(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 119731310 A
(43)申请公布日 2025.03.28

(21)申请号 202380059794.6
(22)申请日 2023.07.07
(30)优先权数据
2022-138789 2022.08.31 JP
(85)PCT国际申请进入国家阶段日
2025.02.14
(86)PCT国际申请的申请数据
PCT/JP2023/025263 2023.07.07
(87)PCT国际申请的公布数据
WO2024/048079 JA 2024.03.07
(71)申请人 富士胶片株式会社
地址 日本
(72)发明人 梅川正夫 铃木贵文 佐藤政宽 长濑雅也 松浦达也 村上裕太
(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 王灵菇
(51)Int.Cl.
C12N 5/10(2006.01)
C12N 5/071(2006.01)
C12P 1/00(2006.01)
C12P 21/08(2006.01)
G16B 40/00(2006.01)

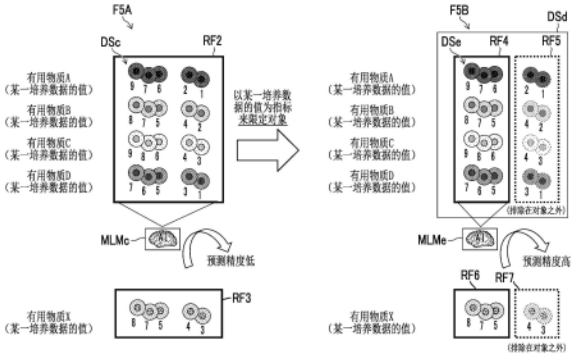
权利要求书2页 说明书18页 附图13页

(54)发明名称

预测产生有用物质的克隆的产生稳定性的方法、信息处理装置、程序及预测模型生成方法

(57)摘要

本发明提供一种能够以高精度且低成本预测产生有用物质的克隆的产生稳定性的方法、信息处理装置、程序及预测模型生成方法。1个以上的处理器执行以下处理:针对产生有用物质的克隆获取1种以上的克隆的培养数据;分析培养数据来限定作为预测对象的克隆;及使用针对作为预测对象的克隆测定出的数据来预测基于作为预测对象的克隆的有用物质的产生稳定性。产生稳定性可以根据培养开始时和规定期间培养后的有用物质的产生量的变化的有无来定义。



CN 119731310 A

1.一种预测产生有用物质的克隆的产生稳定性的方法,其中,

1个以上的处理器执行以下处理:

获取1种以上的所述克隆的培养数据;

分析所述培养数据来限定作为预测对象的克隆;及

使用针对所述作为预测对象的克隆测定出的数据来预测基于所述作为预测对象的克隆的所述有用物质的产生稳定性。

2.根据权利要求1所述的方法,其中,

所述产生稳定性根据培养开始时和规定期间培养后的所述有用物质的产生量的变化有无来定义。

3.根据权利要求1所述的方法,其中,

所述1个以上的处理器进行以下处理:

设定从所述培养数据获得的指标和与所述指标相关的阈值;并

根据所述指标的值和所述阈值来限定所述预测对象。

4.根据权利要求3所述的方法,其中,

所述阈值被调整为所述产生稳定性的预测精度与不限定所述预测对象的情况相比更高。

5.根据权利要求3所述的方法,其中,

所述阈值使用关于所述指标的值的位次来定义。

6.根据权利要求3所述的方法,其中,

所述预测对象为所述指标的值的上位群体。

7.根据权利要求3至6中任一项所述的方法,其中,

所述指标为所述有用物质的产生量。

8.根据权利要求3至6中任一项所述的方法,其中,

所述指标为积分活细胞密度。

9.根据权利要求3至6中任一项所述的方法,其中,

所述指标为乳酸浓度。

10.根据权利要求1至6中任一项所述的方法,其中,

用于所述产生稳定性的预测的所述数据包括1个以上的基因表达水平。

11.根据权利要求1至6中任一项所述的方法,其中,所述1个以上的处理器进行以下处理:

接收所述预测对象的所述数据的输入,并使用进行稳定或不稳定的2类分类的模型来预测所述产生稳定性。

12.根据权利要求11所述的方法,其中,

所述模型是通过使用了多个训练数据的机器学习来训练的模型,所述多个训练数据是针对与所述作为预测对象的克隆进行了相同的限定的训练用克隆的所述数据和正解的稳定性标签建立关联的多个训练数据。

13.根据权利要求12所述的方法,其中,

所述多个训练数据包括针对所产生的有用物质不同的多种克隆的所述训练数据,

所述1个以上的处理器预测针对产生与使用于所述模型的训练的有用物质不同的有用

物质的克隆的产生稳定性。

14.根据权利要求1至6中任一项所述的方法,其中,

所述有用物质为作为医药品原料的蛋白质、肽及病毒中的任一种。

15.根据权利要求1至6中任一项所述的方法,其中,

所述有用物质为抗体或抗体样蛋白。

16.根据权利要求1至6中任一项所述的方法,其中,

所述克隆为源自脊椎动物的细胞。

17.根据权利要求1至6中任一项所述的方法,其中,

所述克隆为源自哺乳类的细胞。

18.根据权利要求1至6中任一项所述的方法,其中,

所述克隆为CHO细胞或HEK细胞。

19.一种信息处理装置,其具备:

1个以上的处理器;及

1个以上的存储装置,存储使所述1个以上的处理器执行的命令,

所述1个以上的处理器进行以下处理:

针对产生有用物质的克隆,获取1种以上的克隆的培养数据;

分析所述培养数据来限定作为预测对象的克隆;并

使用针对所述作为预测对象的克隆测定出的数据,预测基于所述作为预测对象的克隆的所述有用物质的产生稳定性。

20.一种程序,其使计算机实现以下功能:

针对产生有用物质的克隆获取1种以上的克隆的培养数据;

分析所述培养数据来限定作为预测对象的克隆;及

使用针对所述作为预测对象的克隆测定出的数据来预测基于所述作为预测对象的克隆的所述有用物质的产生稳定性。

21.一种记录介质,其为非临时性且计算机可读取的记录介质,且记录有权利要求20所述的程序。

22.一种预测模型生成方法,其生成使计算机实现预测产生有用物质的克隆的产生稳定性的功能的预测模型,在所述预测模型生成方法中,

包含1个以上的处理器的系统包括以下步骤:

获取1种以上的所述克隆的培养数据;

分析所述培养数据来限定作为预测对象的克隆;及

使用针对属于所述预测对象的克隆测定出的数据和正解的稳定性标签建立了关联的多个训练数据来进行机器学习,并以使相对于所述数据的输入的所述预测模型的输出接近所述正解的稳定性标签的方式,训练所述预测模型。

预测产生有用物质的克隆的产生稳定性的方法、信息处理装置、程序及预测模型生成方法

技术领域

[0001] 本发明涉及一种预测产生有用物质的克隆的产生稳定性的信息处理技术及机器学习技术。

背景技术

[0002] 近年来,正在推进使细胞制作在以往的化学合成中难以制作的复杂的有用物质的制造法的产业利用。其中一个例子为生物医药品,在全球医药品销售额排名前10中,半数以上的品种数占据约三分之二的销售额。生物医药品与以往的低分子医药品相比,利用了复杂的蛋白质等,人工化学合成非常困难。因此,作为生物医药品的一例的抗体医药品,例如将与所期望的人类蛋白质对应的基因插入到CHO细胞(Chinese Hamster Ovary cells:中国仓鼠卵巢细胞)等中,通过细胞功能产生所期望的蛋白质,并对其进行提取及纯化来制造抗体医药品的生产方法广泛普及。

[0003] 如上所述的基因向细胞的插入无法进行精细的控制,因此通常向大量的细胞同时插入基因。此时,考虑到所生成的各个细胞的基因插入位置是随机的,为了稳定化作为医药品的抗体并进行质量保证,许多监管机构要求在基因插入后负责抗体产生的细胞源自单一细胞,且通过传代培养使其性质不发生变化,即所谓的单克隆性。

[0004] 因此,从基因的插入位置为随机的各个细胞提取单一细胞,使该单一细胞增殖而制作细胞克隆(以下,称为克隆),通过使该克隆产生抗体来确保单克隆性。本发明的克隆是指基因上相同的细胞的群体或构成该群体的细胞。

[0005] 另一方面,在产业化中,要求具有优质的抗体产生能的克隆。在此,优质的抗体产生能是指,在当前时点具有高抗体产生能力及在长期培养期间抗体产生能力也稳定。如上所述,由基因的插入位置为随机的各个细胞制作的克隆的抗体产生能力存在偏差,需要对每个克隆判别是否为优质的抗体产生能。在当前时点是否为抗体产生能力高的高产克隆能够通过2周的标准试验来判别,但关于在长期培养期间抗体产生能力是否稳定的产生稳定性的判别,实际上需要基于数个月左右的长期培养的实验验证(稳定性试验)。

[0006] 在这种背景下,在专利文献1中,提出了一种根据在当前时点获得的克隆的基因表达数据来预测数个月之后的克隆的重组蛋白质的产生稳定性的方法。并且,在非专利文献1中,提出了一种在克隆开发的早期阶段鉴定能够预测重组蛋白质的稳定表达的标记基因,并在克隆开发的早期阶段预测重组蛋白质的产生稳定性的方法。

[0007] 现有技术文献

[0008] 专利文献

[0009] 专利文献1:国际公开第2016/075216号

[0010] 非专利文献

[0011] 非专利文献1:Uros Jamnikar,Petra Nikolic,Ales Belic,Marjanca Blas,Dominik Gaser,Andrej Francky,Holger Laux,Andrej Blejec,Spela Baebler and

Kristina Gruden,“Transcriptome study and identification of potential marker genes related to the stable expression of recombinant proteins in CHO clones” BMC Biotechnology volume 15,Article number 98(2015).

发明内容

[0012] 发明要解决的技术课题

[0013] 但是,专利文献1中所记载的方法在预测精度方面不能说是充分的。并且,由于针对多个克隆的基因分析等通常需要高额的费用,因此还存在通过预测重组蛋白质的产生稳定性而获得的降本效果因用于预测的基因分析等引起的成本上升而减退的问题。为了抑制成本,考虑缩小作为产生稳定性的预测对象的克隆数,但这样一来,预测对象中的产生稳定性高的克隆数也会减少,其结果会导致得到的产生稳定性高的克隆数变少,单纯地缩小作为预测对象的克隆数也变得困难。

[0014] 本发明要解决的第1课题在于提供一种以高精度预测克隆中的有用物质的产生稳定性的方法。第2课题在于提供一种降低克隆中的有用物质的产生稳定性的预测成本的方法。

[0015] 本发明是鉴于这种情况而完成的,其目的在于提供一种能够以高精度且低成本预测产生有用物质的克隆的产生稳定性的方法、信息处理装置、程序及预测模型生成方法。

[0016] 用于解决技术课题的手段

[0017] 本发明的第1方式所涉及的方法为预测产生有用物质的克隆的产生稳定性的方法,其中,1个以上的处理器执行以下处理:获取1种以上的克隆的培养数据;分析培养数据来限定作为预测对象的克隆;及使用针对作为预测对象的克隆测定出的数据来预测作为预测对象的克隆的有用物质的产生稳定性。

[0018] 根据第1方式,由于根据从培养数据获得的信息来限定预测对象并进行产生稳定性的预测,因此与不限定对象的情况相比,能够以高精度预测产生稳定性。并且,只要限定作为预测对象的克隆并进行预测所需的数据的获取即可,因此能够抑制成本。

[0019] 所预测的产生稳定性可以与实际上通过数个月的长期培养实验性地验证的产生稳定性相同地表示数个月后的未来的克隆的状态。例如,可以从长期培养后是否仍维持初期的产生量的观点来评价产生稳定性。根据第1方式,能够以高精度且低成本预测需要长期培养的稳定性试验的结果。

[0020] 本发明的第2方式所涉及的方法在第1方式所涉及的方法中,可以是以下结构:产生稳定性根据培养开始时和规定期间培养后的有用物质的产生量的变化有无来定义。

[0021] 本发明的第3方式所涉及的方法在第1方式或第2方式所涉及的方法中,可以是1个以上的处理器进行以下处理的结构:包括设定从培养数据获得的指标和与指标相关的阈值,并且根据指标的值和阈值来限定预测对象。

[0022] 本发明的第4方式所涉及的方法在第3方式所涉及的方法中,可以是以下结构:阈值被调整为产生稳定性的预测精度与不限定预测对象的情况相比更高。

[0023] 本发明的第5方式所涉及的方法在第3方式或第4方式所涉及的方法中,可以是使用针对指标的值的位次来定义阈值的结构。另外,“位次”可以是将针对多个克隆的指标的值以降序排列时的位次和以升序排列时的位次。例如,阈值可以定义为包含多个克隆的群

体中的相对位次的前40％等。

[0024] 本发明的第6方式所涉及的方法在第3方式至第5方式中任一方式所涉及的方法中,预测对象可以是指标值的上位群体。

[0025] 本发明的第7方式所涉及的方法在第3方式至第6方式中任一方式所涉及的方法中,指标可以是有用物质的产生量。

[0026] 本发明的第8方式所涉及的方法在第3方式至第6方式中任一方式所涉及的方法中,指标可以是积分活细胞密度。

[0027] 本发明的第9方式所涉及的方法在第3方式至第6方式中任一方式所涉及的方法中,指标可以是乳酸浓度。

[0028] 本发明的第10方式所涉及的方法在第1方式至第9方式中任一方式所涉及的方法中,可以是以下结构:用于产生稳定性的预测的数据可以包括1个以上的基因表达水平。

[0029] 本发明的第11方式所涉及的方法在第1方式至第10方式中任一方式所涉及的方法中,可以是以下结构:1个以上的处理器接收预测对象的数据的输入,并使用进行稳定或不稳定的2类分类的模型来预测产生稳定性。

