(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 118974834 A
(43)申请公布日 2024.11.15

(21)申请号 202380031812.X
(22)申请日 2023.03.24
(30)优先权数据
2022-056626 2022.03.30 JP
(85)PCT国际申请进入国家阶段日
2024.09.29
(86)PCT国际申请的申请数据
PCT/JP2023/011772 2023.03.24
(87)PCT国际申请的公布数据
WO2023/190136 JA 2023.10.05
(71)申请人 富士胶片株式会社
地址 日本
(72)发明人 J·辛格
(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 牛玉婷
(51)Int.Cl.
G16B 40/00(2006.01)
C12M 1/00(2006.01)
C12M 1/34(2006.01)
C12N 15/11(2006.01)
C12Q 1/6844(2006.01)
C12Q 1/6869(2006.01)
G01N 33/50(2006.01)
G06N 20/00(2006.01)
G16B 30/00(2006.01)

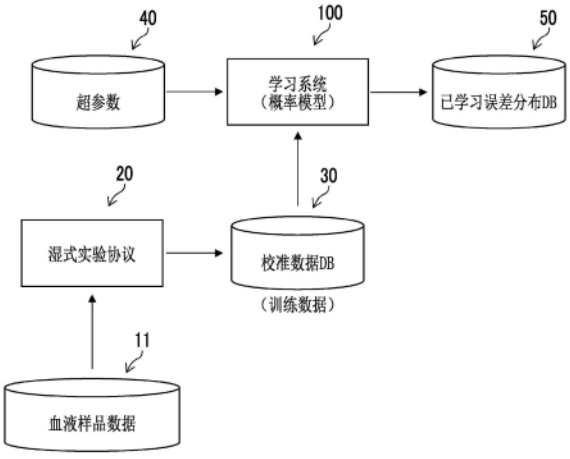
权利要求书3页 说明书12页 附图8页

(54)发明名称

学习系统、确定系统和预测系统以及学习方法、确定方法和预测方法

(57)摘要

本发明的一方式提供一种学习系统、确定系统和预测系统以及学习方法、确定方法和预测方法。在DNA的甲基化测定中,存在亚硫酸氢盐转化的不完全性的问题(问题1)、在几个不同的生物标记序列/基因一起扩增时产生偏差的问题(问题2)及非甲基化信号的过度扩增的程度依赖于基因序列本身和用于测定的化学物质这一问题(问题3)。在本发明的一方式中,提供一种在存在3个问题时学习测定误差特性并且使所学习的误差特性反映到生物标记选择基准的系统和与该系统对应的方法。解决在存在问题1~3的组合时的对DNA的甲基化的测定误差特性评价的问题,形成本发明的主要新颖性。



CN 118974834 A

1.一种学习系统,其学习测定协议变量与作为生物标记序列的结果而产生的误差特性的关系,所述学习系统具备处理器,

所述处理器进行如下处理:

输入以能够针对具有重要性的变量获取适当的数据的方式设计的校准数据;及

使用概率模型,学习针对所述具有重要性的变量遍及各测定协议的误差分布的特性,

所述概率模型包括:

第1参数,为了对亚硫酸氢盐转化的误差进行模型化,用适当选择的先验参数进行了初始化;

第2参数,为了对生物标记序列的扩增的相互依赖性进行模型化,用适当选择的先验参数进行了初始化;及

第3参数,为了对PCR整体的偏差进行模型化,用适当选择的先验参数进行了初始化。

2.根据权利要求1所述的学习系统,其中,

所述第2参数为如下参数:分别获取亚硫酸氢盐转化后的基因的甲基化序列及非甲基化序列的计数,针对所述甲基化序列及所述非甲基化序列的各序列,将所获取的计数以能够分别确定先验变量的多项分布进行了模型化。

3.根据权利要求1或2所述的学习系统,其中,

所述第3参数为如下参数:在使用通用引物同时扩增多个序列的情况下,施加了以多项分布计算的计数的各个计数的合计遵循高斯分布这一结构数据限制。

4.一种确定系统,其具备处理器,其中,

所述处理器进行如下处理:

输入在多重化面板中使用的关注生物标记序列的核苷酸序列和测定协议信息;

从权利要求1至3中任一项所述的学习系统输入所学习的所述误差特性和与所述误差特性建立关联的元数据;

使用利用预先确定的基准进行所述输入的所述核苷酸序列、所述测定协议信息、所学习的所述误差特性和所述元数据来输出用于集合可能的生物标记序列的第1分数;及

考虑针对各集合的所述第1分数的值来确定生物标记序列集。

5.根据权利要求4所述的确定系统,其中,

所述处理器进行如下处理:

针对每个应确定的生物标记序列输入第2分数;及

考虑针对所述生物标记序列集中的各生物标记序列的所述第1分数,将所述第1分数与所述第2分数的平衡最佳化,由此选择所述多重化面板的最佳子集。

6.一种预测系统,其预测基因序列的测定误差特性,所述预测系统具备处理器,

所述处理器进行如下处理:

输入在多重化面板中使用的关注生物标记序列的核苷酸序列和测定协议信息;

从权利要求1至3中任一项所述的学习系统输入所学习的所述误差特性和与所述误差特性建立关联的元数据;

使用用于计算2个基因序列之间的相似性的尺度的测定基准来计算以前包含在校准数据中的生物标记序列与新的生物标记序列的相似度;及

将计算出的所述相似度与其他相关的输入和所学习的所述误差特性组合使用而预测

测定不包含在所述校准数据中的生物标记序列时的误差特性。

7.根据权利要求6所述的预测系统,其中,

所述处理器进行如下处理:

使用所预测的所述误差特性来获取与不包含在所述校准数据中的生物标记序列最相似的、能够在所述校准数据中利用的生物标记序列;及

将所获取的所述生物标记序列的信息反映到权利要求4或5所述的确定系统中的生物标记序列集的确定中。

8.一种学习方法,其由具备处理器且学习测定协议变量与作为生物标记序列的结果而产生的误差特性的关系的学习系统执行,其中,

所述处理器进行如下处理:

输入以能够针对具有重要性的变量获取适当的数据的方式设计的校准数据;及

使用概率模型,学习针对所述具有重要性的变量遍及各测定协议的误差分布的特性,

所述概率模型包括:

第1参数,为了对亚硫酸氢盐转化的误差进行模型化,用适当选择的先验参数进行了初始化;

第2参数,为了对生物标记序列的扩增的相互依赖性进行模型化,用适当选择的先验参数进行了初始化;及

第3参数,为了对PCR整体的偏差进行模型化,用适当选择的先验参数进行了初始化。

9.根据权利要求8所述的学习方法,其中,

所述第2参数为如下参数:分别获取亚硫酸氢盐转化后的基因的甲基化序列及非甲基化序列的计数,针对所述甲基化序列及所述非甲基化序列的各序列,将所获取的计数以能够分别确定先验变量的多项分布进行了模型化。

10.根据权利要求8或9所述的学习方法,其中,

所述第3参数为如下参数:在使用通用引物同时扩增多个序列的情况下,施加了以多项分布计算的计数的各个计数的合计遵循高斯分布这一结构数据限制。

11.一种确定方法,其由具备处理器的确定系统执行,其中,

所述处理器进行如下处理:

输入在多重化面板中使用的关注生物标记序列的核苷酸序列和测定协议信息;

输入作为权利要求8至10中任一项所述的学习方法的结果获得的、所学习的所述误差特性和与所述误差特性建立关联的元数据;

使用利用预先确定的基准进行所述输入的所述核苷酸序列、所述测定协议信息、所学习的所述误差特性和所述元数据来输出用于集合可能的生物标记序列的第1分数;及

考虑针对各集合的所述第1分数的值来确定生物标记序列集。

12.根据权利要求11所述的确定方法,其中,

所述处理器进行如下处理:

针对每个应确定的生物标记序列输入第2分数;及

考虑针对所述生物标记序列集中的各生物标记序列的所述第1分数,将所述第1分数与所述第2分数的平衡最佳化,由此选择所述多重化面板的最佳子集。

13.一种预测方法,其由具备处理器且预测基因序列的测定误差特性的预测系统执行,

其中,

所述处理器进行如下处理:

输入在多重化面板中使用的关注生物标记序列的核苷酸序列和测定协议信息;

输入通过权利要求8至10中任一项所述的学习方法获得的、所学习的所述误差特性和与所述误差特性建立关联的元数据;

使用用于计算2个基因序列之间的相似性的尺度的测定基准来计算以前包含在校准数据中的生物标记序列与新的生物标记序列的相似度;及

将所述计算出的相似度与其他相关的输入和所学习的所述误差特性组合使用而预测测定不包含在所述校准数据中的生物标记序列时的误差特性。

14.根据权利要求13所述的预测方法,其中,

所述处理器进行如下处理:

使用所预测的所述误差特性来获取与不包含在所述校准数据中的生物标记序列最相似的、能够在所述校准数据中利用的生物标记序列;及

将所获取的所述生物标记序列的信息反映到权利要求11或12所述的确定方法中的生物标记序列集的确定中。

学习系统、确定系统和预测系统以及学习方法、确定方法和预测方法

技术领域

[0001] 本发明涉及一种测定生物标记的值的技术。

背景技术

[0002] 在DNA(deoxyribonucleic acid:脱氧核糖核酸)中,已知发生被称为“甲基化”的现象。甲基化是指基于甲基分子与胞嘧啶化学键合的修饰。该胞嘧啶(C:cytosine)与鸟嘌呤(G:guanine)、腺嘌呤(A:adenine)、胸腺嘧啶(T:thymine)一起构成了构成DNA的4个必需核酸碱基。核酸碱基的任意序列被称为“核苷酸序列”,对蛋白质等的重要信息进行编码的核苷酸序列被称为“基因组序列”或“基因”。

