的转录量变化信息结果,或者输出与成为靶标的转录量变化信息匹配的一个以上的药物的信息(S330)。

[0077] 如图2所示,当输入转录量变化信息作为查询时,药物学习模型输出与之匹配的药物候选物质。另外,当输入针对新型物质的化学结构的嵌入向量时,药物学习模型输出与之匹配的转录量变化信息的结果。

[0078] 另一方面,在本发明中,如上所述,基于转录量的变化信息,可输出与之匹配的药物候选物质。特别是,在确定靶蛋白的情况下,基于敲除(knock out,KO)/敲低(knock down,KD)靶蛋白的基因之前的转录量(基准转录量)与敲除/敲低靶蛋白的基因之后的转录量(诱导转录量)的差来确定转录量变化信息。并且,通过将这样获得的转录量变化信息输入到药物学习模型中,可输出药物候选物质。

[0079] 作为另一种方法,在给出了正常组与靶标疾病组的转录量信息的情况下,基于疾病组的转录量(基准转录量)与正常组的转录量(诱导转录量)之间的差,可确定转录量变化信息。并且,通过将这样获得的转录量变化信息输入到药物学习模型中,可输出药物候选物质。

[0080] 另一方面,转录量变化量的绝对值的方差(variance)根据值的观测方法可能与模型学习中使用的转录组变化量的值的方差差得大。在将此种值照原样用作药物学习模型的输入的情况下,由于是与模型学习环境有很大不同的数据,因此可能难以获得预期的结果。为了解决此问题,需要使作为查询输入的转录量变化信息的方差与在药物学习模型的学习过程中使用的转录量变化信息的学习数据的方差大体一致。

[0081] 为此,针对构成新输入的转录量变化信息的基准转录量与诱导转录量,通过T检验(T-test)按照T统计量(T-statistic)的大小顺序列出各个基因。这与对新输入的转录组信息按照转录组变化量的大小顺序列出基因的情况相同。接下来,对于按照转录组变化量的大小顺序列出的基因,按照大的顺序映射学习数据的转录组变化量的值。例如,针对与基准转录量相比在诱导转录量中转录量减少最多的基因,在学习数据的转录组变化量中赋予负绝对值最大的值。通过此种方法,针对新输入的转录组信息,可保持转录组变化量基准基因的的顺序的同时,生成与学习数据的值的方差具有相同水平的方差的输入数据。

[0082] 在药物学习模型中,在输入已调整的转录量变化信息的嵌入向量时,将从先前构建的化合物嵌入数据库中搜索具有最接近向量值的化合物。此时,向量间的距离可使用如欧几里得距离(Euclidean distance)或余弦相似度(Cosine similarity)等一般距离函数。基于此,推导出具有最接近向量值的化合物作为诱导预期的转录量变化效果的候选药物。

[0083] 本发明的一实施例还可以如下形态实现:如由计算机实行的程序模块般包括可由计算机实行的命令语的记录介质。计算机可读介质可为可由计算机存取(access)的任何可用介质,且包括易失性介质及非易失性介质、可移除型及不可移除型介质全部。另外,计算

机可读介质可包括计算机存储介质。计算机存储介质包括利用用于存储如计算机可读命令语、数据结构、程序模块或其他数据等信息的任意方法或技术实现的易失性及非易失性、可移除型及不可移除型介质全部。

[0084] 尽管已经结合特定实施例说明了本发明的方法及系统,但是可使用具有通用硬件架构的计算机系统来实现本发明的方法及系统的一部分或全部构成要素或操作。

[0085] 所述的本申请的说明仅用于例示,且本申请所属技术领域内技术人员将能够理解,可在不改变本申请的技术思想或必要特征的情况下容易地变形为其他具体形态。因此,应当理解,以上记述的实施例在所有方面均为例示性的而非限制性的。例如,说明为单一型的各构成要素也可以分布式方式实施,并且同样地说明为分布式的构成要素也可以组合形态实施。

[0086] 本申请的范围由以上权利要求来表示而并非所述详细的说明,且权利要求的含义及范围以及根据其等同概念导出的所有变更或变形的形态应被解释为包含于本申请的范围