[0030] 本发明的第12方式所涉及的方法在第11方式所涉及的方法中,模型是通过使用了多个训练数据的机器学习来训练的模型,所述多个训练数据是针对与作为预测对象的克隆进行了相同的限定的训练用克隆的数据和正解的稳定性标签建立关联的多个训练数据。

[0031] 本发明的第13方式所涉及的方法在第12方式所涉及的方法中,可以为以下结构:多个训练数据包括针对所产生的有用物质不同的多种类型的克隆的训练数据,1个以上的处理器预测针对产生与使用于模型的训练的有用物质不同的有用物质的克隆的产生稳定性。

[0032] 本发明的第14方式所涉及的方法在第1方式至第13方式中任一方式所涉及的方法中,有用物质可以是作为医药品原料的蛋白质、肽及病毒中的任一种。

[0033] 本发明的第15方式所涉及的方法在第1方式至第14方式中任一方式所涉及的方法中,有用物质可以是抗体或抗体样蛋白。

[0034] 本发明的第16方式所涉及的方法在第1方式至第15方式中任一方式所涉及的方法中,克隆可以是源自脊椎动物的细胞。

[0035] 本发明的第17方式所涉及的方法在第1方式至第15方式中任一方式所涉及的方法中,克隆可以是源自哺乳类的细胞。

[0036] 本发明的第18所涉及的方法在第1方式至第15方式中任一方式所涉及的方法中,克隆可以是CHO细胞或HEK细胞(Human Embryonic Kidney cells:人胚胎肾细胞)。

[0037] 本发明的第19方式所涉及的信息处理装置具备:1个以上的处理器;及1个以上的存储装置,存储使1个以上的处理器执行的命令,1个以上的处理器进行以下处理:关于产生有用物质的克隆,获取1种以上的克隆的培养数据;分析培养数据来限定作为预测对象的克隆;及使用针对作为预测对象的克隆测定出的数据,预测基于作为预测对象的克隆的有用物质的产生稳定性。

[0038] 关于第19方式所涉及的信息处理装置,能够设为包括与第2方式至第18方式中任一方式的方法相同的方式的结构。

[0039] 本发明的第20方式所涉及的程序使计算机实现以下功能:针对产生有用物质的克

隆获取1种以上的克隆的培养数据;分析培养数据来限定作为预测对象的克隆;及使用针对作为预测对象的克隆测定出的数据来预测基于作为预测对象的克隆的有用物质的产生稳定性。

[0040] 关于第20方式所涉及的程序,能够设为包括与第2方式至第18方式中任一方式的方法相同的方式的结构。

[0041] 本发明的第21方式所涉及的预测模型生成方法为生成使计算机实现预测产生有用物质的克隆的产生稳定性的功能的预测模型的预测模型生成方法,其中,包含1个以上的处理器的系统包括以下步骤:获取1种以上的克隆的培养数据;分析培养数据来限定作为预测对象的克隆;及使用针对属于预测对象的克隆测定出的数据和正解的稳定性标签建立关联的多个训练数据来进行机器学习,并以使相对于数据的输入的预测模型的输出接近正解的稳定性标签的方式,训练预测模型。

[0042] 关于第21方式所涉及的预测模型生成方法,能够设为包括与第2方式至第18方式中任一方式的方法相同的方式的结构。

[0043] 发明效果

[0044] 根据本发明,能够基于分析培养数据而获得的信息来适当地限定预测对象,并高精度地预测产生有用物质的克隆的产生稳定性。并且,根据本发明,通过限定预测对象,能够抑制产生稳定性的预测成本,能够以低成本进行预测。

附图说明

[0045] 图1是表示抗体医药品的生产工序的概要的说明图。

[0046] 图2是表示基于克隆的抗体产生量的变化的例子的图表。

[0047] 图3是对通过本实施方式实现的稳定性预测AI(Artificial Intelligence:人工智能)的作用进行概述的说明图。

[0048] 图4是以基因表达数据为基础来预测产生稳定性的机器学习模型的概念图。

[0049] 图5是表示本实施方式所涉及的克隆的产生稳定性预测方法的概要的说明图。

[0050] 图6是表示用于模型的训练及评价的数据集的例子的表格。

[0051] 图7是表示基于培养数据的某一指标对对象进行缩小的例子的图表。

[0052] 图8是表示作为评价样品而准备的5种抗体产生CHO细胞的克隆数与稳定性标签的赋予例的表格。

[0053] 图9是表示在各抗体种类中抗体生产量的值属于相对位次的前40%的克隆数与稳定性标签的赋予例的表格。

[0054] 图10是表示在各抗体种类中积分活细胞密度的值属于相对位次的前60%的克隆数与稳定性标签的赋予例的表格。

[0055] 图11是表示在各抗体种类中乳酸浓度的值属于相对位次的前40%的克隆数与稳定性标签的赋予例的表格。

[0056] 图12是表示实施方式所涉及的信息处理装置的功能结构的框图。

[0057] 图13是表示信息处理装置的硬件结构的例子的框图。

[0058] 图14是表示执行用于生成产生稳定性预测模型的机器学习的处理的机器学习装置的硬件结构的例子的框图。

[0059] 图15是表示机器学习装置所执行的机器学习方法的例子的流程图。

[0060] 图16是表示实施方式所涉及的信息处理装置所执行的信息处理方法的例子的流程图。

具体实施方式

[0061] 以下,根据附图对本发明的优选实施方式进行详细说明。

[0062] 《抗体医药品的生产工序的概要》

[0063] 在生物医药品中,从药效方面和安全方面的兼顾性高的角度考虑,市场正在扩大的抗体医药品使用能够稳定地产生具有复杂结构的蛋白质即抗体的动物细胞的克隆来生产。以下,作为有用物质,以抗体为例进行说明。图1是表示抗体医药品的生产工序的概要的说明图。生产抗体医药品为止的工艺包括[1]克隆制作阶段、[2]工艺开发阶段及[3]GMP(Good Manufacturing Practice:良好生产规范)制造阶段。

[0064] 克隆制作阶段包括以下工序:对适合抗体医药品的生产的动物细胞添加载体来进行基因重组,并制作多个克隆的候选;及从这些多个候选中甄别抗体的产生量、细胞增殖性、即使反复增殖细胞特性也不会发生变化的品质稳定性等方面优异的克隆。

[0065] 工艺开发阶段是使用甄别出的克隆来开发GMP制造所需的生产工艺(培养条件、纯化条件等)的阶段。

[0066] 在GMP制造阶段中,在已确立的生产工艺的基础上培养克隆并使其增殖,使克隆产生抗体。而且,通过纯化该抗体并制剂化来制作抗体医药品。

[0067] 在使克隆产生抗体的情况下,要求其产生性在长期内不发生变化(稳定)。因此,尽量制作多种类型的克隆,并从中选择产生性稳定的克隆,但是在以往,需要进行数个月的连续培养的实验性验证是必需的,因此负荷变高。

[0068] 图2是表示基于克隆的抗体产生量的变化的例子的图表。纵轴表示抗体的产生性,横轴表示经过时间(时间点)。“抗体的产生性”以由克隆产生的抗体的每单位时间的抗体产生量来表示。

[0069] 在图2中示出将由克隆产生的抗体的量在长时间(2～3个月)内变化多少进行了绘制的图表。图表G1是表示针对产生性稳定的克隆的抗体产生量的变化的图表。图表G2是表示针对产生性不稳定的克隆的抗体产生量的变化的图表。如图表G1所示,产生性稳定的克隆即使从当前时点经过2～3个月,产生性也大致不变,能够维持与当前时点大致不变的产生性。相对于此,如图表G2所示,产生性不稳定的克隆在2～3个月期间产生性逐渐下降。

[0070] 在本发明中,“当前时点”是指2周的标准试验时点或标准试验结束的时点,即用于判别产生稳定性的培养开始的时点。并且,“当前时点的抗体的产生性”是指2周的标准试验中的由克隆产生的每单位时间的抗体产生量。

[0071] 在通过基因导入来制作产生抗体的细胞的情况下,如图2所示,制作出稳定的克隆和不稳定的克隆这两者。因此,在克隆制作阶段中,制作多种类型的克隆,并从中选择显示出如图表G1那样的行为的产生性稳定的克隆。

[0072] 如图2所示的产生性的行为根据克隆的种类而异,以往每当使克隆制作的抗体的种类改变时,必须进行与图2相同的实验来评价各个克隆的产生稳定性。

[0073] 相对于此,在本发明的实施方式中,提出一种基于从当前时点的克隆获得的信息,

精度良好地预测数个月之后的抗体的产生稳定性的机制。在此,“从当前时点的克隆获得的信息”是指在2周的标准试验中从克隆获得的信息。作为预测的目标变量的抗体的产生稳定性,能够根据当前时点与培养数个月期间后的抗体的产生量有无变化来定义。这里的“数个月”例如是指2个月以上的期间,例示性地可以是2～3个月。并且,也可以设为直到进行规定次数的传代为止的期间。关于期间的设定,可以根据克隆的增殖能力来确定,也可以根据实际进行抗体的制造时的克隆的培养期间来确定。“当前时点”为图2的图表的左端所示的培养初期的时点,即2周的标准试验结束的时点,是用于判别抗体的产生稳定性的培养开始的时点。产生性“稳定”是指抗体的产生量在当前时点与数个月之后没有变化。“没有变化”包括变化量在允许范围内,且可以视为实质上没有变化的情况。产生性“不稳定”是指抗体的产生量在当前时点与数个月之后发生变化,大多数情况下是指产生量下降。被视为发生产生性的变化的阈值能够任意设定,例如相对于当前时点的产生量可以是±30%或±20%。

[0074] 《关于向未知克隆的泛化性能》

[0075] 图3是对通过本实施方式实现的稳定性预测AI(Artificial Intelligence)的作用进行说明的说明图。如图3所示,在克隆制作阶段中,对宿主细胞进行导入欲制作的有用物质的基因的设计图的基因导入。例如,在对宿主细胞将制作有用物质A的设计图进行基因导入的情况下,可获得产生有用物质A的细胞。由于这样的产生细胞能够概率性的形成,因此也会制作出不产生有用物质A的细胞或产生量不充分的细胞。因此,首先在该阶段进行简单的试验,选择能够充分产生有用物质A的高产的克隆。

[0076] 之后,若按照以往,则如图2中所说明的那样,进行2～3个月的稳定性试验,确认是否能够经数个月持续制作有用物质A,并选择具有产生稳定性的克隆。

[0077] 在本实施方式中,作为代替以往的稳定性试验的方法,构建稳定性预测AI,基于测定当前时点的克隆的状态即2周的标准试验中的克隆的状态而获得的特征数据(profile),通过稳定性预测AI预测2～3个月后的状态(产生性的变化)。

[0078] 细胞中产生的有用物质(例如抗体)的种类根据目的而多种多样,因此期望构建一种无关乎细胞产生的有用物质的种类而能够预测产生稳定性的模型。即,优选对未知的抗体种类稳健地预测抗体产生稳定性的模型。

[0079] 在学习适用于稳定性预测AI的模型时,无法事先知道作为对象的有用物质,在学习模型时使用的有用物质的种类与学习后使模型预测的对象的克隆所产生的有用物质可以成为不同的种类。即,优选对未知的有用物质种类稳健且精度良好地预测产生稳定性的模型,且优选构建以有用物质种类作为域的具有域泛化(domain generalization)性的预测模型。

[0080] 《预测产生稳定性的机器学习模型的概要》

[0081] 在本实施方式中,在克隆制作阶段,构建能够根据当前时点的克隆的信息来估计(预测)2～3个月之后的产生性有无变化,即能够预测有用物质的产生稳定性的稳定性预测AI。更具体而言,构建接收克隆的当前时点(标准试验时)的基因表达数据的输入,并输出表示有用物质的产生稳定性的稳定性标签的模型。更详细而言,将克隆的一部分作为使用于标准试验的克隆,将另一部分作为用于获取基因表达数据的基因分析的克隆,由此获取使用于标准试验的克隆的基因表达数据。稳定性标签能够由表示“稳定”的值的“1”或表示“不稳定”的值的“0”这2个值表示。预测产生稳定性的预测模型可以是进行“稳定”或“不稳定”

的类别分类的2类别分类模型。

[0082] 基因表达数据包括1个以上的基因水平。本实施方式中使用的基因表达数据包括将多个基因的各个的基因表达水平数值化而得的数据。基因表达数据例如能够通过RNA(ribonucleic acid:核糖核酸)序列分析来获得。表示基因表达量的值例如为取正整数的计数值,能够进行对数转换而用作特征量。

[0083] 图4是基于基因表达数据预测产生稳定性的机器学习模型MLM的概念图。在图4的矩形框RF1的内侧示出训练数据的数据集的例子。在图4中,作为将多个克隆A～N的各个的当前时点(标准试验时)的基因表达数据通过热图来可视化的基因表达模式GEP来表示。基因表达模式GEP的横轴表示基因的种类,多个基因的各个的基因表达水平由双色灰阶(热图)表示。基因表达数据中所包含的基因a、b、c、d……的种类数例如优选如下:关于稳定的克隆和不稳定的克隆获取所有基因表达数据,使用稳定的克隆与不稳定的克隆之间的2组统计学上的显著概率而选择的300～400种。而且,在缩小基因的种类数的情况下,优选使用已选择的基因一边使基因的种类数增减一边实际训练机器学习模型MLM,搜索预测性能变高的种类数,并缩小到例如50～100种基因。另外,在此获取了所有基因表达数据,但并不一定需要获取所有基因表达数据,也可以随机选择一部分基因,并获取其基因表达数据。由于图示的限制,无法表现热图的颜色,因此代替地将红色显示为“R”,将蓝色显示为“B”,将白色显示为“W”。红色(R)表示基因表达水平相对高,蓝色(B)表示基因表达水平相对低。白色(W)表示基因表达水平为中间值。