[0003] 在人体中,在DNA链上胞嘧啶紧与鸟嘌呤连接的位置(称为“CpG位点”)上,甲基化尤其常见。甲基化状态影响基因的激活或抑制化,某些基因的CpG位点的甲基化状态形成许多疾病的重要的生物标记。通常,为了制作疾病诊断的定量模型,使用从几个生物标记候选序列的组合获得的数据。因此,测定生物标记的DNA甲基化变得重要。

[0004] 在DNA的测定过程中,数据出现错误,对任何推测/预测的可靠性都会带来影响。用于将生物标记的选择最佳化的以往的研究在测定工艺中假设仅有极少的错误,仅将焦点集中在可利用的数据的预测值上。作为这种方法的例子,已知有一种特征选择算法,其依赖于来自(Artificial Intelligence(人工智能)分类器的性能之类的)定量模型的输出信号来确定是否使用生物标记序列作为用于分类的特征。

[0005] 关于这种以往的技术,例如在专利文献1中记载有从代表性的生物标记数据选择生物标记集进行评价的内容。并且,在非专利文献1中记载有PCR偏差(PCR:polymerase chain reaction(聚合酶链式反应))的测定及松弛。

[0006] 以往技术文献

[0007] 专利文献1:日本特表2017-523437号公报

[0008] 非专利文献

[0009] 非专利文献1:“Measuring and Mitigating PCR Bias in Microbiome Data”、Justin D.Silverman等、[2022年3月22日检索]、互联网(https://www.biorxiv.org/content/10.1101/604025v1)

发明内容

[0010] 发明要解决的技术课题

[0011] 在下一部分中,对所要学习生物标记序列(sequence:序列)的测定误差特性的现有研究进行详细讨论。对这些现有研究和与它们相关的问题进行讨论,并进行各个阶段的详细说明。

[0012] [DNA的甲基化测定]

[0013] 将甲基化测定的概要示于图1中。在甲基化的测定中,血液样品10被亚硫酸氢盐转

化,通过PCR装置扩增基因/信号,并通过新时代测序仪等进行测定。这些一系列的测定顺序构成湿式实验协议20(wet experiment protocol)。

[0014] [STEP1:亚硫酸氢盐转化]

[0015] 为了区分Cm(甲基化胞嘧啶)和Cu(非甲基化胞嘧啶),使用亚硫酸氢盐转化(Bisulfite conversion)的追加步骤。在亚硫酸氢盐转化中,Cu转化为尿嘧啶(U:uracil),Cm仍为Cm。若所转化的样品被序列化,则Cm作为C(胞嘧啶)而读出,另一方面,尿嘧啶作为胸腺嘧啶而读出。由此,能够区分胞嘧啶的甲基化状态。

[0016] [问题1:亚硫酸氢盐转化中的问题]

[0017] 该顺序的理想结果为Cu被转化为100%的尿嘧啶、Cm完全不会转化为尿嘧啶(转化为0%,Cm仍为Cm)。但是,在化学反应的性质上,转化的成功(或不成功)的程度是概率论的,定量研究是困难的。以下将这种亚硫酸氢盐转化的不完全性称为“问题1”。

[0018] [STEP2:PCR扩增]

[0019] 该阶段可以理解为测定的信号扩增阶段。标准上(即,不是为了进行甲基化而是为了进行亚硫酸氢盐转化),各个“信号”是感兴趣的基因或序列。在原始数据中,这种序列的数量非常少,因此所派生的信号较弱。因此,认为通过多次复制原始序列,能够增加序列数而扩增信号。例如,将PCR前的基因1的信号强度称为G1_pre,将PCR后的信号强度称为G1_post。另外,实际上,将焦点集中在同时扩增多个基因/信号上。因此,关于基因2,以与基因1相同的方式定义G2_pre和G2_post。

[0020] 现在,若首先进行上述STEP1,则即使是仅1个基因,也可获得2个信号。例如,基因1具有几个被转化为包含尿嘧啶的其他序列的、作为非甲基化具有CpG的序列。同样地,CpG被甲基化的序列不会被转化。这是常见的,在肝脏和胃的DNA的混合物中被发现。在这种混合物中,有可能对肝脏重要的基因在肝细胞中未被甲基化,但在胃细胞中被甲基化(因此被抑制)。因此,关于基因1,将PCR前信号的强度和PCR后信号的强度设为G1_U_Pre及G1_U_post(在未被甲基化的情况)、设为G1_M_Pre及G1_M_post(在被甲基化的情况),将所解密的序列设为G1_M_Pre及G1_M_post。

[0021] [问题2:单亚硫酸氢盐协议中的PCR偏差]

[0022] 即使扩增相同基因的信号,亚硫酸氢盐转化也会成为2个信号类型。因此,G1_U_post/G1_U_pre=G1_M_post/G1_M_pre不成立。已知即使在G1_U_pre=G1_M_pre的情况下,扩增后成为G1_U_post/G1_U_pre>G1_M_post/G1_M_pre(即,非甲基化基因相对于甲基化基因过度扩增)。但是,这种非甲基化信号的过度扩增的程度依赖于基因序列本身和测定中使用的化学物质。以下,将该问题称为“问题2”。

[0023] [问题3:PCR扩增中的问题]

[0024] PCR的理想结果为G1_post/G1_pre=G2_post/G2_pre。但是,实际上,某种基因序列比其他基因序列容易测定而该等价性不成立。将这种在一起扩增了几个不同的生物标记序列/基因的情况下产生的偏差称为多重化协议中的“PCR偏差”(以下,称为“问题3”)。

[0025] [以往技术中的对应]

[0026] 对上述的问题1~3的以往技术中的对应进行说明。在以往技术中,关于问题1,定量研究中通常不需要极端的准确度,因此未考虑亚硫酸氢盐转化的成功程度。并且,关于问题3,在迄今为止的微生物学的研究中,以乘法的方式考虑PCR的效果。即,在以往的技术中,

认为若1次PCR循环后的基因1的信号强度为j,则2次循环后的信号强度为j^2,x次循环后的信号强度同样为j^x。使用该假设,PCR被模型化为使用了多项逻辑-通常线性模型的对数线性过程。“批次效应”(针对每个批次显示出稍微不同的偏差特性的样本的PCR)等其他协变量也包含在概率论方法中。模型在“训练”所生成的校准数据之后,用于校准PCR偏差。

[0027] 并且,关于问题2,单一协议设定中的测定误差和偏差的特征化更简单,因此在一部分的PCR数据中,为了发现偏差的程度而进行线性回归。在计算出线性回归估计量之后,能够使用该方程式来校准这种偏差。

[0028] DNA甲基化的准确测定的重要性已经叙述。在如疾病诊断的应用领域中,使用多个生物标记序列的数据并输入到定量模型中并不少见。在设计同时测定多个生物标记的指标值的测定工艺时,问题1、问题2、问题3被组合在一起,这些全部成为问题。因此,误差的定量化和学习误差的特性变得非常困难。在本发明中,对在存在3个问题的情况下学习测定误差特性并且使所学习的误差特性反映到生物标记选择基准的系统进行研究。在存在该问题的组合的情况下,解决针对DNA的甲基化的测定误差特性评价的问题,形成本发明的主要新颖性。

[0029] 本发明在如液体生命学(liquid biopsy)那样需要同时且非常准确地测定来自多个基因的DNA甲基化的情况下尤为重要。尤其,已知为了准确鉴定如癌症的疾病,某种癌细胞基因与健康细胞中的相同基因相比显示出高的甲基化。在这种情况下,问题2是指测定时从癌症和正常的DNA的混合低估了真实的甲基化比(负的偏差)。问题1和问题3进一步使低估的程度恶化。

[0030] 本发明是鉴于上述情况而完成的,其一方式提供一种学习生物标记序列的测定误差特性的学习系统及学习方法。并且,本发明的一方式提供一种反映所学习的误差特性来确定序列集的确定系统和确定方法、以及使用通过学习系统或学习方法获得的数据来预测基因序列的测定误差特性的预测系统和预测方法。

[0031] 用于解决技术课题的手段

[0032] 本发明的第1方式所涉及的学习系统为学习测定协议变量与作为生物标记序列的结果而产生的误差特性的关系的学习系统,其具备处理器,处理器进行如下处理:输入以能够针对具有重要性的变量获取适当的数据的方式设计的校准数据;及使用概率模型,学习针对具有重要性的变量遍及各测定协议的误差分布的特性,概率模型包括:第1参数,为了对亚硫酸氢盐转化的误差进行模型化,用适当选择的先验参数进行了初始化;第2参数,为了对生物标记序列的扩增的相互依赖性进行模型化,用适当选择的先验参数进行了初始化;及第3参数,为了对PCR整体的偏差进行模型化,用适当选择的先验参数进行了初始化。第1方式所涉及的学习系统为学习测定协议变量与作为生物标记序列的结果而产生的误差特性之间的关系(被定义为模板对产物比)的系统。

[0033] 在第1方式及以下各方式中,“具有重要性的变量”为由实验室的专家已知对信号扩增性能带来影响的变量,对于这种变量,PCR装置进行调整。例如,如后述的图2所示的PCR温度或PCR循环数为“具有重要性的变量”的一例。若温度过高,则DNA被分解,不会发生为了复制目标基因序列而需要的反应。并且,关于第1～第3参数,可以使用相同的参数作为“适当地选择的先验参数”。并且,关于“校准数据(Calibration data)的输入”,例如在PCR温度的情况下,要求能够在通常的PCR中使用的温度的范围内适当的显示。