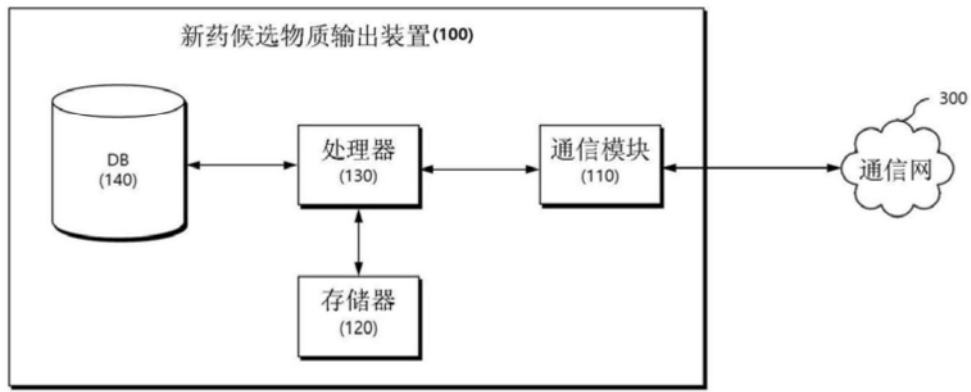


图1

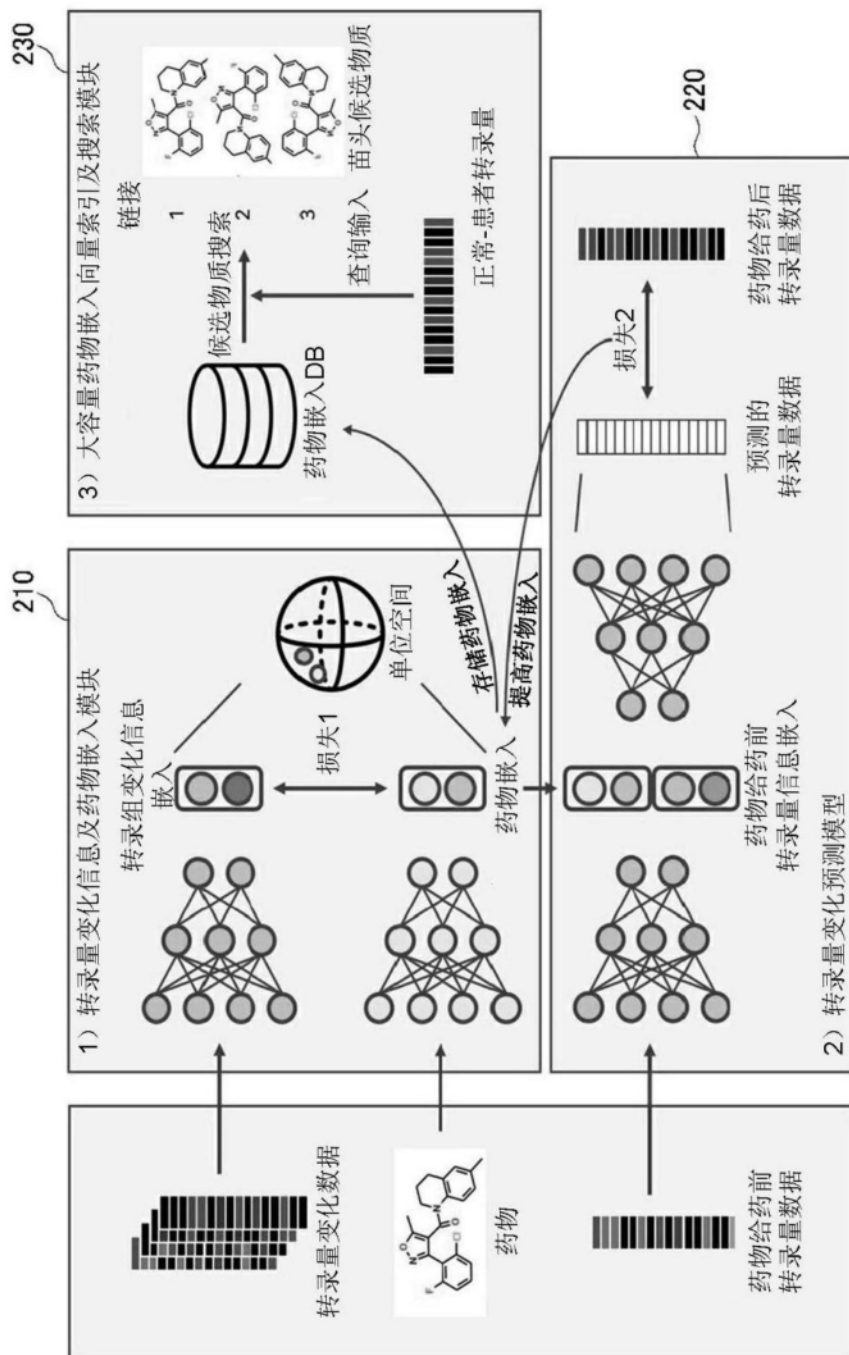


图2

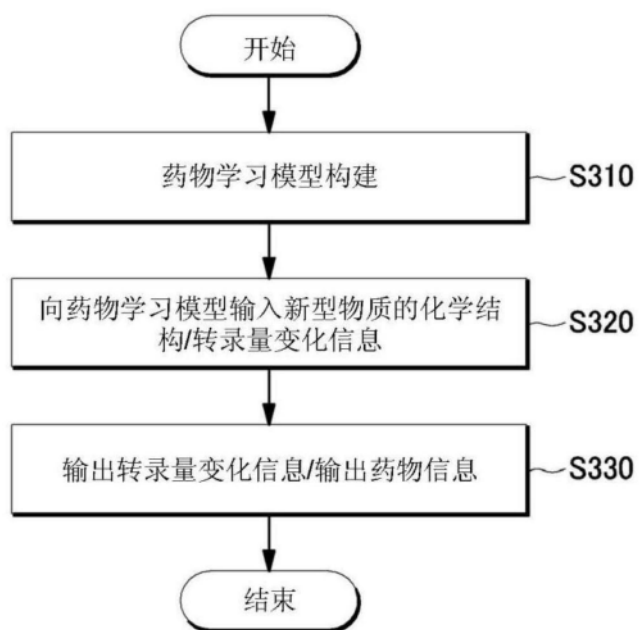


图3