[0084] 关于多个克隆A～N的各个,根据基于标准试验后的数个月期间的培养的实验验证,可确认为“稳定”或“不稳定”,对各克隆A～N赋予表示“稳定”或“不稳定”的稳定性标签(正解标签)。如此,准备包括多个克隆A～N各自的当前时点的基因表达数据与作为正解的稳定性标签建立了关联(联系)的多个训练数据的数据集。之后,使用多个训练数据来训练机器学习模型MLM,并使机器学习模型MLM学习稳定或不稳定的基因模式。若对如此训练的学习完毕(训练完毕)的机器学习模型MLM输入未知的克隆X的当前时点(标准试验时)的基因表达数据,则机器学习模型MLM根据所输入的基因表达数据预测产生稳定性,并输出“稳定”或“不稳定”的标签作为预测结果。另外,在图4中,示出了机器学习模型MLM对未知的克隆X预测为“稳定”的例子。

[0085] 《实施方式的概要:限定预测对象来构建预测有用物质的产生稳定性的模型》

[0086] 由于产生有用物质的克隆具有各种特性,因此难以不问种类而高精度地预测所有克隆的产生稳定性。在本实施方式中,通过根据从当前时点(标准试验时)的各克隆的培养数据获得的指标来限定预测对象,实现高精度的预测。在此,培养数据是指,关于克隆,能够使用培养装置或对包含细胞的培养液进行一部分采样的专用装置来进行测定的一般数据。

[0087] 图5是表示本实施方式所涉及的克隆的有用物质的产生稳定性预测方法的概要的说明图。在图5的左图F5A中示出不限定预测对象的情况的比较例,在图5的右图F5B中示出基于本实施方式的方法的概要。

[0088] 对左图F5A的不限定预测对象的情况进行说明。在左图F5A的矩形框RF2内示意性地示出包含产生有用物质A～D的多种类型的克隆的训练数据的数据集DSc。该数据集DSc包含针对各有用物质A～D各5个克隆,共计20个克隆的训练数据。在此,训练数据是关于20个克隆的各个的标准试验时的基因表达数据与正解的稳定性标签建立关联(联系)的数据。在

图5的各克隆的下部显示的“9”、“7”、“6”等值表示标准试验时的各克隆的某一培养数据的测定值。另外,也可以不表示测定值,而是表示能够从测定值获取的各克隆中的相对水平。在此,对左图F5A进行了说明,但右图F5B也相同。

[0089] 在左图F5A中,示出不限定训练数据而使用数据集DSc的所有训练数据来训练机器学习模型MLMc,并使用学习完毕(训练完毕)的模型来预测产生未知的有用物质X的多种类型的克隆的产生稳定性。在该情况下,对于作为预测对象的产生未知的有用物质X的多种类型的克隆也没有特别限定,如矩形框RF3内所示,对产生未知的有用物质X的5种克隆全部进行产生稳定性的预测。产生稳定性的预测是通过对产生未知的有用物质X的5种克隆的全部获取当前时点(标准试验时)的基因表达数据,并输入到学习完毕(训练完毕)模型中来进行的,但其预测精度低。

[0090] 接着,对右图F5B的基于本实施方式的方法进行说明。与左图F5A的不限定预测对象的方法相比,在右图F5B所示的方法中,以标准试验时的某一培养数据的值为指标来限定预测对象。首先,关注某一培养数据的值来确定阈值,并将数据集DSd中所包含的克隆群体进行分组。例如,将作为指标的培养数据的值分为相对于阈值相对大和小这2个组。在此,示出将阈值设为“5”,将作为指标的培养数据的值为“5”以上的群体设为训练对象,且将培养数据的值小于“5”的群体排除在对象之外的例子。通过该阈值处理,如矩形框RF4内所示那样,针对各有用物质A～D保留各3个克隆共计12个克隆的训练数据来作为对象,并将包括这些有限的群体的训练数据的数据集DSe用于机器学习模型MLMe的训练。另一方面,虚线的矩形框RF5内所示的8个克隆的训练数据即不满足阈值的条件的克隆的训练数据排除在处理对象之外。

[0091] 如此,使用对象被限定的数据集DSe来训练机器学习模型MLMe。并且,在使用学习完毕的模型来预测产生未知的有用物质X的克隆的产生稳定性时,该作为预测对象的克隆与在模型的训练中使用的数据集DSe的克隆的群体同样地,对作为指标的培养数据的值适用阈值,并限定为满足基于阈值的限定条件的群体(指标的值高于阈值的群体)来进行预测。矩形框RF6内所示的3种克隆表示与预测对象对应的克隆。并且,虚线的矩形框RF7内所示的2种克隆表示排除在预测对象之外的克隆。如此通过限定预测对象来进行预测,能够实现高预测精度。而且,虚线的矩形框RF7内所示的排除在预测对象之外的克隆不需要获取基因表达数据,因此能够抑制基因分析的成本。

[0092] 《在训练及评价中使用的数据集的例子》

[0093] 图6中示出用于模型的训练及评价的数据集的例子。在图6的上部示出关于产生作为有用物质的抗体A的克隆的数据集DSA的例子,在下部示出关于产生作为有用物质的抗体B的克隆的数据集DSB的例子。虽然省略图示,但关于产生作为有用物质的其他种类的抗体的克隆的数据集也相同。

[0094] 数据集DSA包括针对多个克隆ACLj的各个测定出的标准试验时的培养数据、标准试验时的基因表达数据、及通过稳定性试验获得的作为正解的稳定性标签。词尾j表示识别克隆的索引编号。培养数据例如可以包括抗体产生量、积分活细胞密度(integral viable cell density:IVCD)、乳酸浓度、pH等1个以上的项目。培养数据可以是能够对培养装置或包含细胞的培养液进行一部分采样并使用专用装置进行测定的通常的数据,例如可以包含细胞总数、细胞分泌物质的量、细胞产物的量、细胞代谢物质的量及培养基成分量中的1个

以上。图6所示的表的各单元格内的文字符号(附加了词尾j的符号)表示对应的数据项目的值。

[0095] 关于数据集DSB也相同。数据集DSA中所包含的克隆ACLj的个数na与数据集DSB中所包含的克隆BCLj的个数nb可以不同。

[0096] 从如此准备的多个域(有用物质种类)的数据集中,关注某一培养数据的指标来进行对象的缩小(限定)。

[0097] 《预测对象的缩小的例子》

[0098] 图7是表示基于培养数据的某一指标的预测对象的缩小的例子的图表。在横轴上排列有产生多个有用物质A～E的各个的多种类型的克隆。纵轴为由标准试验时的培养数据获得的某一指标的值。另外,图7所示的克隆为用于模型的训练(学习)的克隆。

[0099] 如图7所示,有时根据所产生的有用物质种类不同的克隆,由培养数据获得的某一指标的分布范围不同。在这种情况下,若如图5中已说明那样,相对于指标的值确定阈值,根据与该阈值的相对大小关系来将克隆分为2个群体,并确定在训练中使用的克隆的群体和设为排除在训练对象之外的克隆的群体,则根据产生的有用物质种类,成为训练对象的克隆的数量会出现偏差。例如,设为将指标的阈值设为2.5且将阈值以上的值的克隆群体使用于训练的情况下,导致产生有用物质B的克隆不使用于训练。

[0100] 因此,例如如图7所示,关于产生各有用物质A～D的克隆,可以将训练对象限定为由培养数据获得的某一指标的相对的前X%(Top-X%)。在此,相对的前X%是指在产生各有用物质A～D的各个的克隆的群体中,对于由培养数据获得的某一指标按降序排列时的前X%(Top-X%)。相当于成为限定条件的阈值的“X%”这一基准优选调整为来自各有用物质A～E的采样的数量成为大致相同的数量。相对的前X%为本发明中的“使用指标的值的位次来定义的阈值”的一例。

[0101] 即使如此限定训练对象,其也可能存在产生性稳定的克隆和不稳定的克隆。并且,在使用学习(训练)完毕模型来预测产生未知的有用物质Y的克隆的产生稳定性时,该作为预测对象的克隆也与模型的训练中使用的克隆同样地,对由培养数据获得的某一指标限定为前X%的克隆来进行预测。

[0102] 在此,将产生多个有用物质A～E的各个的多种类型的克隆用于模型的训练,但无需一定要使用多种产生不同的有用物质的克隆,例如也可以仅将产生有用物质A的克隆使用于训练。在这种情况下,在训练中使用的克隆群体的限定方法可以关注标准试验时的某一培养数据的值来设定阈值,并根据与阈值的相对大小关系来进行,也可以关于由标准试验时的培养数据获得的某一指标的值设为前X%。并且,设为相对的前X%,但根据由培养数据获得的某一指标,也可以设为相对后X%。

[0103] 限定预测对象时的培养数据的指标和阈值可以通过对所准备的数据集以试错的方式反复进行假设和验证的作业来确定。又或者,限定预测对象时的培养数据的指标和阈值能够通过对所准备的数据集进行探索性分析来确定。

[0104] 例如,在存在如图7所示的5个有用物质A～E(域)的数据集的情况下,通过包含处理器的信息处理装置,使用滤波法(Filter Method)等特征选择的方法,在5个域中的各个中分别评价各特征量与目标变量(稳定性标签)的相关性,将在5个域中例如4个域以上中相关性高的特征量设为域普遍性高的特征量。信息处理装置从所有数据中着眼于某一指标将

满足特定条件的数据作为子集来提取,并针对所提取的子集基于域普遍性高的特征量的个数进行域泛化性评价。在域普遍性高的特征量的个数多的情况下,评价为域泛化性高的子集。通过将域泛化性高的子集的数据用作训练数据来进行预测模型的学习(训练),学习完毕的模型对于在与学习时相同的条件下限定对象的群体(子集),也能够对其他域(有用物质种类)稳健地预测产生稳定性。

[0105] 在抗体产生克隆的情况下,对预测对象的限定有效的培养数据的指标例如为抗体产生量、积分活细胞密度、乳酸浓度等,确认到通过将这些中的任一个指标的值的上位群体作为对象,能够进行高精度的产生稳定性预测。

[0106] 《有用物质的例子》

[0107] 有用物质不限于抗体,也可以是抗体样蛋白。有用物质可以是作为医药品原料的蛋白质、肽及病毒中的任一种。

[0108] 《克隆的例子》

[0109] 产生有用物质的克隆可以是源自脊椎动物的细胞。克隆例如可以是源自哺乳类的细胞。克隆可以是CHO细胞或HEK细胞。

[0110] 《实施例》

[0111] 以下,对适用了本发明的技术的实施例1～3进行说明。各实施例1～3中共同的结构如下。即,将有用物质设为抗体,且将产生细胞设为CHO细胞。示出以下例子:作为评价样品,针对5种抗体产生CHO细胞的克隆,分别准备多种克隆,通过RNA测序(RNA-Seq)分析,从在2周的标准试验中测定的所有基因表达水平选择100种基因表达水平来作为解释变量,并实施将分类为稳定或不稳定的2个类别的逻辑回归模型作为学习器的5折交叉验证来进行预测模型的训练(学习),性能评价则使用了PRAUC(Area Under the Precision-Recall Curve:精确率-召回率曲线下面积)。关于用于解释变量的基因表达水平的种类,在实施例1～3中,使用通过统计学上的显著概率而选择的300～400种基因,一边使种类数增减一边实际进行预测模型的训练(学习)来搜索预测性能变高的种类数,由此设为100种。另外,关于标准试验,克隆(CHO细胞)的播种数为5×10^5cells/mL,在40mL的烧瓶中进行了悬浮培养。

[0112] 5折交叉验证按照5种抗体种类进行分割,并评价了基于未学习的抗体种类的性能。即,将4种抗体种类的数据集用作训练(学习)用数据,将剩余的1种抗体种类的数据集用作性能评价用测试数据。

[0113] 图8是表示作为评价样品而准备的5种抗体产生CHO细胞的克隆数与稳定性标签的赋予例的表格。作为评价样品,准备182个克隆的5种抗体产生细胞,在与标准试验相同的条件下进行2个月细胞培养,由此对各克隆赋予了稳定性标签(“稳定”或“不稳定”)(参考图8)。例如,产生抗体A的克隆为共计24个克隆,其中被赋予“稳定”标签的克隆为7个,被赋予“不稳定”标签的克隆为17个。并且,关于182个克隆的各个,在标准试验时获取培养数据和基因表达数据,针对每一克隆将基因表达数据与稳定性标签等建立关联,构成训练数据。

[0114] [实施例1]

[0115] 在实施例1中,对进行将预测对象限定为“相对高产的克隆”的稳定性预测的例子进行说明。在此,“相对高产的克隆”是指有用物质的产生量相对高的克隆。

[0116] 对进行将预测对象限定为“相对高产的克隆”的稳定性预测时的、针对作为用于预测模型的训练的训练对象的克隆的限定方法进行说明。另外,限定作为训练对象的克隆相

当于限定在训练中作为使预测模型预测的对象的克隆,即相当于限定作为基于预测模型的预测对象的克隆。

[0117] 作为训练对象的克隆的限定方法设为如下方法:从所有182个克隆的标准试验时的培养数据中关注“抗体产生量”,以预测模型中的预测性能变高的方式搜索阈值,限定为各抗体种类中相对位次为前40%的克隆。在此,“抗体产生量”例如能够设为标准试验中的2周(14天)的抗体产生量的累计量。或者,也可以设为标准试验中的某一期间,例如期间,例如10天的抗体产生量的累计量,也可以设为以除以测量期间而得的每单位时间的抗体产生量。“前40%”是阈值的一例。图9是表示在各抗体种类中抗体生产量的值属于相对位次的前40%的克隆数与稳定性标签的赋予例的表格。