[0034] 第2方式所涉及的学习系统在第1方式中,第2参数为如下参数:分别获取亚硫酸氢盐转化后的基因的甲基化序列及非甲基化序列的计数,针对甲基化序列及非甲基化序列的各序列,将所获取的计数以能够分别确定先验变量的多项分布进行了模型化。第2方式规定用于对应于上述的问题2的第2参数的具体方式,将亚硫酸氢盐转化的误差进行模型化及修正,以使能够正确地评价生物标记序列的甲基化。在第2方式中,能够根据经验性数据分析来选择更优异的先验变量。另外,在第2方式中获取的计数能够根据如碱基序列的GC比(鸟嘌呤与胞嘧啶之比)的因素来进行模型化。

[0035] 第3方式所涉及的学习系统在第1或第2方式中,第3参数为如下参数:在使用通用引物同时扩增多个序列的情况下,施加了以多项分布计算的计数的各个计数的合计遵循高斯分布这一结构数据限制。第3方式规定用于对应于上述的问题3的第2参数的具体方式,在多个生物标记的数量多且以能够计算结构数据限制的方式简化模型化参数的情况下,多个分散计数中的各计数的合计遵循高斯分布。并且,在同时扩增多个序列的情况下,进行如各标记的计数值并非独立且合计值成为大致恒定的扩增方式,因此适合如上所述的基于多项分布的模型化。此外,在伴随亚硫酸氢盐转化的甲基化测量中,各标记物具有甲基化、非甲基化这2种状态,因此成为相对于标记数×2的计数值的模型化。

[0036] 本发明的第4方式所涉及的确定系统为具备处理器的确定系统,处理器进行如下处理:输入在多重化面板中使用的关注生物标记序列的核苷酸序列和测定协议信息;从第1至第3方式中任一项所涉及的学习系统输入所学习的误差特性和与误差特性建立关联的元数据;使用利用预先确定的基准进行输入的核苷酸序列、测定协议信息、所学习的误差特性和元数据来输出用于集合可能的生物标记序列的第1分数;及考虑针对各集合的第1分数的值来确定生物标记序列集。在第4方式所涉及的确定系统中,为了确定在多重面板中是否使用生物标记序列,使用来自第1方式所涉及的系统的输出。第1分数为源自测定精度的分数,并且为测定误差越小值越高的“低误差分数”。

[0037] 第5方式所涉及的确定系统在第4方式中,处理器进行如下处理:针对每个应确定的生物标记序列输入第2分数;及考虑针对生物标记序列集中的各生物标记序列的第1分数,将第1分数与第2分数的平衡最佳化,由此选择多重化面板的最佳子集。在第5方式所涉及的确定系统中,通过考虑多重化面板(multiplex panel)的最终目标,为了能够进行生物标记序列的更加平衡的取舍的选择,增强第4方式。例如与所要预测的疾病的相关性越大,第2分数为越高的分数(相关性分数)。并且,例如通过算出由第1分数与第2分数的相加平均或相乘平均规定的第3分数并且将该第3分数最大化,能够使“第1分数与第2分数的平衡”最佳化。

[0038] 本发明的第6方式所涉及的预测系统为预测基因序列的测定误差特性的预测系统并且具备处理器,处理器进行如下处理:输入在多重化面板中使用的关注生物标记序列的核苷酸序列和测定协议信息;从第1至第3方式中任一项所涉及的学习系统输入所学习的误差特性和与误差特性建立关联的元数据;使用用于计算2个基因序列之间的相似性的尺度的测定基准来计算以前包含在校准数据中的生物标记序列与新的生物标记序列的相似度;及将计算出的相似度与其他相关的输入和所学习的误差特性组合使用而预测测定不包含在校准数据中的生物标记序列时的误差特性。第6方式所涉及的预测系统能够将第1～第3方式所涉及的学习系统用于不包含在校准数据中的生物标记序列。

[0039] 另外,在第6方式中,“其他相关的输入”例如是指与生物标记序列对应的元数据。例如,若基因类型为“启动子或增强子”、CpG类型为“岛屿、岛岸、岛架”且CG的丰富程度为“高、低”,则针对某一生物标记序列G1的这些信息的组合(元数据的一例)能够表示为“启动子、岛屿、低”这一矢量。

[0040] 第7方式所涉及的预测系统在第6方式中,处理器进行如下处理:使用所预测的误差特性来获取与不包含在校准数据中的生物标记序列最相似的、能够在校准数据中利用的生物标记序列;及将所获取的生物标记序列的信息反映到第4或第5方式所涉及的确定系统中的生物标记序列集的确定中。在第7方式中,使用第4或第5方式所涉及的确定系统,在生物标记序列集选择中,能够使用不包含在校准数据中的生物标记序列。

[0041] 本发明的第8方式所涉及的学习方法为由具备处理器且学习测定协议变量与作为生物标记序列的结果而产生的误差特性的关系的学习系统执行的学习方法,其中,处理器进行如下处理:输入以能够针对具有重要性的变量获取适当的数据的方式设计的校准数据(校准数据输入步骤);及

[0042] 使用概率模型,学习针对具有重要性的变量遍及各测定协议的误差分布的特性(学习步骤),概率模型包括:第1参数,为了对亚硫酸氢盐转化的误差进行模型化,用适当选择的先验参数进行了初始化;第2参数,为了对生物标记序列的扩增的相互依赖性进行模型化,用适当选择的先验参数进行了初始化;及第3参数,为了对PCR整体的偏差进行模型化,用适当选择的先验参数进行了初始化。第8方式规定与上述的第1方式对应的学习方法。

[0043] 第9方式所涉及的学习方法在第8方式中,第2参数为如下参数:分别获取亚硫酸氢盐转化后的基因的甲基化序列及非甲基化序列的计数,针对甲基化序列及非甲基化序列的各序列,将所获取的计数以能够分别确定先验变量的多项分布进行了模型化。第9方式规定与上述的第2方式对应的学习方法。

[0044] 第10方式所涉及的学习方法在第8或第9方式中,第3参数为如下参数:在使用通用引物同时扩增多个序列的情况下,施加了以多项分布计算的计数的各个计数的合计遵循高斯分布这一结构数据限制。第10方式规定与上述的第3方式对应的学习方法。

[0045] 本发明的第11方式所涉及的确定方法为由具备处理器的确定系统执行的确定方法,其中,处理器进行如下处理:输入在多重化面板中使用的关注生物标记序列的核苷酸序列和测定协议信息(序列信息输入步骤);输入作为第8至第10方式中任一项所涉及的学习方法的结果获得的、所学习的误差特性和与误差特性建立关联的元数据(学习结果输入步骤);使用利用预先确定的基准进行输入的核苷酸序列、测定协议信息、所学习的误差特性和元数据来输出用于集合可能的生物标记序列的第1分数(分数输出步骤);及考虑针对各集合的第1分数的值来确定生物标记序列集(序列集确定步骤)。第11方式规定与上述的第4方式对应的确定方法。

[0046] 第12方式所涉及的确定方法在第11方式中,处理器进行如下处理:针对每个应确定的生物标记序列输入第2分数(分数输入步骤);及考虑针对生物标记序列集中的各生物标记序列的第1分数,将第1分数与第2分数的平衡最佳化,由此选择多重化面板的最佳子集(子集选择步骤)。第12方式规定与上述的第5方式对应的确定方法。

[0047] 本发明的第13方式所涉及的预测方法为由具备处理器且预测基因序列的测定误差特性的预测系统执行的预测方法,其中,处理器进行如下处理:输入在多重化面板中使用

的关注生物标记序列的核苷酸序列和测定协议信息(序列信息输入步骤);输入通过第8至第10方式中任一项所涉及的学习方法获得的、所学习的误差特性和与误差特性建立关联的元数据(学习结果输入步骤);使用用于计算2个基因序列之间的相似性的尺度的测定基准来计算以前包含在校准数据中的生物标记序列与新的生物标记序列的相似度(相似度计算步骤);及将计算出的相似度与其他相关的输入和所学习的误差特性组合使用而预测测定不包含在校准数据中的生物标记序列时的误差特性(误差特性预测步骤)。第13方式规定与上述的第6方式对应的预测方法。

[0048] 第14方式所涉及的预测方法在第13方式中,处理器进行如下处理:使用所预测的误差特性来获取与不包含在校准数据中的生物标记序列最相似的、能够在校准数据中利用的生物标记序列(序列获取步骤);及将所获取的生物标记序列的信息反映到第11或第12方式所涉及的确定方法中的生物标记序列集的确定中(信息反映步骤)。第14方式规定与上述的第7方式对应的预测方法。

[0049] 另外,由处理器执行上述的方式的学习方法、确定方法和预测方法的程序(学习程序、确定程序、预测程序)和记录有这些程序的计算机可读代码的非暂时性记录介质也包括在本发明的范围内。

[0050] 发明效果

[0051] 如以上说明,本发明所涉及的学习系统、确定系统和预测系统以及学习方法、确定方法和预测方法具有以下效果。

[0052] (1)能够处理一起测定多个基因序列而多重化的面板。

[0053] (2)能够处理经亚硫酸氢盐转化的样品。

[0054] (3)能够使用序列参数和协议参数作为输入来预测测定误差。

[0055] (4)能够确定是否将序列用于分析/分类的目的。

附图说明

[0056] 图1是表示测定DNA的甲基化的情况的图。

[0057] 图2是表示制作校准数据的情况的图。

[0058] 图3是表示学习系统和与学习系统相关的数据的图。

[0059] 图4是表示学习系统的结构的图。

[0060] 图5是表示概率模型的实施方式的图。

[0061] 图6是表示确定系统与预测系统的关系的图。

[0062] 图7是表示确定系统的结构的图。

[0063] 图8是表示预测系统的结构的图。

具体实施方式

[0064] 以下,对本发明的实施方式进行说明。在说明中,根据需要参考附图。另外,在附图中,为了便于说明,有时会省略一部分构成要件的记载。

[0065] [校准数据的制作]

[0066] 在本发明中,如图2所示,首先,需要从血液样品10通过湿式实验协议20(wet experiment protocol)制作根据如PCR温度或PCR循环数的重要的测定协议变量构成的校

准数据。该校准数据优选设计为能够针对具有重要性的变量获取适当的数据。最终,将序列的测定结果与协议信息一并保存于校准数据DB30(DB:database,数据库)中(以下,有时将数据库记载为“DB”)。另外,在图2中,为了明确化,省略了校准数据制作顺序的一部分。

[0067] 为了学习这种协议变量与其测定特性之间的关系,学习算法(学习系统、学习方法)使用该校准数据。接着,对于所给定的测定协议变量的集合(不包含在校准数据中),该系统(预测系统、预测方法)能够预测所给定的生物标记序列的测定误差特性。使用该预测,系统(确定系统、确定方法)能够确定生物标记序列是否适合用于某些定量研究中。最后,即使不存在于校准数据中的生物标记序列,本发明的系统(确定系统、确定方法)也找出测定误差特性已知的最相似的序列,并且能够使用该序列对新的序列进行相似的确定。

[0068] 具体而言,本发明对概率模型的作用进行详细说明,通过估计“模板对产物”比,对生物标记序列的测定误差进行特征化。“模板”是指生物标记序列的最初的量(PCR扩增前的量),“产物”是指PCR扩增后的相同的生物标记序列的最终量(PCR扩增后的量)。

[0069] [学习系统的结构]

[0070] 图3示出本发明的一方式所涉及的学习系统100和与其相关的数据等。用于学习用于多重亚硫酸氢盐PCR协议的DNA甲基化测定误差特性的这种学习系统的适用是用于保证本发明的新颖性的最小限度的要件。另外,如后述,在学习系统100中,可以附带利用其结果(已学习误差分布DB50)的确定系统200(确定系统)和预测系统300(预测系统)。

[0071] 图4是表示学习系统100的结构例的图。如图4所示,学习系统100具备处理器110(处理器、计算机)、概率模型120(概率模型)、存储部130、ROM140(ROM:Read Only Memory,只读存储器)、RAM150(RAM:Random Access Memory,随机存取存储器)。处理器110进行学习系统100的各部所进行的处理的总括控制,因此具有校准数据输入部112和学习部114。

[0072] 处理器110除了图4所示的要件以外,还可以包括未图示的显示控制部或通信控制部、输出控制部等。

[0073] 处理器110例如由CPU(Central Processing Unit:中央处理器)、GPU(Graphics Processing Unit:图形处理器)、FPGA(Field Programmable Gate Array:现场可编程门阵列)、PLD(Programmable Logic Device:可编程逻辑器件)等各种处理器或电路构成。在这些处理器或电路执行软件(程序)时,将能够由执行的软件的计算机(例如,构成处理器的各种处理器或电路和/或它们的组合)读取的代码存储于ROM140等非暂时性且有形的记录介质中,计算机参考该软件。非暂时性且有形的记录介质中存储的软件包括用于执行本发明所涉及的学习方法、预测方法、确定方法的程序(学习程序、预测程序、确定程序)和在执行时使用的数据。可以在各种光磁记录装置、半导体存储器等非暂时性且有形的记录介质中而不是在ROM140中记录代码。在使用了软件的处理时,例如RAM150用作暂时存储区域,并且也能够参考例如存储于未图示的EEPROM(Electronically Erasable and Programmable Read Only Memory,电子可擦除可编程只读存储器)或闪存等非暂时性且有形的记录介质中的数据。也可以使用存储部130作为“非暂时性且有形的记录介质”。

[0074] 存储部130由硬盘、半导体存储器等各种存储器件及其控制部构成,能够存储上述的校准数据或学习方法的执行条件和执行结果(已学习误差分布的数据)等。

[0075] 学习系统100除了图4所示的要件以外,还可以包括未图示的显示装置(例如,液晶监视器)或操作装置(例如,鼠标或键盘)。显示装置中能够显示校准数据、误差分布的数据

等,并且,用户能够经由操作部进行本发明所涉及的学习方法(学习程序)的执行所需的操作。

[0076] 上述的图3示出血液样品数据11,但这是包括组织样品的任意生物学数据。如图1所示,血液样品数据11通过上述的STEP 1、STEP 2以及加上DNA序列确定的测定顺序来测定,其本身具有几个PCR的循环数等对其有效性带来影响的变量(具有重要性的变量)。由于需要根据这种变量的几个值来获得数据,因此最先识别相关的变量,并在这些值的范围内进行测定。例如,若在PCR循环数为唯一的重要的变量的情况下,能够生成5、10及15PCR循环的相同的血液样品的数据。这是所谓的校准数据。

[0077] [概率模型]

[0078] 学习系统100利用存储于校准数据DB30中的校准数据(训练用数据)(校准数据输入步骤),对概率模型进行训练(学习步骤)。图5中,通过贝叶斯层次模型示出了作为这种概率模型的一例的概率模型120。本发明的重要的新颖性在于,使用(i)亚硫酸氢盐转化误差的先验信息(先验参数;以下相同)、(ii)亚硫酸氢盐转化的协变量的先验信息及(iii)生物标记序列的扩增的相互依赖性的先验信息。这些先验信息(i)~(iii)与本发明的第1~第3参数对应,因此与上述问题1~3对应。若综合以上3个要件,则本发明与如上述的专利文献1或非专利文献1的以往的模型不同。

[0079] 并且,通过这3个因素,能够解决上述的问题1+问题2+问题3成为一体的问题。学习系统100按照最佳化方法(用于最小化的损失函数等)并且通过一系列的超参数(超参数40)来调整。这种调整通过确认系统的最终性能并选择使其最大化的超参数来进行。

[0080] 另外,上述的第1~第3参数为概率模型120的一部分(因此,学习系统100的一部分),这些参数的值在训练工艺中被更新。并且,第1~第3参数为学习系统100的一部分,因此在图3中未显示。

[0081] 另一方面,若使用超参数,则能够控制概率模型120的某一个侧面。但是,超参数的值由用户设定,值在训练工艺中不会被更新。并且,在学习系统100和确定系统200(参考图6)中,超参数不同。

[0082] 更具体而言,关于二项分布的生物标记,能够选择将亚硫酸氢盐转化的误差模型化。因此,能够将亚硫酸氢盐转化误差的先验概率(先验参数的一例)作为[0,1]之间的值来选择。在先验概率为0的情况下,假设其生物标记的完全转化(Cu与尿嘧啶的100%的转化和Cm的0%的转化),在先验概率大于0的情况下,假设其生物标记的不完全转化(仅Cu的一部分转化为尿嘧啶,Cm的一部分也转化为尿嘧啶)。理想的是,应该根据经验性数据分析来设定先验变量。亚硫酸氢盐协变量中包含在以纳克计测定的样品中添加的亚硫酸盐的量及初始DNA量。如此,以先验概率初始化的亚硫酸氢盐转化误差为第1参数。

[0083] 同样地,PCR误差分布能够以多项分布模型化,能够设定适当的先验概率。在该阶段中,PCR后的序列计数(序列的数量)在将所选择的生物标记的数量设为x的情况下能够表示为N1、N2、...、Nx,其序列计数能够作为多项分布而模型化。在此,“Ni”为第i个生物标记的序列计数。PCR协变量中有时包括如PCR温度或PCR循环数的为了制作校准数据而选择的因素。

[0084] 本发明的新颖性在于,从相同的序列考虑2个不同的计数的可能性的能力。一个是亚硫酸氢盐转化后的某一序列的碱基化的计数(甲基化序列的计数),另一个是该序列的非

碱基化类型的计数(非甲基化序列的计数)。由此,考虑N1_M、N1_U、N2_M、N2_U等的可能性的数量成为2倍。在此,“Ni_M”表示针对第i个生物标记的甲基化序列的计数,“Ni_U”表示针对第i个生物标记的非甲基化序列的计数。Nx_M与Nx_U为了彼此施加自然的限制(一个平均次数多是指另一个次数少),使用这种限制(相互依赖性)能够将模型化问题简化。如此,以先验概率初始化的生物标记序列的扩增的相互依赖性为第2参数。

[0085] 最后,整体分布模型(表示PCR整体的偏差的模型)是对所有生物标记即N1+N2+....+Nx的总数的序列进行量化,甚至可以用于通过生物标记计数而施加相互依赖的限制(结构数据限制)(例如,在N1过高的情况下,N3过低)。N1、N2等各自(各个计数)为多项分布,因此认为在所选择的生物标记的数量多(例如,30以上)的条件下,它们的合计(以多项分布计算的计数的各个计数的合计)满足中心极限定理,因此遵循高斯分布。序列类型之间的这种相互依赖性(结构数据限制)可能不会立即明确,但已知在使用通用引物同时扩增多个序列(多个生物标记序列)的情况下,存在这种相互依赖性。如此,以先验概率初始化的PCR整体的偏差为第3参数。

[0086] 这种通用引物的使用仅能够在适当的适配器序列配置于目标生物标记序列的两端之后进行。在该阶段追加的通用引物的有限量产生生物标记序列之间的组成依赖性,影响纯信号扩增。对于对PCR扩增中的多重化面板施加结构数据限制并使用通用引物对相对的生物标记序列的丰富程度进行模型化,形成本发明的第二新颖性。

[0087] [确定系统和预测系统的定位]