[0118] 在图9中示出在各抗体种类中属于相对位次的前40%的共计73个克隆的例子。使用包括图9所示的73个克隆的标准试验时的基因表达数据与稳定性标签建立了关联的训练数据的每个抗体种类的数据集实施了5折交叉验证。如此限定了训练对象的结果,学习完毕的预测模型的预测性能的PRAUC的值成为0.743。另外,关于作为使用学习完毕的预测模型来预测未知的有用物质的产生稳定性时的预测对象的克隆,也与训练对象的限定同样地,设为分析标准试验时(当前时点)的培养数据,并限定为“抗体产生量”的前40%来进行预测。

[0119] [比较例]

[0120] 相对于此,不限定训练对象,使用包括图8所示的182个克隆的所有数据的数据集进行相同的学习,并实施5折交叉验证时所获得的比较例所涉及的预测模型的预测性能的PRAUC的值为0.503。另外,预测对象设为与训练对象同样地不限定对象而进行。确认到通过实施例1限定了预测对象的预测模型的性能与比较例所涉及的预测模型的性能相比为高精度。

[0121] 该结果表示,使用通过实施例1的方法生成的预测模型,能够对未知的有用物质进行高精度的预测,同时,限定为相对高产的克隆在产生有用物质的克隆的选择工序中完全不会成为障碍,通过限定为能够以高精度进行预测的对象,能够以低成本来实施,因此认为基于本发明的稳定性预测能够实际使用。

[0122] [实施例2]

[0123] 在实施例2中,对进行将预测对象限定为“相对细胞密度高的克隆”的稳定性预测的例子进行说明。首先,对于在进行将预测对象限定为“相对细胞密度高的克隆”的稳定性预测时的、作为针对用于训练预测模型的训练对象的克隆的限定方法进行说明。与实施例1同样地,设为以下方法:从图8所示的所有182个克隆的标准试验时的培养数据中着眼于“积分活细胞密度(IVCD)”,以预测模型中的预测性能变高的方式搜索阈值,并限定为各抗体种类中相对位次的前60%的克隆。在此,“细胞密度相对高的克隆”例如能够根据标准试验中的2周(14天)的“积分活细胞密度(IVCD)”来获取。或者,也可以根据标准试验中的某一期间,例如10天的“积分活细胞密度(IVCD)”来获取。“前60%”是阈值的一例。图10是表示在各抗体种类中积分活细胞密度的值属于相对位次的前60%的克隆数与稳定性标签的赋予例的表格。

[0124] 在图10中示出在各抗体种类中属于相对位次的前60%的共计109个克隆的例子。使用包括图10所示的109个克隆的标准试验时的基因表达数据与稳定性标签建立了关联的

训练数据的每个抗体种类的数据集实施了5折交叉验证。如此限定了训练对象的结果,学习完毕的预测模型的预测性能的PRAUC的值成为0.647。即,确认到通过实施例2限定了预测对象的预测模型的性能与不限定对象的比较例所涉及的预测模型的PRAUC(0.503)相比为高精度。另外,关于作为使用学习完毕的预测模型来预测未知的有用物质的产生稳定性时的预测对象的克隆,也与训练对象的限定同样地,设为分析标准试验时(当前时点)的培养数据,并限定为“积分活细胞密度(IVCD)”的前60%来进行预测。

[0125] 该结果表示,使用通过实施例2的方法生成的预测模型,能够对未知的有用物质进行高精度的预测,同时,限定为活细胞密度相对高的克隆在产生有用物质的克隆的选择工序中不会成为障碍,通过限定为能够以高精度进行预测的对象,能够以低成本来实施,因此认为基于本发明的稳定性预测能够实际使用。

[0126] [实施例3]

[0127] 在实施例3中,对限定为“乳酸浓度相对高的克隆”的稳定性预测的例子进行说明。首先,对于在进行将预测对象限定为“相对乳酸浓度高的克隆”的稳定性预测时的、针对作为用于训练预测模型的训练对象的克隆的限定方法进行说明。与实施例1同样地,从图8所示的共182个克隆的2周的标准试验的培养数据中着眼于培养克隆的培养液的“乳酸浓度”,将2周(14天)内的各时点,例如每天测定的培养液的“乳酸浓度”的中央值作为代表值,获取各克隆的“乳酸浓度”。之后,设为以预测模型中的预测性能变高的方式搜索阈值,并限定为各抗体种类中相对位次的前40%的克隆的方法。“前40%”是阈值的一例。图11是表示在各抗体种类中乳酸浓度的值与相对位次的前40%对应的克隆数与稳定性标签的赋予例的表格。

[0128] 在图11中示出在各抗体种类中与相对位次的前40%对应的共计72个克隆的例子。另外,与图9相比,克隆的数量少了1个克隆的原因是,在乳酸浓度的测定中,存在1个克隆的数据缺失。

[0129] 使用包括图11所示的72个克隆的标准试验时的基因表达数据与稳定性标签建立了关联的训练数据的每个抗体种类的数据集实施了5折交叉验证。如此限定了对象的结果,学习完毕的预测模型的预测性能的PRAUC的值成为0.613。即,确认到通过实施例3限定了预测对象的预测模型的性能与不限定对象的比较例所涉及的预测模型的PRAUC(0.503)相比为高精度。另外,关于作为使用学习完毕的预测模型来预测未知的有用物质的产生稳定性时的预测对象的克隆,也与训练对象的限定同样地,设为分析标准试验时(当前时点)的培养数据,并限定为“乳酸浓度”的前40%来进行预测。

[0130] 该结果表示,能够对未知的有用物质进行高精度的预测,同时,限定为乳酸浓度相对高的克隆在产生有用物质的克隆的选择工序中不会成为障碍,通过限定为能够以高精度进行预测的对象,能够以低成本来实施,因此认为基于本发明的稳定性预测能够实际使用。

[0131] 《信息处理装置的结构例》

[0132] 图12是表示实施方式所涉及的信息处理装置10的功能结构的框图。信息处理装置10具备数据获取部12、预测对象限定部14、产生稳定性预测模型16及处理结果输出部18。信息处理装置10的各种功能能够通过计算机的硬件与软件的组合来实现。信息处理装置10的物理形态并无特别限定,可以是服务器计算机,也可以是工作站,也可以是个人计算机或平板终端等。

[0133] 数据获取部12获取包括关于产生有用物质的克隆的1种以上的克隆的培养数据及基因表达数据的各种数据。

[0134] 预测对象限定部14包括培养数据分析部20和限定条件判定部22,分析所输入的1种以上的克隆的培养数据来限定作为预测对象的克隆。培养数据分析部20进行培养数据的分析。限定条件判定部22根据培养数据的分析结果,通过阈值来限定对象。另外,为了便于说明,将培养数据分析部20和限定条件判定部22分开记载,但限定条件判定部22也可以包括在培养数据分析部20中。并且,也可以理解为培养数据分析部20作为预测对象限定部14发挥作用。

[0135] 培养数据分析部20能够执行从所输入的数据集确定用于限定预测对象的指标和阈值的处理。另外,关于成为预测对象的限定条件的指标和阈值,可以根据基于培养数据分析部20的分析结果来设定,也可以作为通过使用未图示的其他信息处理装置等的搜索处理的结果等来事先掌握的已知信息,设定于预测对象限定部14。

[0136] 在产生稳定性预测模型16中适用机器学习模型。产生稳定性预测模型16可以是接收作为预测对象的克隆的当前时点的基因表达数据的输入,根据所输入的基因表达数据预测克隆的产生稳定性并输出稳定性标签的2类分类模型。产生稳定性预测模型16使用通过在图5的右图F5B中说明的方法限定对象的训练数据来进行训练。输入到产生稳定性预测模型16中的基因表达数据包括1个以上的基因表达水平。输入到产生稳定性预测模型16中的基因表达数据中可以包含多个基因的表达水平的数据。用作解释变量的特征量可以通过公知的特征量选择方法来选择。

[0137] 处理结果输出部18输出包括产生稳定性预测模型16的预测结果的处理结果。处理结果输出部18例如可以是进行显示处理结果的处理、将处理结果记录于数据库等的处理、及打印处理结果的处理中的至少1个处理的结构。

[0138] 图13是表示信息处理装置10的硬件结构的例子的框图。在此,叙述使用1台计算机来实现信息处理装置10的处理功能的例子,但信息处理装置10的处理功能也可以通过使用多台计算机构成的计算机系统来实现。

[0139] 信息处理装置10具备处理器102、非临时性有形计算机可读介质104、通信接口106、输入输出接口108及总线110。处理器102通过总线110与计算机可读介质104、通信接口106及输入输出接口108连接。

[0140] 处理器102包括CPU(Central Processing Unit:中央处理器)。处理器102也可以包括GPU(Graphics Processing Unit:图形处理器)。计算机可读介质104包括作为主存储装置的存储器112及作为辅助存储装置的存储设备114。计算机可读介质104例如可以是半导体存储器、硬盘(Hard Disk Drive:HDD)装置、或固态硬盘(Solid State Drive:SSD)装置、或者它们的多个的组合。计算机可读介质104为本发明中的“存储装置”的一例。

[0141] 计算机可读介质104包括存储1种以上的克隆的培养数据及基因表达数据等各种数据的数据存储区域120。并且,在计算机可读介质104中存储有包括预测对象限定程序140、产生稳定性预测模型16、处理结果输出程序180及显示控制程序190的多个程序以及数据等。“程序”这一术语包括程序模块的概念,且包括遵照程序的命令。处理器102通过执行存储于计算机可读介质104中的程序的命令,作为各种处理部发挥功能。

[0142] 预测对象限定程序140包括执行以下处理的命令:分析培养数据来限定预测对象。

预测对象限定程序140可以构成为包括培养数据分析程序142和限定条件判定程序144。培养数据分析程序142包括执行以下处理的命令:对1种以上的克隆的培养数据进行分析。培养数据分析程序142可以包括执行以下处理的命令:搜索用于从数据集缩小预测对象的指标和阈值。

[0143] 限定条件判定程序144包括执行以下处理的命令:利用培养数据分析程序142的分析结果,并根据作为限定条件而设定的指标和阈值来限定预测对象。

[0144] 产生稳定性预测模型16包括执行以下处理的命令:接收满足限定条件的预测对象所涉及的克隆的基因表达数据的输入来预测产生稳定性。

[0145] 处理结果输出程序180包括执行以下处理的命令:输出包括通过产生稳定性预测模型16预测的产生稳定性的处理结果。显示控制程序190包括执行以下处理的命令:生成显示装置154的显示输出所需的显示用信号,并进行显示装置154的显示控制。

[0146] 通信接口106通过有线或无线进行与外部装置的通信处理,并在与外部装置之间进行信息的交换。信息处理装置10经由通信接口106与未图示的通信线路连接。通信线路可以是局域网,也可以是广域网,也可以是它们的组合。通信接口106能够担任接收数据输入的数据获取部的角色。

[0147] 信息处理装置10可以具备输入装置152和显示装置154。输入装置152例如由键盘、鼠标、多点触摸面板或其他指示设备或语音输入装置或它们的适当的组合构成。显示装置154例如由液晶显示器、有机EL(organic electro-luminescence(有机电致发光):OEL)显示器或投影仪或它们的适当的组合构成。输入装置152和显示装置154经由输入输出接口108与处理器102连接。另外,可以如触摸面板那样,输入装置152和显示装置154构成为一体,也可以如触摸面板式平板终端那样,信息处理装置10、输入装置152及显示装置154构成为一体。

[0148] 《机器学习装置的结构例》

[0149] 图14是表示执行用于生成产生稳定性预测模型16的机器学习的处理的机器学习装置300的硬件结构的例子的框图。在此,叙述使用1台计算机来实现机器学习装置300的处理功能的例子,但机器学习装置300的处理功能也可以通过使用多台计算机构成的计算机系统来实现。

[0150] 机器学习装置300具备处理器302、非临时性有形计算机可读介质304、通信接口306、输入输出接口308及总线310。计算机可读介质304包括存储器312及存储设备314。处理器302通过总线310与计算机可读介质304、通信接口306及输入输出接口308连接。输入装置352及显示装置354经由输入输出接口308与总线310连接。

[0151] 机器学习装置300的硬件结构可以与在图6中说明的信息处理装置10的对应的要件相同。机器学习装置300的方式可以是服务器计算机,也可以是个人计算机,也可以是工作站。机器学习装置300为本发明中的“包括1个以上的处理器的系统”的一例。

[0152] 机器学习装置300经由通信接口306与未图示的通信线路连接,且与数据保存部550等外部装置以能够通信的方式连接。数据保存部550包括保存有包含多个训练数据的数据集的存储设备。数据保存部550中可以保存有包括如图6中所例示的多个域的所有数据的数据集,也可以保存有仅包括作为预测对象而被限定的对象的样品的数据的数据集。另外,数据保存部550也可以构建于机器学习装置300内的存储设备314。

[0153] 在计算机可读介质304中存储包括预测对象限定程序320、学习处理程序330及显示控制程序340的多个程序以及数据等。预测对象限定程序320可以与在图12中说明的预测对象限定程序140相同。显示控制程序340可以与在图12中说明的显示控制程序190相同。

[0154] 计算机可读介质304包括预测对象数据存储区域322。在预测对象数据存储区域322中存储与所限定的预测对象对应的训练数据。可以通过预测对象限定程序320从保存于数据保存部550的数据集中适当采样对应的训练数据,也可以预先提取仅为预测对象的数据集来作为子集。

[0155] 学习处理程序330包括数据获取程序400、作为机器学习模型的预测模型410、损失计算程序430及优化器440。数据获取程序400包括执行以下处理的命令:从预测对象数据存储区域322获取训练数据。经由数据获取程序400获取的训练数据输入到预测模型410。

[0156] 损失计算程序430包括执行以下处理的命令:计算出表示从预测模型410输出的稳定性标签的预测值与正解的稳定性标签之间的误差的损失。优化器440包括执行以下处理的命令:根据计算出的损失来计算出预测模型410的参数的更新量,并更新预测模型410的参数。优化器440例如可以通过随机梯度下降法(Stochastic Gradient Descent:SGD)等方法进行参数的最佳化。