[0088] 如图6所示,在上述的学习系统100中,可以附带确定系统200(确定系统)和预测系统300(预测系统)。对于将这些确定系统和预测系统附加于学习系统100中,作为选项而被推荐。通过附加确定系统200和预测系统300,例如能够使用通过学习系统100学习的误差特性,通过确定系统200找到候选生物标记的最佳子集(包括学习结果输入步骤或分数输入步骤、子集选择步骤等,基于本发明所涉及的确定方法的执行),由此对生物标记序列的选择基准赋予信息(包括信息反映步骤等,基于本发明所涉及的预测方法的执行;预测系统300),能够帮助有效利用学习系统100。

[0089] 上述的学习系统100具备通过统计学方法将最佳化基准最大化或最小化而学习的概率模型120,对于如此进行学习,广泛涵盖了将算法进行“训练”的含义。另一方面,确定系统200在学习系统100结束训练之后发挥作用。确定系统200本身没有所要最大化或最小化的已定义的最佳化基准,因此未进行“训练”,系统未学习。但是,确定系统200包括使系统进行“可调整”的超参数而构成。

[0090] [确定系统和预测系统的结构]

[0091] 图7是表示确定系统200的结构的图。如图7所示,确定系统200具备处理器210(处理器)、ROM230(非暂时性且有形的记录介质)及RAM240。处理器210具备序列信息输入部212、学习结果输入部214、分数输出部216及序列集确定部218。确定系统200除了这些要件以外,还可以具有未图示的显示控制部或显示装置、存储装置、操作部等。

[0092] 图8是表示预测系统300的结构的图。如图8所示,预测系统300具备处理器310(处理器)、ROM330(非暂时性且有形的记录介质)及RAM340。处理器310具备序列信息输入部312、学习结果输入部314、相似度计算部316、误差特性预测部318及序列信息反映部320。预测系统300除了这些要件以外,还可以具有未图示的显示控制部或显示装置、存储装置、操

作部等。

[0093] 与学习系统100同样地,确定系统200和预测系统300的这些要件例如由CPU、GPU、FPGA、PLD等各种处理器或电路构成。在这些处理器或电路执行软件(程序)时,将能够由执行的软件的计算机读取的代码存储于ROM230或ROM330等非暂时性且有形的记录介质中,计算机参考该软件。非暂时性且有形的记录介质中存储的软件包括用于执行本发明所涉及的预测方法、确定方法的程序(预测程序、确定程序)和在执行时使用的数据。可以在各种光磁记录装置、半导体存储器等非暂时性且有形的记录介质中而不是在ROM230或ROM330中记录代码。在使用了软件的处理时,例如RAM240、RAM340用作暂时存储区域,并且也能够参考例如存储于未图示的EEPROM或闪存等非暂时性且有形的记录介质中的数据。

[0094] [基于分数的生物标记序列集的确定]

[0095] 以下,对作为“分数的最佳化工艺(最佳化方法)”这两大分类的二进制基的方法和组合基的方法进行说明。

[0096] 在二进制基的最佳化基准中,确定系统200的序列信息输入部212(处理器)输入关注生物标记序列的核苷酸序列和测定协议信息(序列信息输入步骤),学习结果输入部214(处理器)从学习系统100输入所学习的误差特性和与误差特性建立关联的元数据(学习结果输入步骤)。分数输出部216(处理器)能够独立考虑所学习的测定误差特性,根据作为结果产生的测定误差图的斜率,例如将{+1,0,-1}的分数(测定误差分数;第1分数的一例)分配到各生物标记序列(分数输出步骤)。序列集确定部218能够由各生物标记的顺序来合计分数(第1分数),并确定是否使用该组合(生物标记序列集)(序列集确定步骤)。

[0097] 若更耐用地安装它们,则能够通过与特征选择算法相同的方法来设计组合基的最佳化基准。特征选择算法依赖于定量模型的输出,为了将定量模型的性能最优化而更新这些基准。该以往的特征选择算法的观点考虑由测定错误特性产生的分数(第1分数),并且对从所赋予的生物标记序列集的子集选择带来最好的信息,因此为了使用与信号相同的分数(第1分数)而进行修正,由此能够设计本发明中所使用的组合基的最佳化基准。在组合基的最佳化基准的情况下,也与二进制基的最佳化基准的情况同样地,能够使用确定系统200的各要件来确定生物标记序列集(执行序列信息输入步骤~序列集确定步骤)。

[0098] 另外,在上述的二进制基的最佳化基准中,针对在各生物标记序列中独立地分配分数,在组合基的最佳化基准的情况下,针对生物标记序列的组合赋予分数。因此,在可以在各标记序列中独立地处理测定误差的大小的情况下,适合二进制基的最佳化基准,在相互依赖性特别大的情况下,适合组合基的最佳化基准。相互依赖性例如是指“在生物标记序列1与生物标记序列2同时测定的情况下测定误差小,但在与生物标记序列3同时测定的情况下测定误差大”这一情况。

[0099] 在本发明中,不仅能够考虑上述的测定误差分数(第1分数),还能够考虑“与所要预测的疾病的相关性越大分数越高”(相关性分数;第2分数的一例),通过将这些分数的平衡最佳化来确定生物标记序列集。在同时使用这种相关性分数(第2分数)的情况下,将上述的测定误差分数(第1分数)与相关性分数(第2分数)的平衡最佳化(例如,将测定误差分数与相关性分数的算术平均或几何平均最大化),由此能够确定生物标记序列集。在该情况下,相关性分数也能够与测定误差分数的情况同样地,按生物标记序列独立地分配,也能够赋予到生物标记序列的组合。例如,在标记1、2、3均与疾病相关的情况下,标记1、2的相关性

小且标记1、3的相关性大的情况下,标记1、2的组合对疾病预测中更有效,相关性分数变高。

[0100] 任何最佳化基准(如二进制基、特征选择算法或组合基)最好依赖于应用领域、用户及时间的限制,能够根据这些条件适当地选择。该输出为系统共同考虑的生物标记序列的集合,在用于多重化PCR序列确定的被赋予的测定误差的协议中存在最小限度的误差。由于考虑本发明的第5、第12方式,实施方式能够根据确定系统的实施来变更(无论是否考虑已平衡的序列选择)。

[0101] 另外,确定系统200依赖于在学习系统100中获得的误差分布(在图6中为已学习误差分布数据库50),其本身无法计算不包含在原始校准数据中的生物标记序列的分数。针对这种不包含在校准数据中的生物标记序列的测定误差特性的预测,如图6所示并且如以下说明,需要本发明的第6、第7方式所涉及的预测系统300(和本发明的第13、第14方式所涉及的预测方法)。该预测系统300与对于确定系统200上述的内容同样地,是对本发明所涉及的学习系统100(学习系统)的追加,依赖于这种新的生物标记序列的使用事例、存在、重要性。

[0102] [不包含在校准数据中的生物标记序列的测定误差特性的预测]

[0103] 在关注序列数据库60中所含的关注生物标记序列判明为不包含在校准数据中的生物标记序列的情况(图6的判断“是否为训练数据中所含的序列”中为“是”的情况)下,对预测该关注生物标记序列的测定误差特性的方法进行说明。在该情况下,首先将输入传递到预测系统300。具体而言,预测系统300的序列信息输入部312(处理器)输入关注生物标记序列的核苷酸序列和测定协议信息(序列信息输入步骤),并且学习结果输入部314(处理器)从学习系统100输入例如已学习误差特性和与误差特性建立关联的元数据(学习结果输入步骤)。在此,“元数据”例如为基因的类型(启动子或增强子)、基因的区域(转录起始位点等),但并不限定于此。

[0104] 并且,相似度计算部316(处理器)使用如莱文斯坦距离或GC含量(GC-content;DNA分子中的氮碱基中鸟嘌呤与胞嘧啶的比例)的2个基因序列之间的相似度的测定基准(相似性的尺度),计算以前校准数据中所含的生物标记序列(在校准数据中能够利用的生物标记序列)与新的生物标记序列(关注生物标记序列)的相似度(相似度计算步骤)。相似度计算部316从存在于已学习误差分布数据库50中的生物标记序列找出关注生物标记序列和“最相似的”生物标记序列(相似度计算步骤)。预测系统300能够使用已检测的“最相似的序列”的信息来获取与该“最近的序列”对应的已学习误差特性(从已学习误差分布数据库50)(误差特性预测步骤、序列获取步骤),由此能够完全实施本发明的第6、第13方式。序列信息反映部320还能够将该信息与实施本发明的第4、第5方式的确定系统200(和本发明的第11、第12方式所涉及的确定方法)同时使用,反映到确定系统200中的生物标记序列集的确定中(信息反映步骤)。

[0105] [实施例]

[0106] 认为基因序列的候选集首先显示出如GC含量的测定相关因素的充分变化。例如,假设仅序列GC内容是重要的,由于能够同时测定而确定“高”的3基因序列及“低”的GC含量的3基因序列。接着,确定一系列重要的测定协议相关变量,并考虑范围。然后,针对考虑到所有值和所有变量的全部范围,执行湿式实验顺序。例如,考虑{5、10、15}的PCR循环,将甲基化比率视为{5%、10%},则从同一生物标本的等分试样(aliquot)中进行3*2=6个的6个基因的独立测定。接着,对学习系统中所使用的上述概率模型进行训练,并调整(调谐)了确

定系统的超参数。现在,如癌症诊断的情况那样,能够一边寻找更多的基因生物标记,一边为了评价该基因的测定特性的良好程度而使用确定模型。

[0107] 若考虑在100个基因序列测定中进行了训练的Artificial Intelligence癌症分类模型以70%的灵敏度进行,则工作性能低的理由的一部分有可能是在一部分基因中测定噪声高。若使用上述系统重新考虑100个基因并去除测定困难的基因,则有可能通过避免测定误差而使Artificial Intelligence的性能上升至80%,具有更良好的鲁棒性。

[0108] 以上说明的实施方式具有以下效果。

[0109] (1)能够处理一起测定多个基因序列而多重化的面板。

[0110] (2)能够处理经亚硫酸氢盐转化的样品。

[0111] (3)能够使用序列参数和协议参数作为输入来预测测定误差。

[0112] (4)能够确定是否将序列用于分析/分类的目的。

[0113] (5)根据适当地学习的误差特性、适当地选择的生物标记序列集(和其子集)、高精度地预测的生物标记序列集的测定误差,能够高精度地进行使用了生物标记序列的分析或诊断(例如,上述的基于AI的癌症的分类)等。

[0114] 以上对本发明的实施方式进行了说明,但本发明并不限定于上述的方式,能够进行各种变形。

[0115] 符号说明

[0116] 10-血液样品,11-血液样品数据,20-湿式实验协议,30-校准数据DB,40-超参数,50-已学习误差分布数据库,60-关注序列数据库,100-学习系统,110-处理器,112-校准数据输入部,114-学习部,120-概率模型,130-存储部,200-确定系统,210-处理器,212-序列信息输入部,214-学习结果输入部,216-分数输出部,218-序列集确定部,300-预测系统,310-处理器,312-序列信息输入部,314-学习结果输入部,316-相似度计算部,318-误差特性预测部,320-序列信息反映部。