[0157] 《机器学习方法的流程图》

[0158] 图15是表示机器学习装置300所执行的机器学习方法的例子的流程图。在此,作为准备有如图6中例示的用于机器学习的数据集的情况进行说明。在步骤S102中,处理器302从所准备的数据集中获取培养数据。

[0159] 在步骤S104中,处理器302分析培养数据,并限定训练对象。处理器302可以根据预先指定的培养数据的指标和阈值来筛选是满足限定条件的对象样品的数据还是不满足限定条件的并非对象的样品的数据,也可以根据培养数据搜索作为限定条件的指标和阈值,并筛选对象样品的数据和并非对象的样品的数据。

[0160] 在步骤S106中,处理器302仅使用满足限定条件的克隆的数据进行机器学习,并训练预测模型410。即,处理器302将满足限定条件的样品的基因表达数据输入到预测模型410中,并计算出表示从预测模型410输出的稳定性标签的预测值与正解的稳定性标签之间的误差的损失。处理器302根据计算出的损失来计算出预测模型410的参数的更新量,并更新参数。如此,处理器302以相对于输入到预测模型410中的数据,使来自预测模型410的输出(预测值)接近正解的稳定性标签的方式,训练预测模型410。另外,预测模型410的参数的更新可以以小批量为单位来实施。

[0161] 在步骤S108中,处理器302判定是否结束学习。学习的结束条件可以根据损失值来确定,也可以根据参数的更新次数来确定。作为基于损失值的方法,例如可以将损失收敛到规定的范围内作为学习结束条件。并且,作为基于更新次数的方法,例如可以将更新次数达到规定次数作为学习结束条件。或者,也可以除了训练数据以外另外准备模型的性能评价用数据集,并根据使用了评价用数据的评价值来判定可否结束学习。

[0162] 在步骤S108的判定结果判定为“否”的情况下,处理器302返回到步骤S106,继续学习处理。另一方面,在步骤S108的判定结果判定为“是”的情况下,处理器302结束图12的流程图。

[0163] 学习完毕的预测模型410作为产生稳定性预测模型16被组装到信息处理装置10

中。机器学习装置300所执行的机器学习方法能够理解为生成产生稳定性预测模型16的方法,是本发明中的预测模型生成方法的一例。

[0164] 《进行产生稳定性的预测的信息处理方法的流程图》

[0165] 图16是表示信息处理装置10所执行的信息处理方法的例子的流程图。在步骤S202中,处理器102获取针对产生有用物质的克隆测定出的培养数据。处理器102可以从未图示的数据保存服务器等自动获取数据,也可以经由用户界面接收数据的指定输入,并获取关于所指定的克隆的数据。

[0166] 在步骤S204中,处理器102分析培养数据,并限定预测对象。处理器102适用与在训练产生稳定性预测模型16时限定训练对象的条件相同的限定条件来限定预测对象。另外,通过该步骤S204限定预测对象之后,对属于预测对象的克隆实施基因表达数据的测量,由此与实施所有克隆的基因分析的情况相比,能够降低作业负荷及成本。

[0167] 在步骤S206中,处理器102将属于预测对象的克隆的基因表达数据输入到产生稳定性预测模型16中,并通过产生稳定性预测模型16预测稳定性。

[0168] 在步骤S208中,处理器102输出从产生稳定性预测模型16输出的预测结果。基于该产生稳定性的预测结果,能够进行产生克隆的选择。

[0169] 在步骤S208之后,处理器102结束图16的流程图。

[0170] 《关于运行计算机的程序》

[0171] 能够将使计算机实现实施方式所涉及的信息处理装置10及机器学习装置300的各装置中的处理功能的一部分或全部的程序记录于光盘、磁盘或者半导体存储器、其他作为有形非临时性信息存储介质的计算机可读介质,并通过该信息存储介质提供程序。

[0172] 并且,也能够代替将程序存储于这种有形非临时性计算机可读介质中来提供的方式,利用互联网等电信线路将程序信号作为下载服务来提供。

[0173] 而且,也可以通过云计算来实现上述各装置中的处理功能的一部分或全部,并且,也能够作为SaaS(Software as a Service:软件即服务)来提供。

[0174] 《关于各处理部的硬件结构》

[0175] 信息处理装置10中的数据获取部12、预测对象限定部14、包含产生稳定性预测模型16的稳定性预测部、处理结果输出部18、培养数据分析部20、限定条件判定部22、机器学习装置300中的包含预测模型410的学习部、损失计算部、参数更新量计算部、参数更新部等执行各种处理的处理部(processing unit)的硬件结构例如为如下所示的各种处理器(processor)。

[0176] 各种处理器中,包括执行程序而作为各种处理部发挥作用的通用的处理器即CPU、GPU、FPGA(Field Programmable Gate Array:现场可编程门阵列)等在制造之后能够变更电路结构的处理器即可编程逻辑器件(Programmable Logic Device:PLD)、ASIC(Application Specific Integrated Circuit:专用集成电路)等具有为了执行特定处理而专门设计的电路结构的处理器即专用电路等。

[0177] 1个处理部可以由这些各种处理器中的1个构成,也可以由相同种类或不同种类的2个以上的处理器构成。例如,1个处理部可以通过复数个FPGA或者CPU与FPGA的组合或CPU与GPU的组合而构成。并且,也可以由1个处理器构成多个处理部。作为由1个处理器构成多个处理部的例子,首先,有如以客户端、服务器等计算机为代表那样,由1个以上的CPU和软

件的组合构成1个处理器,该处理器作为多个处理部发挥功能的方式。第二,具有如以片上系统(System On Chip:SoC)等为代表那样使用由1个IC(Integrated Circuit:集成电路)晶片实现包含复数个处理部的整个系统的功能的处理器装置的形态。如此,各种处理部作为硬件结构而使用1个以上的上述各种处理器构成。

[0178] 而且,更具体而言,这些各种处理器的硬件结构为将半导体元件等电路元件组合而成的电路(circuitry)。

[0179] 《实施方式的优点》

[0180] 根据上述实施方式所涉及的预测产生克隆的产生稳定性的方法及执行该方法的信息处理装置10,可获得如下效果。

[0181] [1]根据当前时点(标准试验时)的培养数据的指标来适当地限定作为预测对象的克隆,因此能够对作为预测对象的克隆以高精度预测产生稳定性。

[0182] [2]由于仅限定为作为预测对象的克隆来进行基因分析(RNA-Seq分析)即可,因此与对所有克隆进行基因分析的情况相比,能够抑制成本。

[0183] [3]通过适用本实施方式所涉及的方法来代替以往的稳定性试验,能够实现产生细胞的开发工序的期间缩短及低成本化。

[0184] 《其他》

[0185] 本发明并不限定为上述实施方式,在不脱离本发明的技术思想的主旨的范围内,可以进行各种变形。

[0186] 符号说明

[0187]	10	信息处理装置
[0188]	12	数据获取部
[0189]	14	预测对象限定部
[0190]	16	产生稳定性预测模型
[0191]	18	处理结果输出部
[0192]	20	培养数据分析部
[0193]	22	限定条件判定部
[0194]	102	处理器
[0195]	104	计算机可读介质
[0196]	106	通信接口
[0197]	108	输入输出接口
[0198]	110	总线
[0199]	112	存储器
[0200]	114	存储装置
[0201]	120	数据存储区域
[0202]	140	预测对象限定程序
[0203]	142	培养数据分析程序
[0204]	144	限定条件判定程序
[0205]	152	输入装置
[0206]	154	显示装置

[0207]	180	处理结果输出程序
[0208]	190	显示控制程序
[0209]	300	机器学习装置
[0210]	302	处理器
[0211]	304	计算机可读介质
[0212]	306	通信接口
[0213]	308	输入输出接口
[0214]	310	总线
[0215]	312	存储器
[0216]	314	存储设备
[0217]	320	预测对象限定程序
[0218]	322	预测对象数据存储区域
[0219]	330	学习处理程序
[0220]	340	显示控制程序
[0221]	352	输入装置
[0222]	354	显示装置
[0223]	400	数据获取程序
[0224]	410	预测模型
[0225]	430	损失计算程序
[0226]	440	优化器
[0227]	550	数据保存部
[0228]	DSA、DSB	数据集
[0229]	DSc、DSd、DSe	数据集
[0230]	F5A	左图
[0231]	F5B	右图
[0232]	G1	图表
[0233]	G2	图表
[0234]	GEP	基因表达模式
[0235]	MLM	机器学习模型
[0236]	MLMc、MLMe	机器学习模型
[0237]	RF1～RF7	矩形框
[0238]	S102～S108	机器学习方法的步骤
[0239]	S202～S208	预测产生稳定性的信息处理方法的步骤