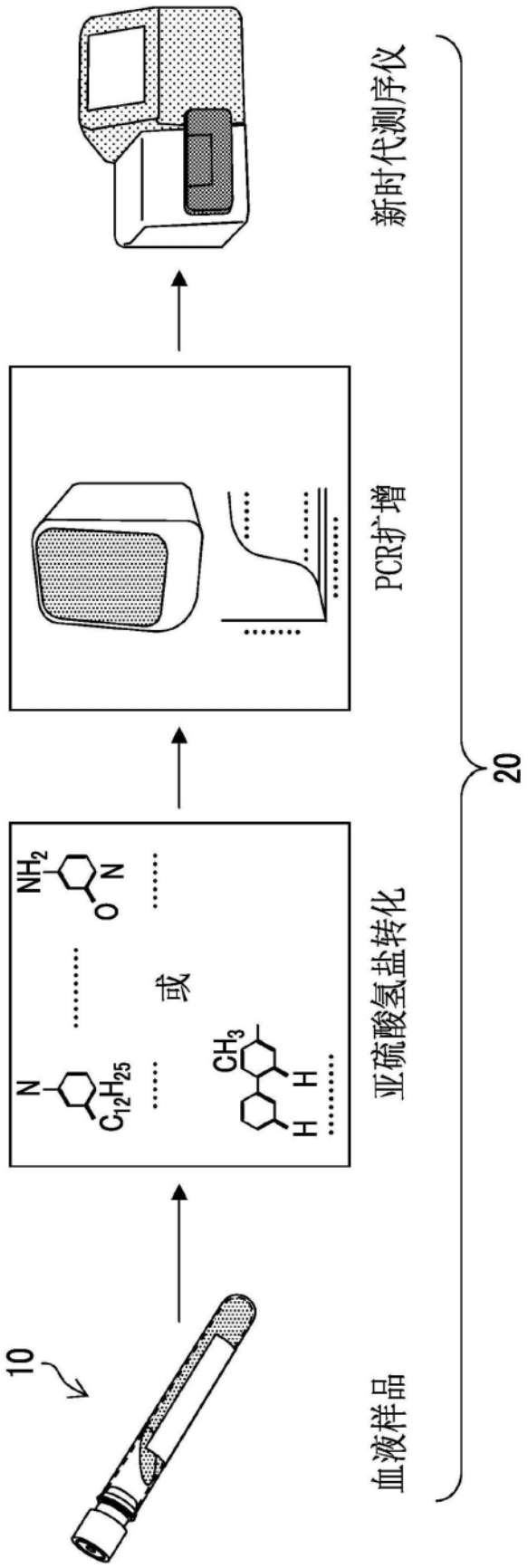


图1

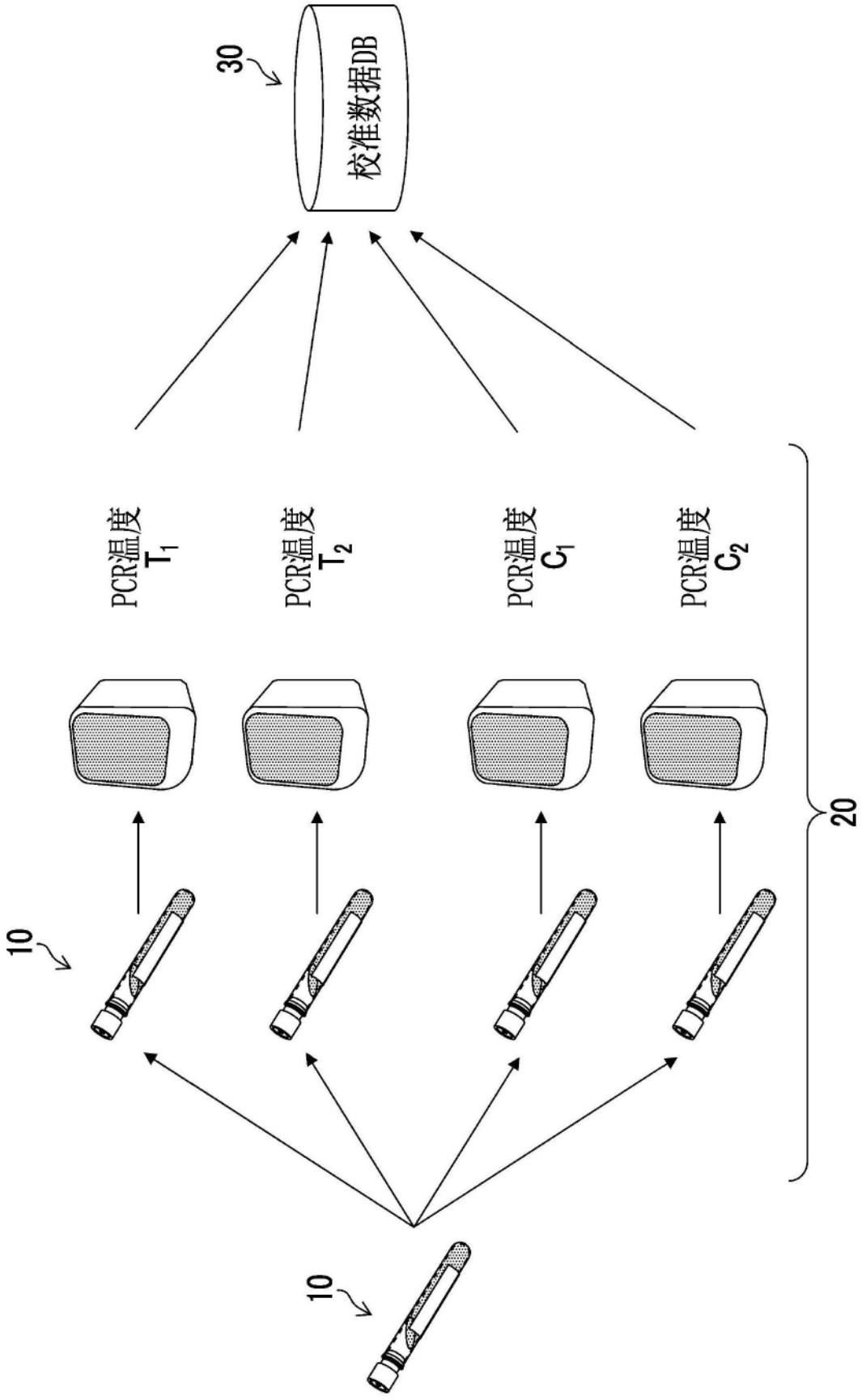


图2

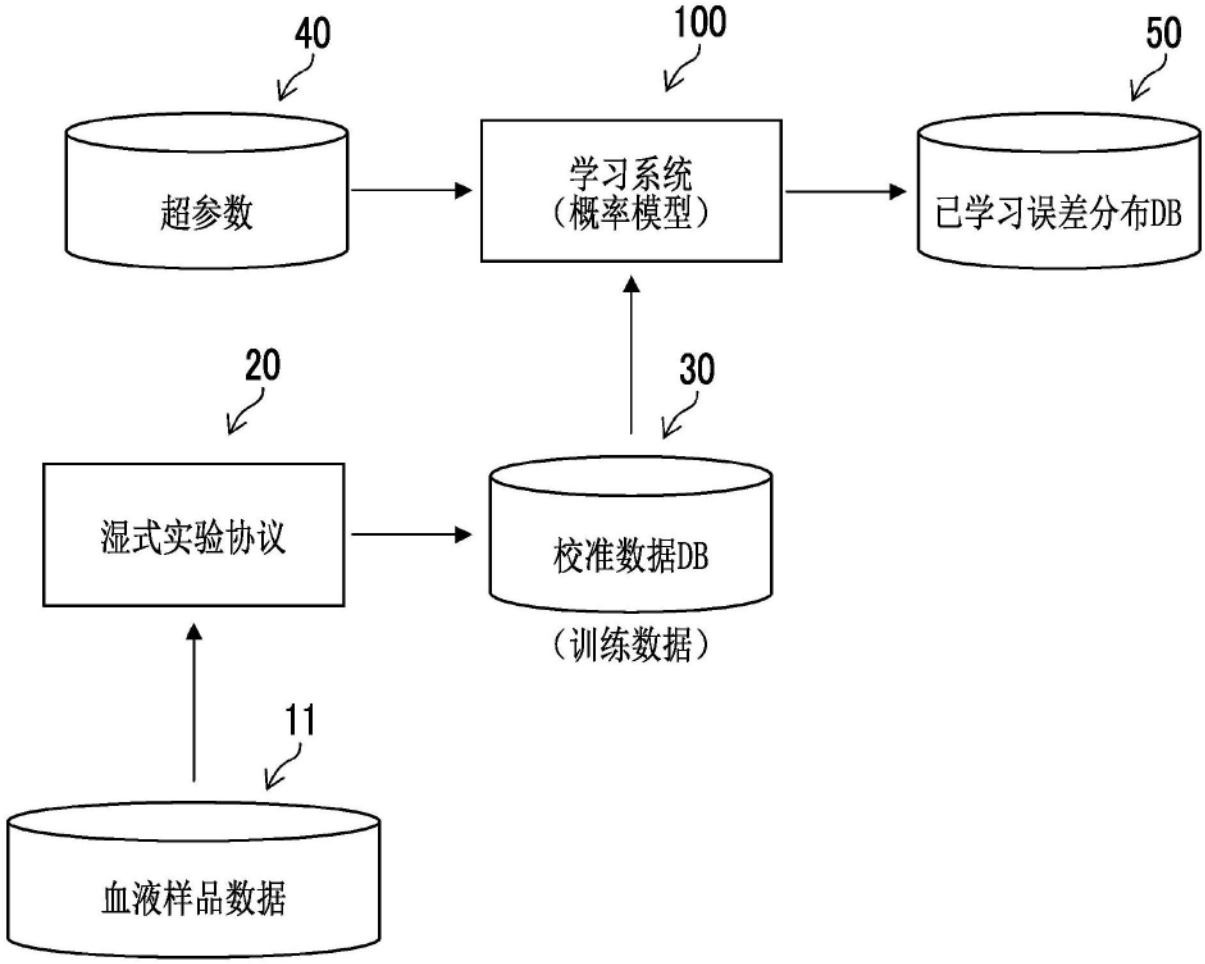

40
100
50
超参数
学习系统
（概率模型）
已学习误差分布DB
20
30
湿式实验协议
校准数据DB
（训练数据）
11
血液样品数据

图3

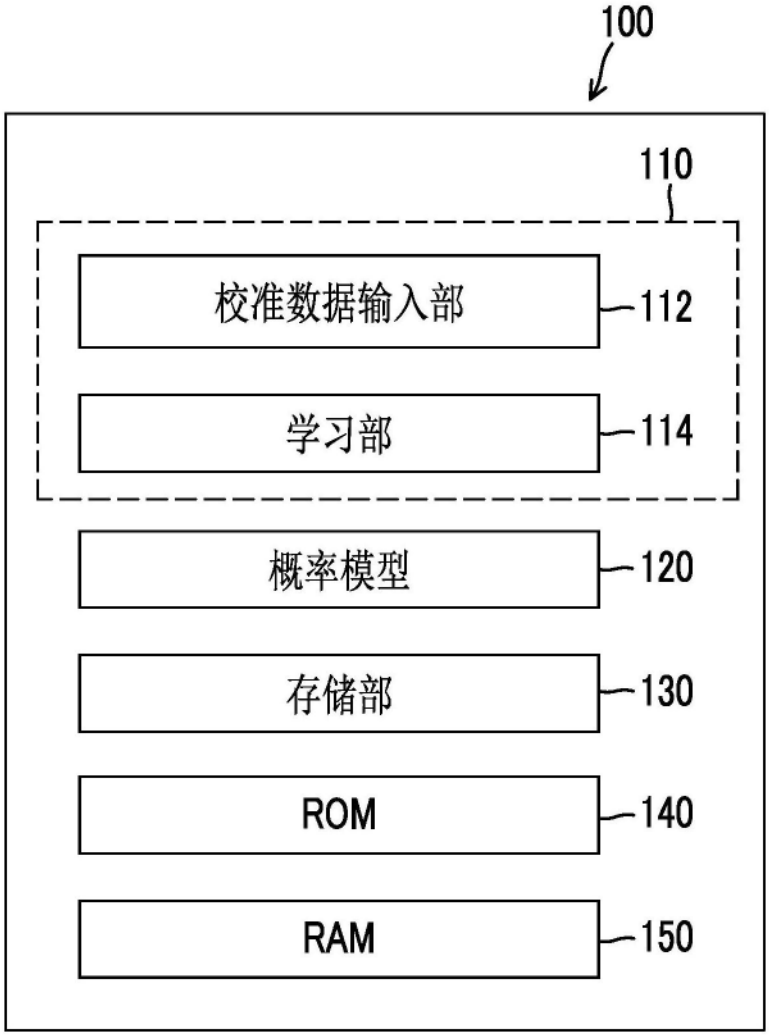


图4

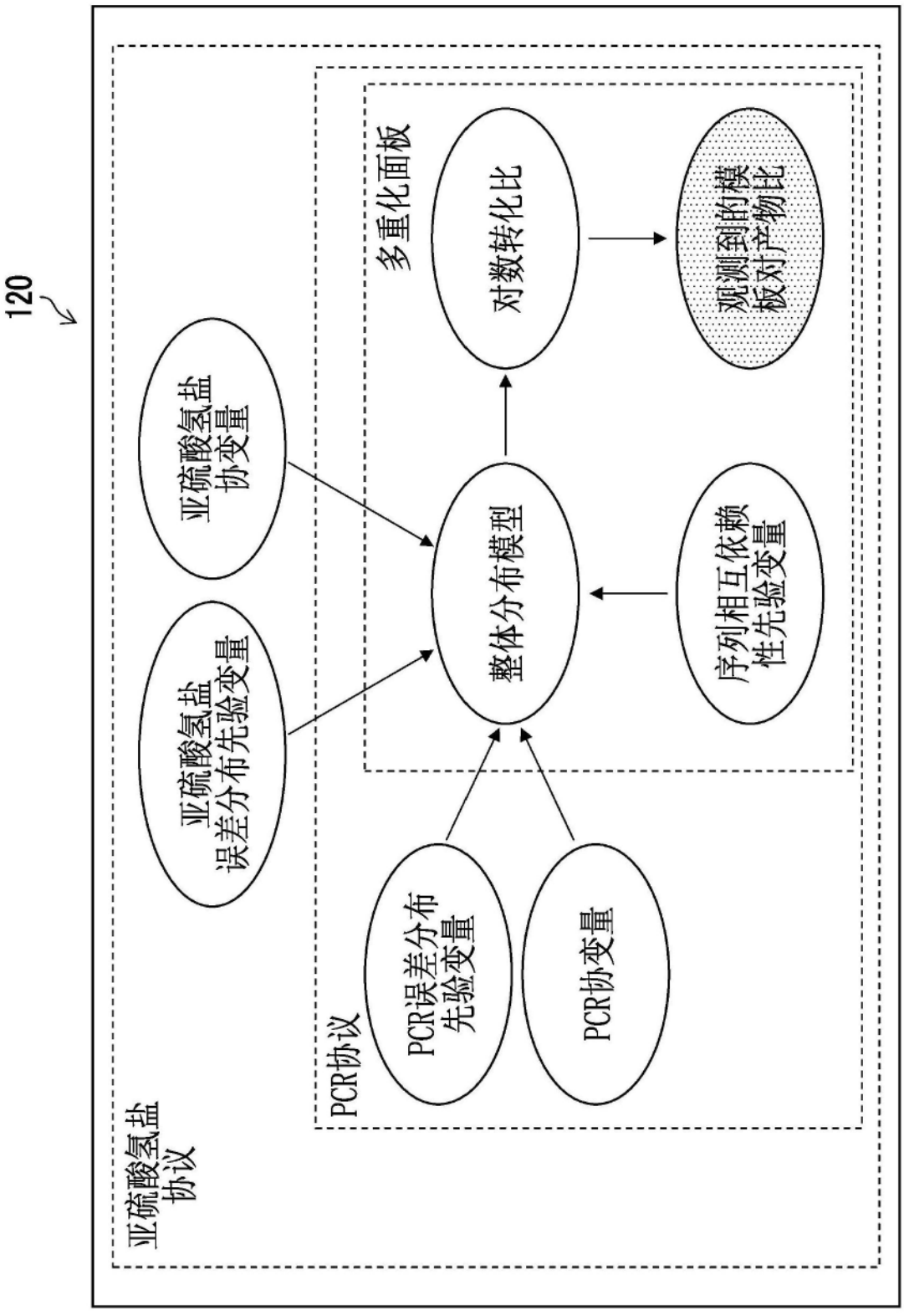

120
亚硫酸氢盐
协议
PCR协议
多重化面板
亚硫酸氢盐
协变量
亚硫酸氢盐
误差分布先验变量
PCR误差分布
先验变量
PCR协变量
整体分布模型
序列相互依赖
性先验变量
对数转化比
观测到的模
板对产物比