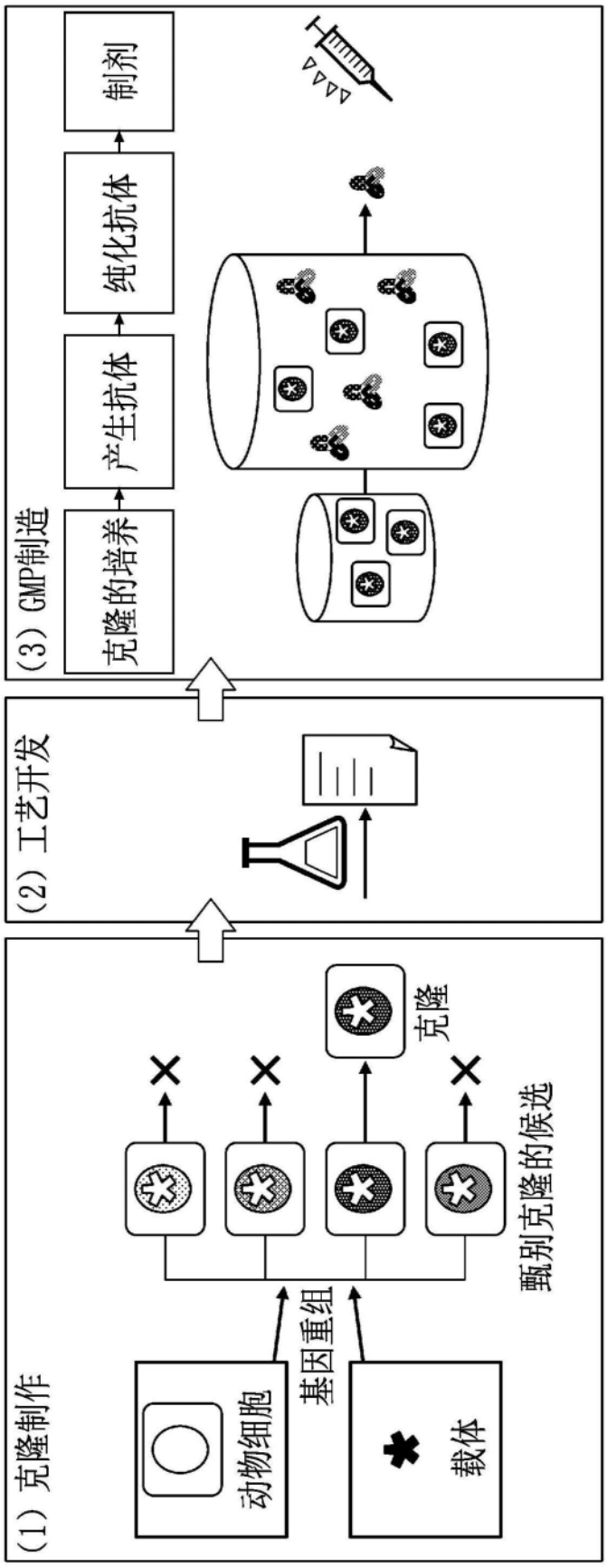

(1) 克隆制作
动物细胞
基因重组
载体
克隆
甄别克隆的候选
(2) 工艺开发
(3) GMP制造
克隆的培养
产生抗体
纯化抗体
制剂

图1

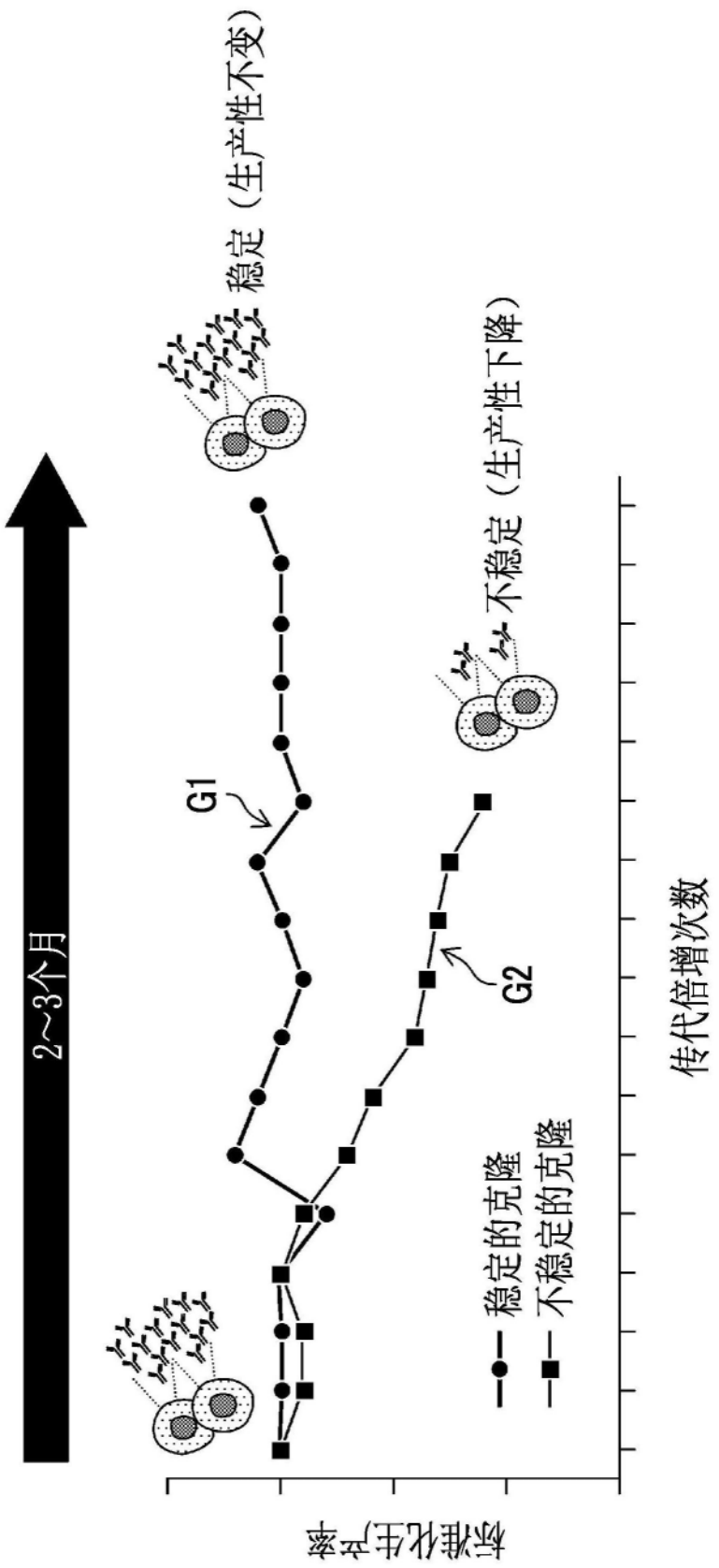


图2

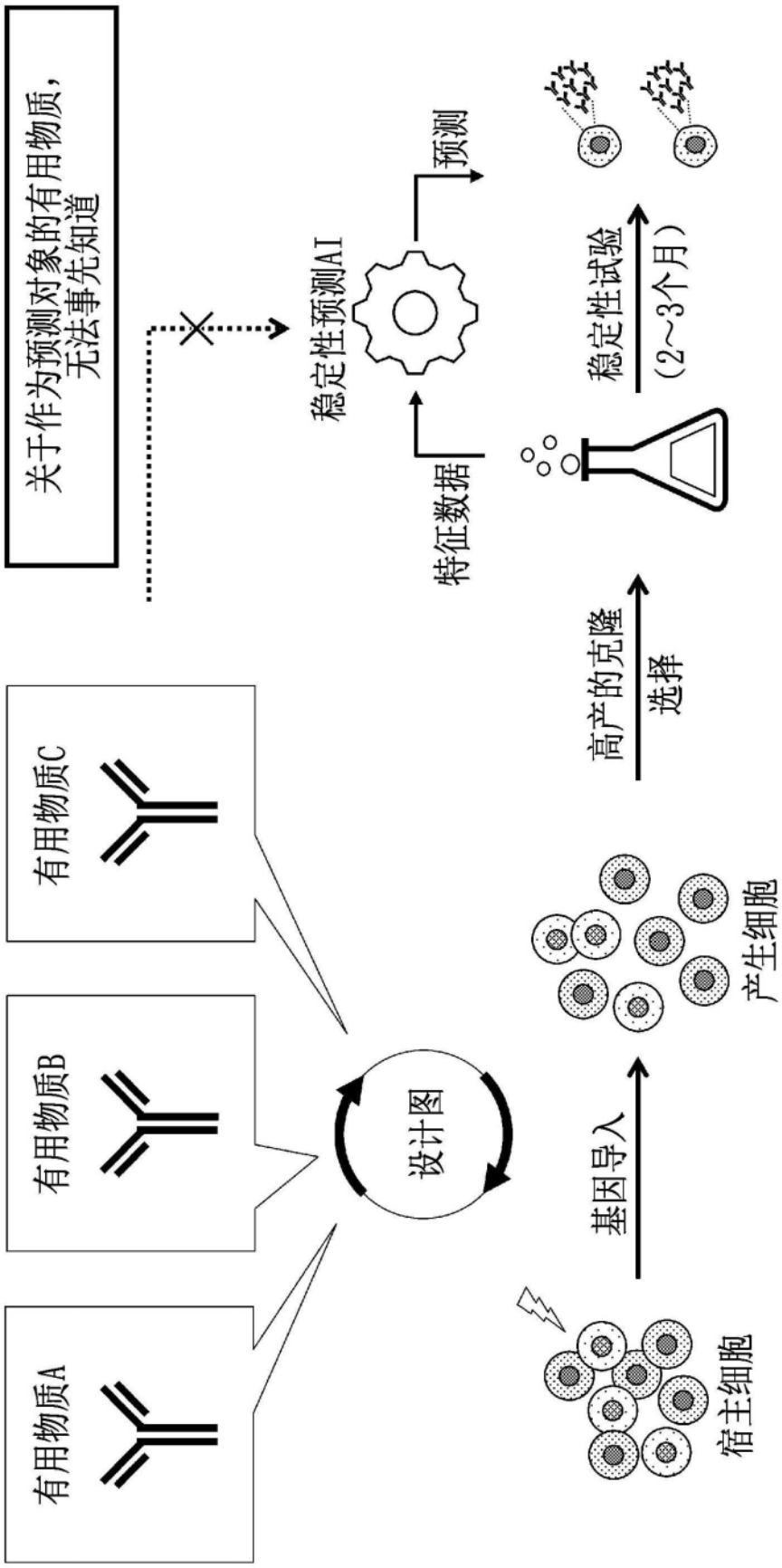

关于作为预测对象的有用物质，
无法事先知道
稳定性预测AI
预测
特征数据
稳定性试验
（2~3个月）
高产的克隆
选择
有用物质A
有用物质B
有用物质C
设计图
基因导入
产生细胞
宿主细胞

图3

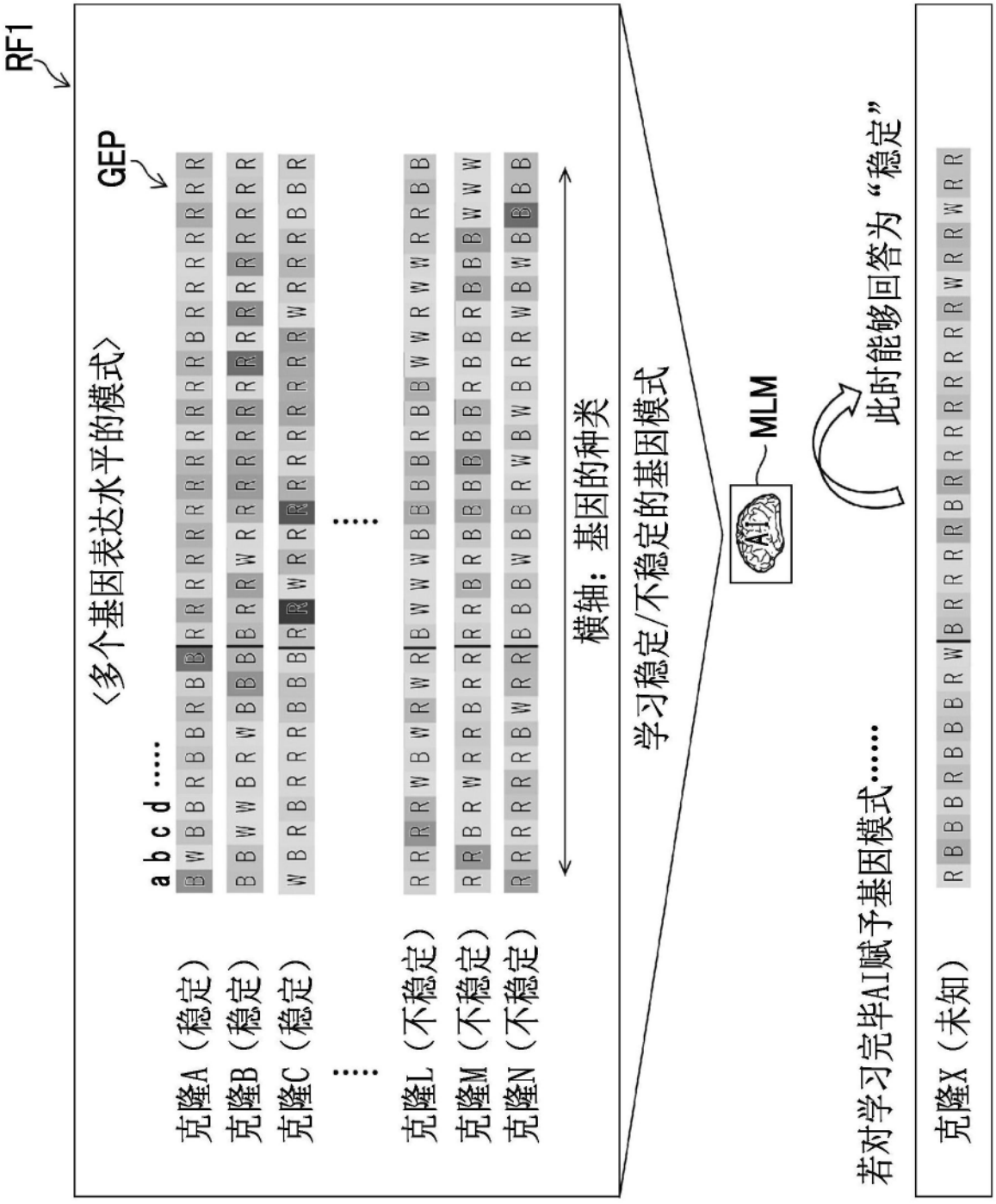

RF1
〈多个基因表达水平的模式〉
GEP
a b c d ……
克隆A（稳定）
克隆B（稳定）
克隆C（稳定）
……
克隆L（不稳定）
克隆M（不稳定）
克隆N（不稳定）
横轴：基因的种类
学习稳定/不稳定的基因模式
MLM
AI
若对学习完毕AI赋予基因模式……
此时能够回答为“稳定”
克隆X（未知）

图4

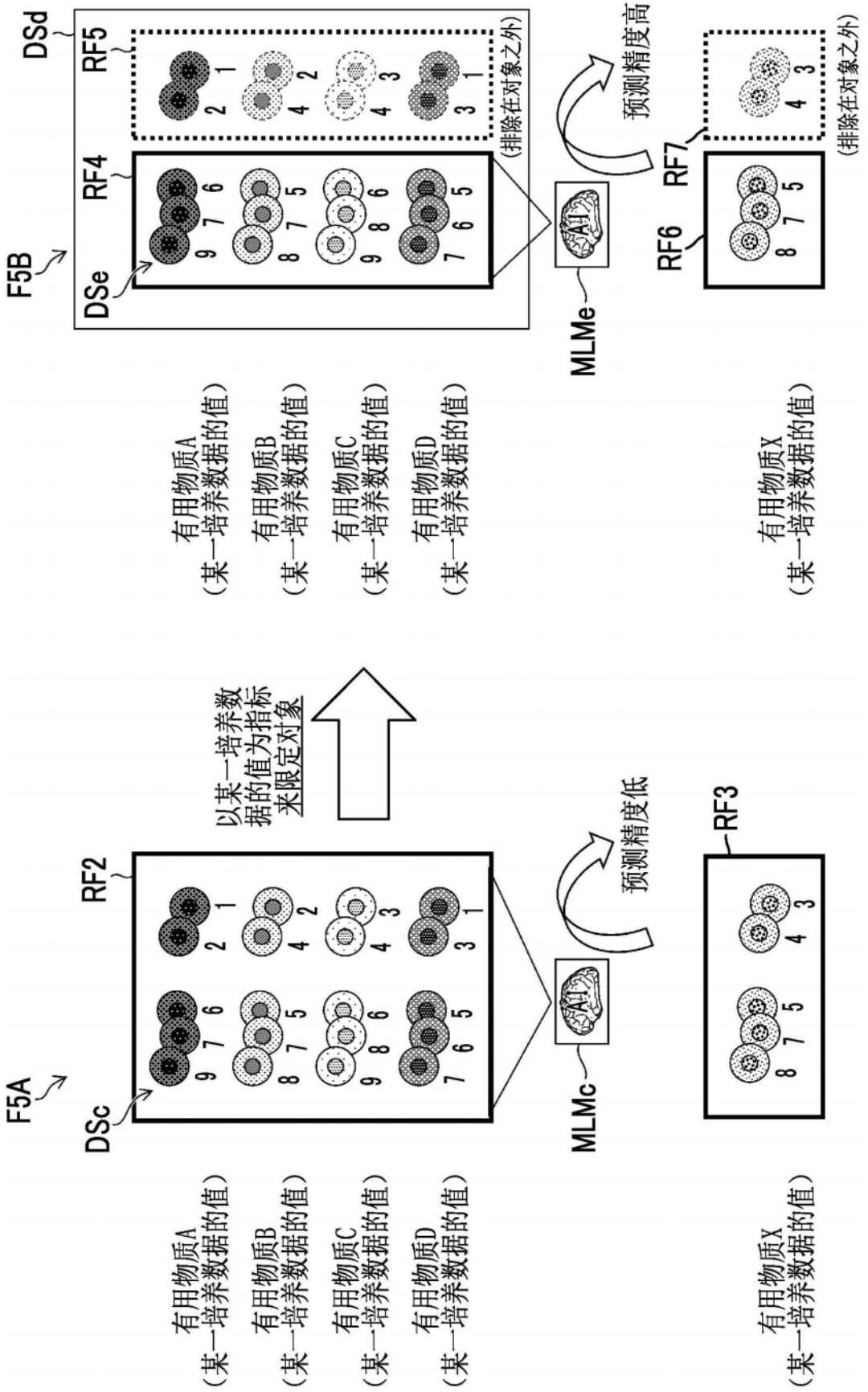

F5A
DSc
RF2
9 7 6
8 7 5
9 8 6
7 6 5
2 1
4 2
4 3
3 1
有用物质A
（某一培养数据的值）
有用物质B
（某一培养数据的值）
有用物质C
（某一培养数据的值）
有用物质D
（某一培养数据的值）
MLMc
AI
预测精度低
RF3
8 7 5
4 3
有用物质X
（某一培养数据的值）
以某一培养数据的值为指标来限定对象
F5B
DSe
RF4
DSd
RF5
(排除在对象之外)
MLMe
预测精度高
RF6
RF7
(排除在对象之外)