图5

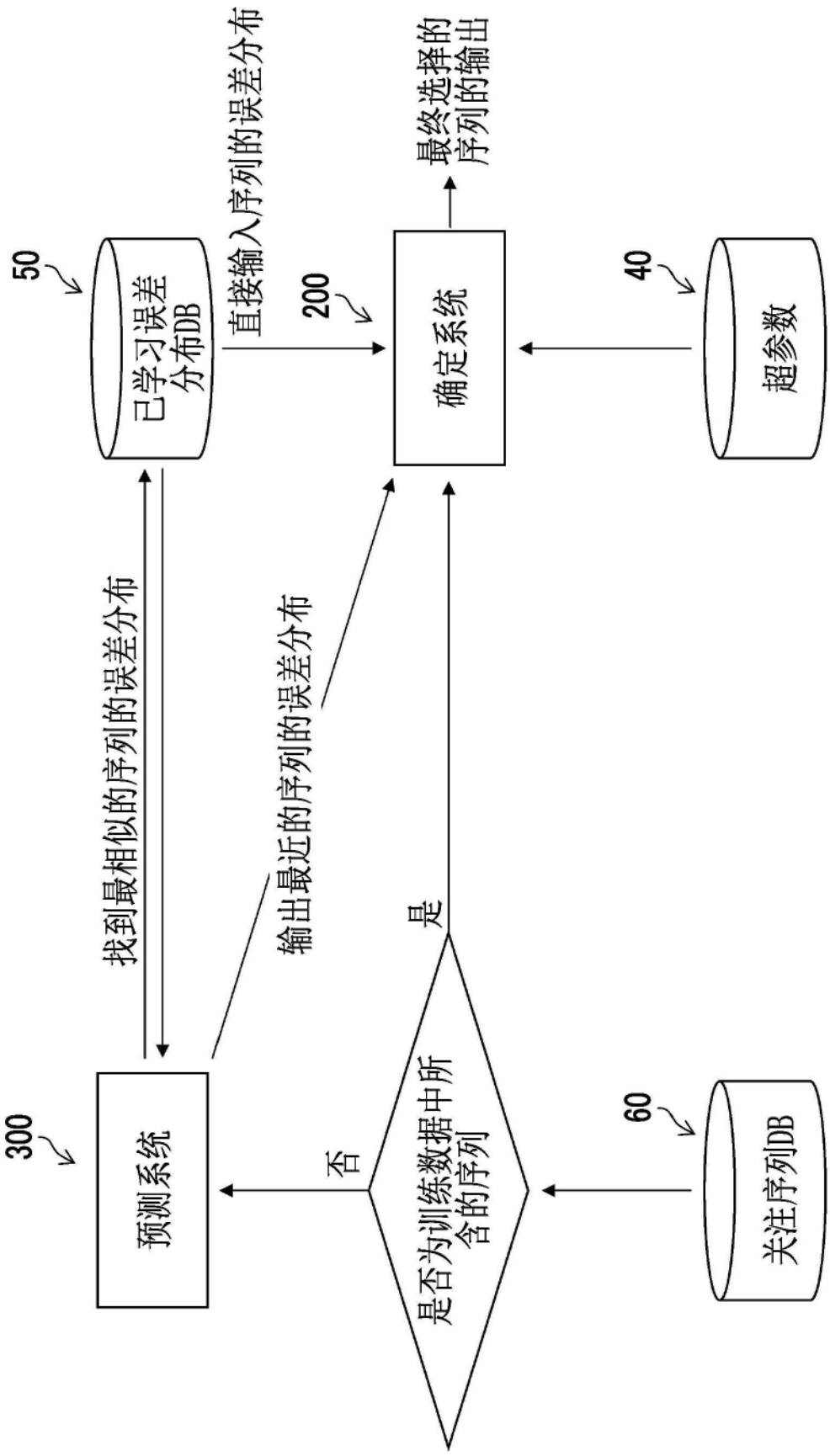

50
已学习误差分布DB
直接输入序列的误差分布
200
确定系统
最终选择的序列的输出
40
超参数
找到最相似的序列的误差分布
输出最近的序列的误差分布
是
300
预测系统
否
是否为训练数据中所含的序列
60
关注序列DB

图6

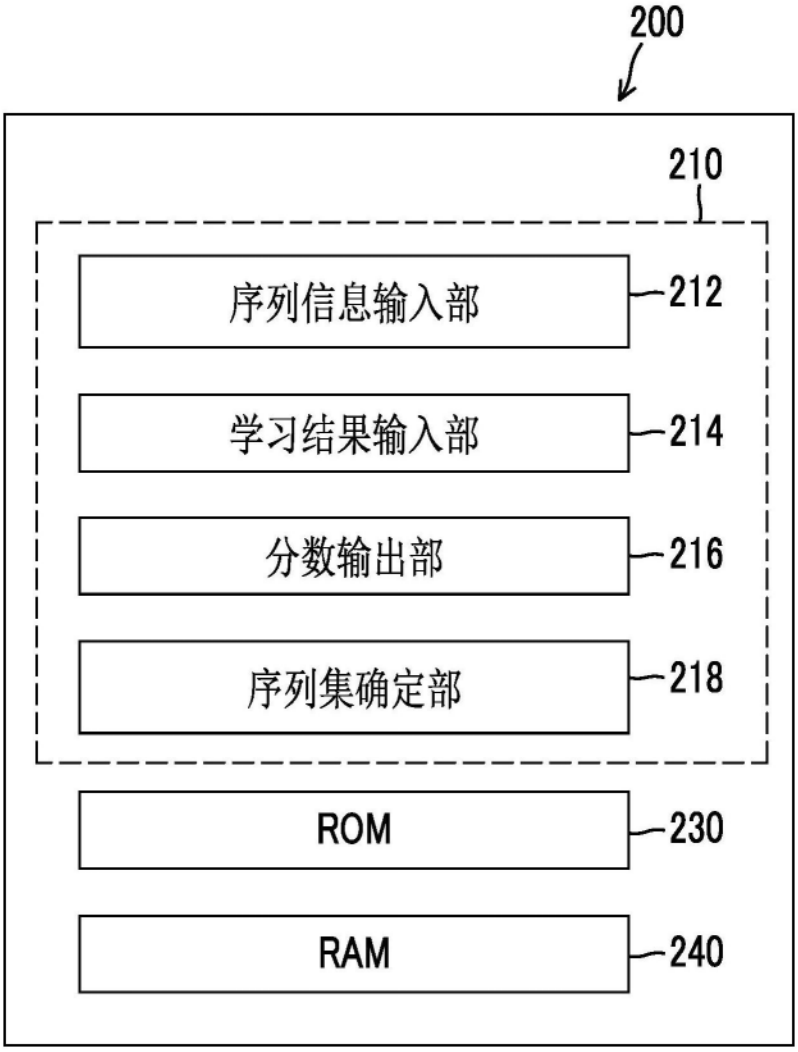

200
210
序列信息输入部
212
学习结果输入部
214
分数输出部
216
序列集确定部
218
ROM
230
RAM
240

图7

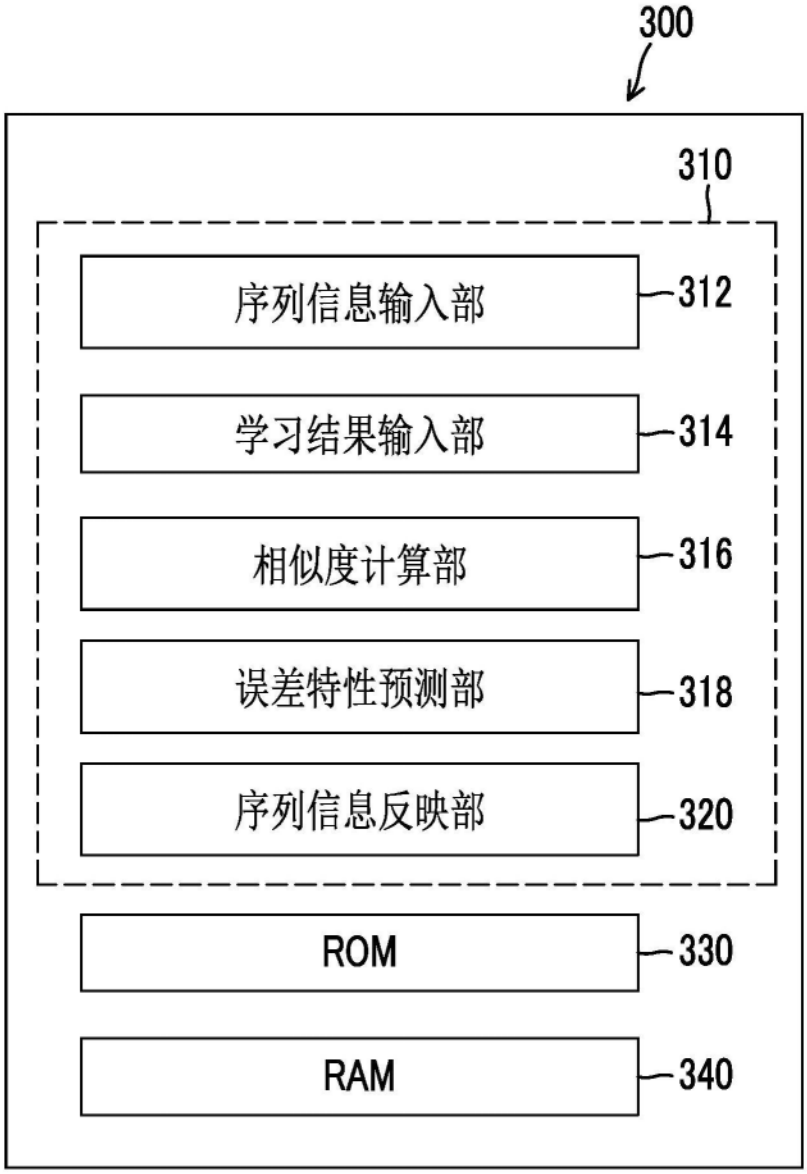

300
310
序列信息输入部
312
学习结果输入部
314
相似度计算部
316
误差特性预测部
318
序列信息反映部
320
ROM
330
RAM
340

图8