图5

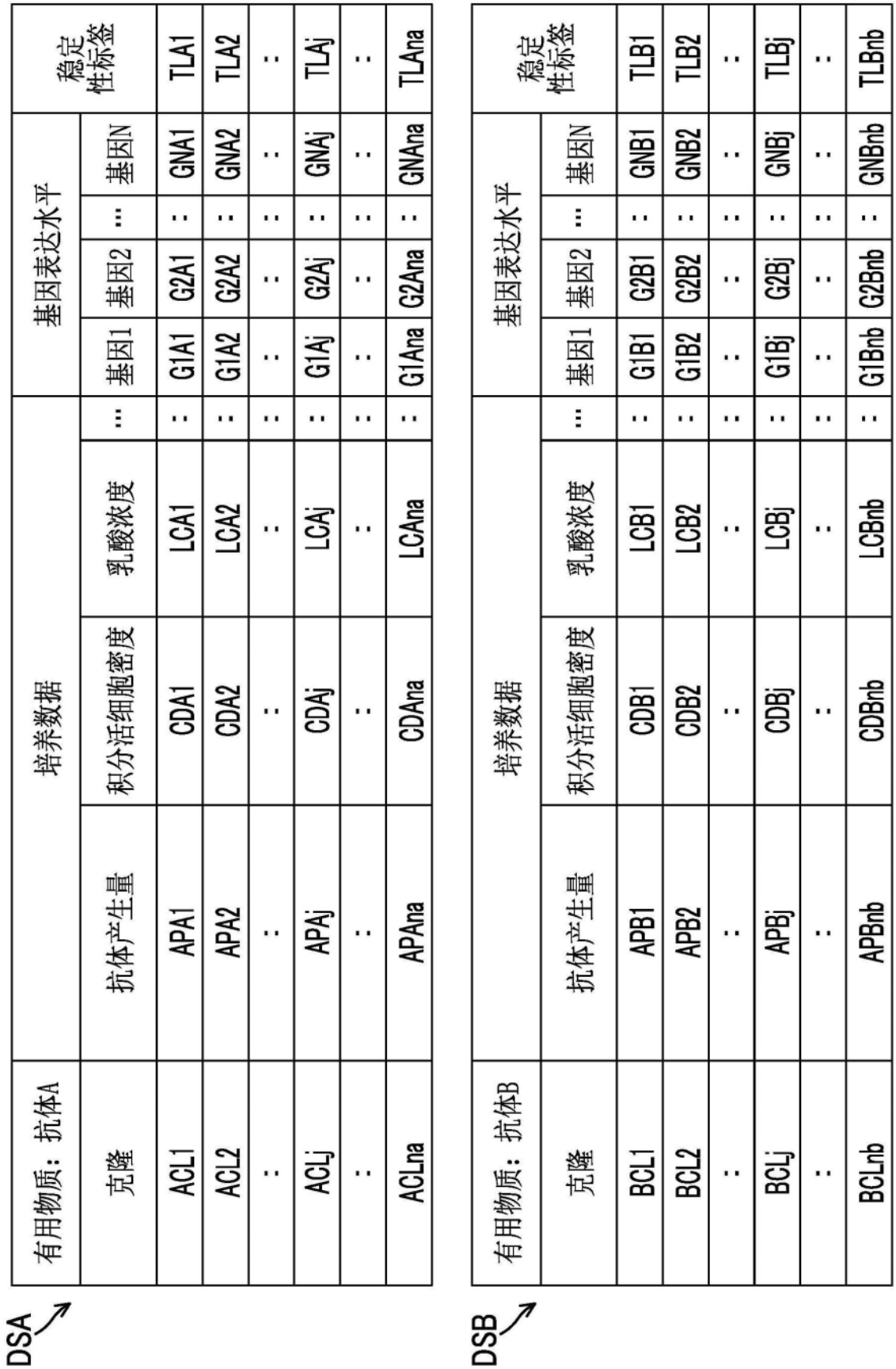

DSA
有用物质：抗体A
培养数据
基因表达水平
稳定性标签
克隆
抗体产生量
积分活细胞密度
乳酸浓度
...
基因1
基因2
...
基因N
ACL1 APA1 CDA1 LCA1 ... G1A1 G2A1 ... GNA1 TLA1
ACL2 APA2 CDA2 LCA2 ... G1A2 G2A2 ... GNA2 TLA2
... ...
ACLj APAj CDAj LCAj ... G1Aj G2Aj ... GNAj TLAj
... ...
ACLna APAna CDAna LCAna ... G1Ana G2Ana ... GNAna TLAna
DSB
有用物质：抗体B
培养数据
基因表达水平
稳定性标签
克隆
抗体产生量
积分活细胞密度
乳酸浓度
...
基因1
基因2
...
基因N
BCL1 APB1 CDB1 LCB1 ... G1B1 G2B1 ... GNB1 TLB1
BCL2 APB2 CDB2 LCB2 ... G1B2 G2B2 ... GNB2 TLB2
... ...
BCLj APBj CDBj LCBj ... G1Bj G2Bj ... GNBj TLBj
... ...
BCLnb APBnb CDBnb LCBnb ... G1Bnb G2Bnb ... GNBnb TLBnb

图6

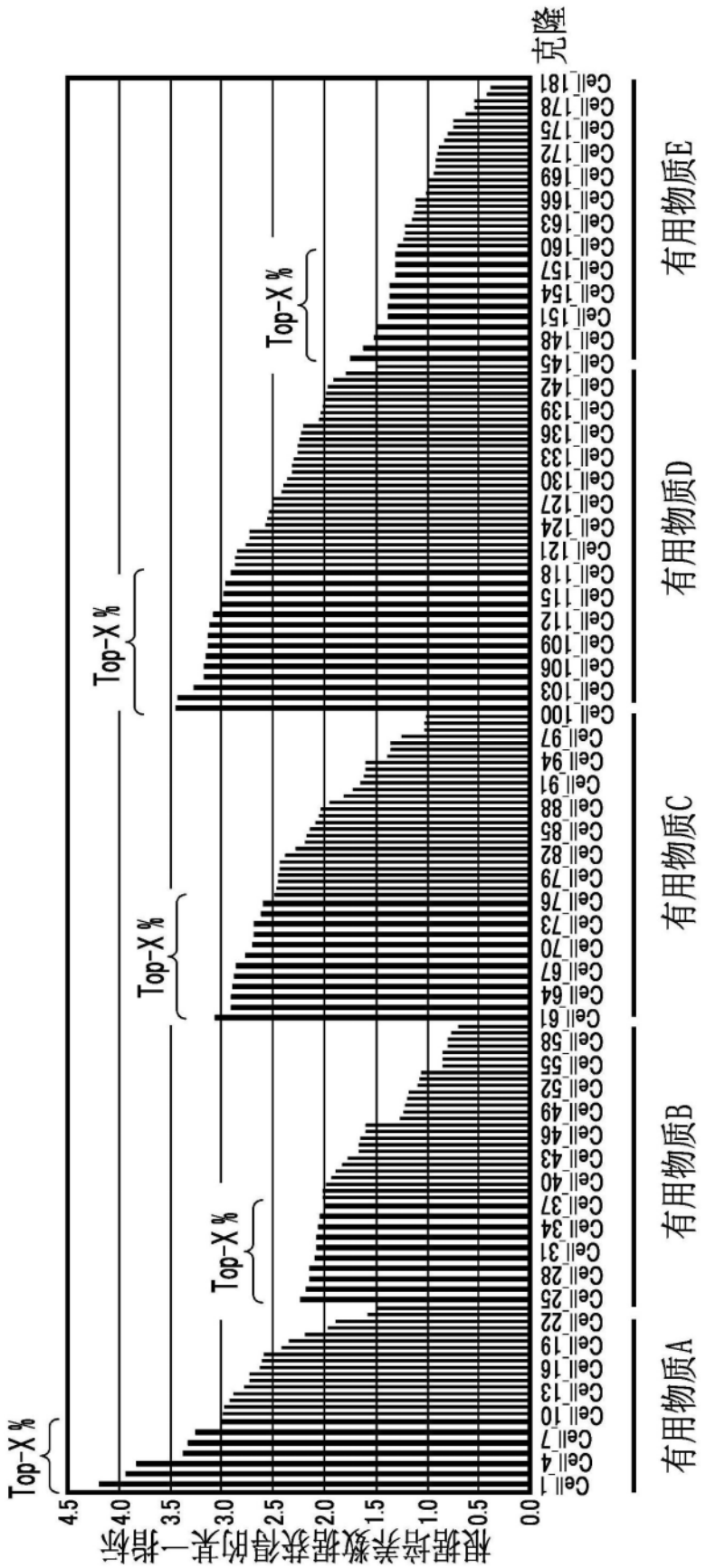

克隆
根据后续数据获得的其一指标
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Top-X %
Top-X %
Top-X %
Top-X %
有用物质A
有用物质B
有用物质C
有用物质D
有用物质E

图7

抗体产生CHO细胞的克隆数与稳定性标签

	抗体A	抗体B	抗体C	抗体D	抗体E	共计
稳定	7	5	20	21	23	76
不稳定	17	32	20	24	13	106
共计	24	37	40	45	36	182

图8

限定为相对高产时的克隆数与稳定性标签

	抗体A	抗体B	抗体C	抗体D	抗体E	共计
稳定	5	0	11	3	13	32
不稳定	5	14	5	15	2	41
共计	10	14	16	18	15	73

图9

限定为相对高细胞密度时的克隆数与稳定性标签

	抗体A	抗体B	抗体C	抗体D	抗体E	共计
稳定	7	5	15	15	15	57
不稳定	9	17	9	11	6	52
共计	16	22	24	36	21	109

图10

限定为相对高乳酸浓度时的克隆数与稳定性标签

	抗体A	抗体B	抗体C	抗体D	抗体E	共计
稳定	5	1	7	15	8	36
不稳定	5	12	9	4	6	36
共计	10	13*	16	19	14	72

(*数据缺失1个克隆)

图11

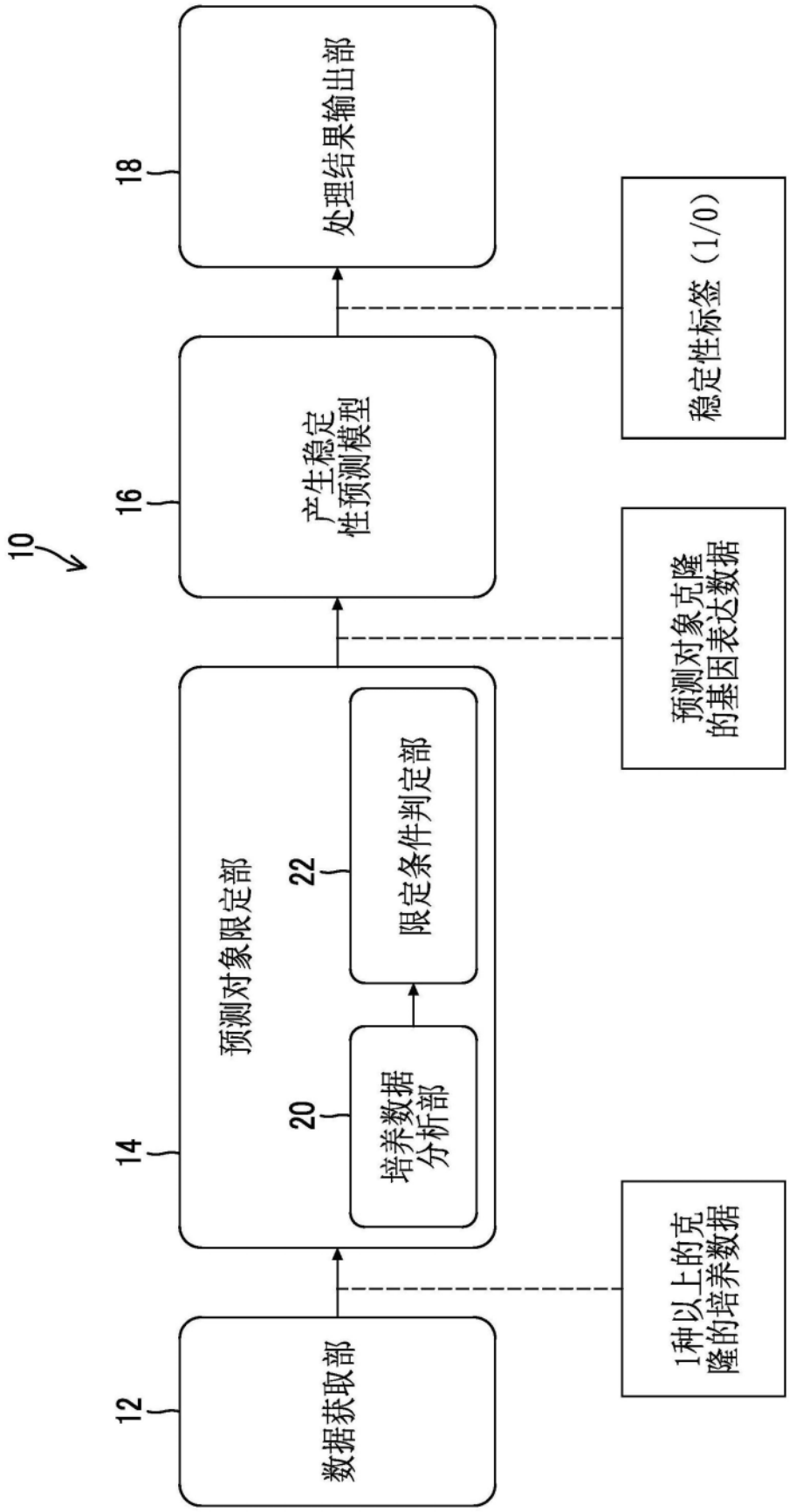

10
12
数据获取部
14
预测对象限定部
20
培养数据分析部
22
限定条件判定部
16
产生稳定性预测模型
18
处理结果输出部
1种以上的克隆的培养数据
预测对象克隆的基因表达数据
稳定性标签（1/0）

图12

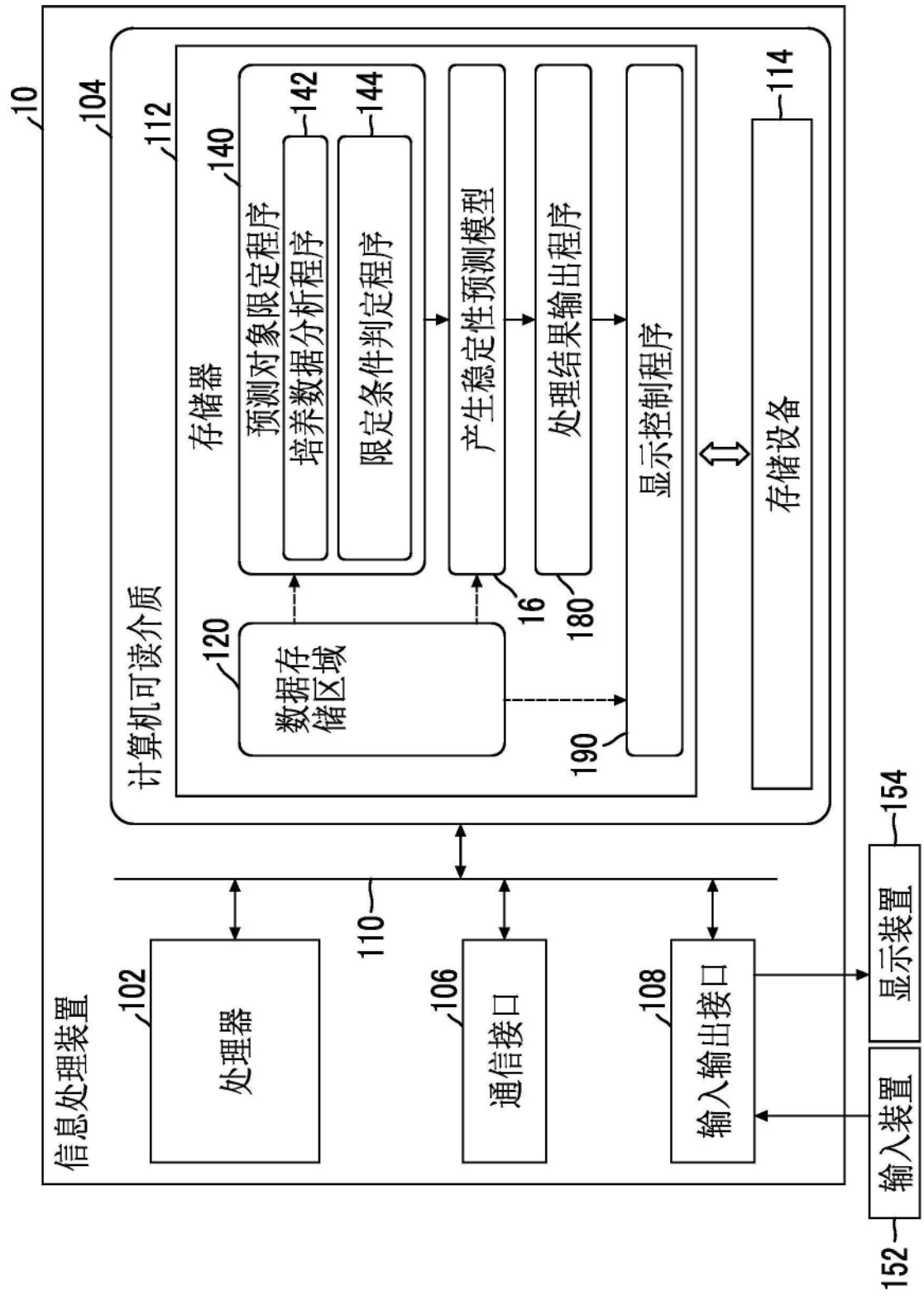

10
信息处理装置
104
计算机可读介质
112
存储器
140
预测对象限定程序
142
培养数据分析程序
144
限定条件判定程序
产生稳定性预测模型
16
处理结果输出程序
180
显示控制程序
190
120
数据存储区域
114
存储设备
110
102
处理器
106
通信接口
108
输入输出接口
154
显示装置
152
输入装置

图13

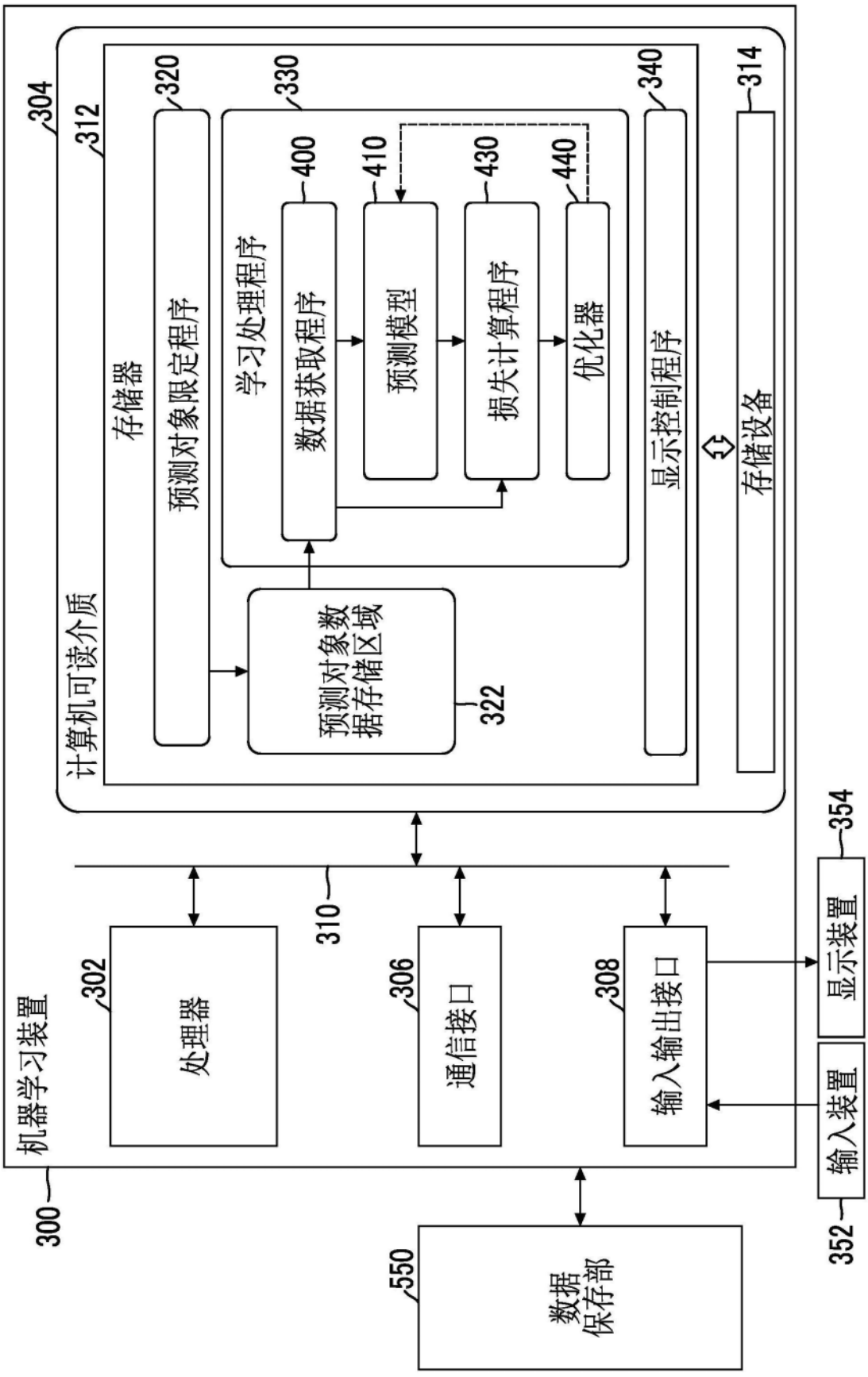

机器学习装置
300
304
计算机可读介质
存储器
312
预测对象限定程序
320
学习处理程序
330
数据获取程序
400
预测模型
410
损失计算程序
430
优化器
440
预测对象数据存储区域
322
显示控制程序
340
存储设备
314
处理器
302
310
通信接口
306
输入输出接口
308
显示装置
354
输入装置
352
数据保存部
550

图14

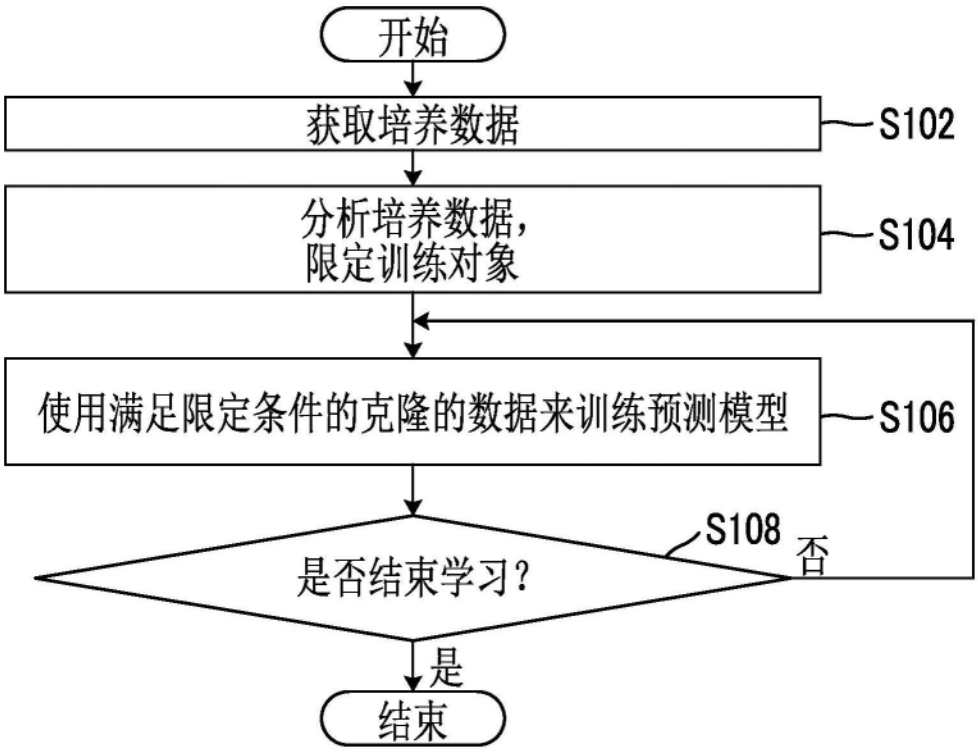

开始
获取培养数据
S102
分析培养数据，
限定训练对象
S104
使用满足限定条件的克隆的数据来训练预测模型
S106
S108
是否结束学习？
否
是
结束

图15

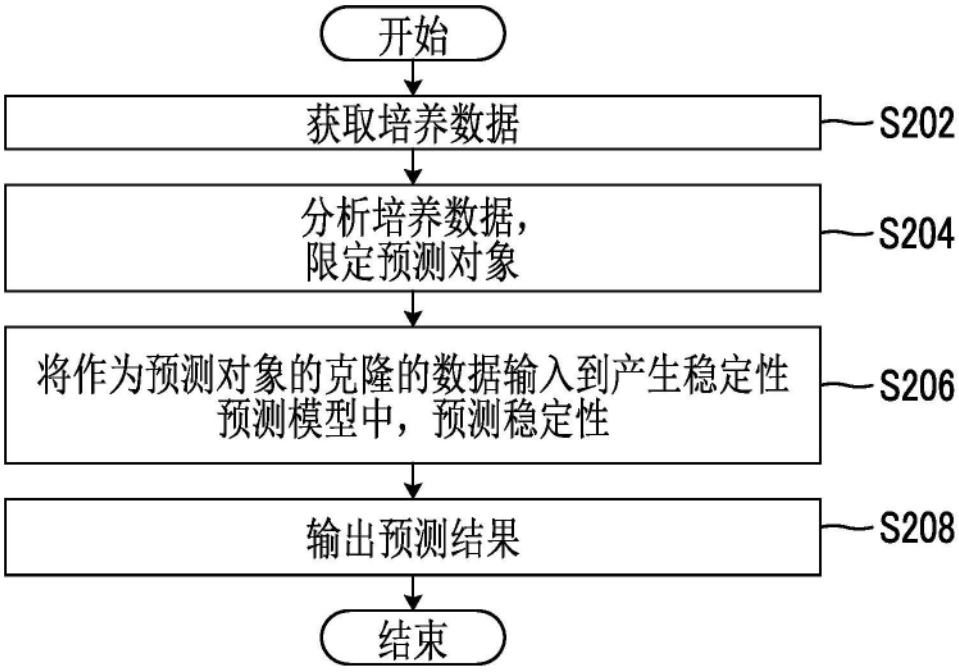

开始
获取培养数据
S202
分析培养数据，
限定预测对象
S204
将作为预测对象的克隆的数据输入到产生稳定性
预测模型中，预测稳定性
S206
输出预测结果
S208
结束

图16