(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 102349075 B
(45)授权公告日 2014.12.17

(21)申请号 201080012005.6
(22)申请日 2010.03.16
(30)优先权数据
2009-063273 2009.03.16 JP
(85)PCT国际申请进入国家阶段日
2011.09.14
(86)PCT国际申请的申请数据
PCT/JP2010/001867 2010.03.16
(87)PCT国际申请的公布数据
WO2010/106794 JA 2010.09.23
(73)专利权人 学校法人明治大学
地址 日本东京
专利权人 公立大学法人滋贺县立大学
(72)发明人 矢野健太郎 清水顕史
(74)专利代理机构 北京德恒律师事务所 11306
代理人 陆鑫 高雪琴
(51)Int.Cl.
G06F 19/20(2011.01)
(56)对比文件
JP 2006277611 A,2006.10.12,
CN 101103272 A,2008.01.09,
JP 2007048027 A,2007.02.22,
审查员 吴瑶

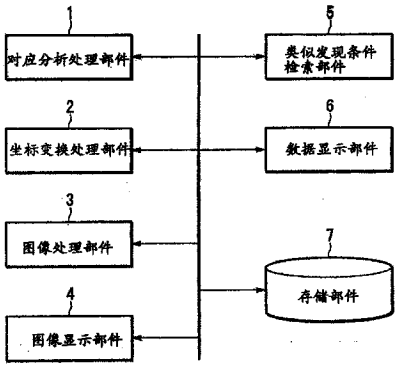
权利要求书2页 说明书13页 附图7页

(54)发明名称

发现曲线分析系统及其程序

(57)摘要

本发明提供一种发现曲线分析系统，即通过通常的计算机高速地分析从下一代高速时序产生器、类似的实验方法等得到的大量的发现曲线数据，将遗传基因的发现模式可视化，容易地对新遗传基因是否具有与任意的遗传基因接近的功能进行分析。本发明的发现曲线分析系统是对遗传基因的发现曲线数据进行分析的发现曲线分析系统，包括：存储部件，与评价遗传基因对应地，将遗传基因的多个发现条件的每一个的从评价对象的每个评价遗传基因中发现的mRNA的计数数目存储为发现数据；对应分析处理部件，针对每个评价遗传基因，从存储部件中读出发现数据，根据发现数据的每个发现条件的计数数目，进行对应分析；坐标变换处理部件，根据通过对应分析所得到的n(n是自然数)维的分数，将各评价遗传基因变换为配置为m(m是自然数，m≤n)维的坐标值；图像处理部件，描绘为与每个遗传基因对应的坐标值而显示在图像显示部件上。



CN 102349075 B

1. 一种发现曲线分析系统，是对遗传基因的发现曲线数据进行分析的发现曲线分析系统，其特征在于，包括：

存储部件，与评价遗传基因对应地，将遗传基因的多个发现条件的每一个的从评价对象的所述评价遗传基因中发现的 mRNA 的计数数目存储为发现数据；

对应分析处理部件，针对每个所述评价遗传基因，从所述存储部件中读出所述发现数据，根据发现数据的每个发现条件的计数数目，进行对应分析；

坐标变换处理部件，根据通过对应分析所得到的 n 维的分数，将各评价遗传基因变换为配置为 m 维的坐标值，其中，n 维的分数为与各发现条件的坐标值对应的各遗传基因的坐标，n 是自然数，m 是小于等于 n 的自然数；

图像处理部件，描绘为与每个所述遗传基因对应的坐标值而显示在图像显示部件上，

在对应分析的处理中包含功能已知的已知遗传基因，根据该已知遗传基因和与所述评价遗传基因的 n 维坐标的距离，而进行功能与所述已知遗传基因类似的评价遗传基因的抽出处理。

2. 根据权利要求 1 所述的发现曲线分析系统，其特征在于：

在对应分析的处理中作为虚拟遗传基因而包含只根据各发现参数发现的所述已知遗传基因，将该虚拟遗传基因的坐标作为表示通过所述 n 维所显示的图形中的只有任意一个发现参数的发现条件的顶点。

3. 根据权利要求 2 所述的发现曲线分析系统，其特征在于，还包括：

类似发现条件检索部件，求出配置在所述顶点的所述虚拟遗传基因的坐标与所述评价遗传基因的坐标的距离，针对所述顶点的坐标，抽出位于预定距离内的坐标的评价遗传基因。

4. 根据权利要求 1 所述的发现曲线分析系统，其特征在于，还包括：

数据显示部件，通过选择与所述评价遗传基因、所述已知遗传基因对应的坐标，而从所述存储部件中读出被配置在该选择出的遗传基因的图像的坐标位置上的遗传基因相关的信息并显示。

5. 根据权利要求 2 或 3 所述的发现曲线分析系统，其特征在于：

所述坐标变换处理部件在由对应分析处理部件所求出的各维中，从行分数的贡献率高的维开始对该贡献率进行累计，将累计结果的累计贡献率与预先设置的阈值进行比较，由此通过 1 维、2 维和 3 维的任意一个来显示由所述顶点构成的图形。

6. 一种发现曲线分析方法，是对遗传基因的发现曲线数据进行分析的发现曲线分析方法，其特征在于，通过计算机执行以下的处理：

对应分析处理步骤，分析处理部件针对每个评价遗传基因从存储部件中读出发现数据，根据发现数据的每个发现条件的计数数目，进行对应分析，其中，存储部件，与所述评价遗传基因对应地，将遗传基因的多个发现条件的每一个的从评价对象的所述评价遗传基因中发现的 mRNA 的计数数目存储为发现数据；

坐标变换处理步骤，坐标变换处理部件根据通过对应分析所得到的 n 维的分数，将各评价遗传基因变换为配置为 m 维的坐标值，其中，n 维的分数为与各发现条件的坐标值对应的各遗传基因的坐标，n 是自然数，m 是小于等于 n 的自然数；

图像处理步骤，图像处理部件描绘为与每个所述遗传基因对应的坐标值而显示在图像

显示部件上；

在对应分析的处理中包含功能已知的已知遗传基因，根据该已知遗传基因和与所述评价遗传基因的 n 维坐标的距离，而进行功能与所述已知遗传基因类似的评价遗传基因的抽出处理。

发现曲线分析系统及其程序

技术领域

[0001] 本发明涉及对遗传基因的发现曲线进行分析等的发现曲线分析系统及其程序。本申请根据在 2009 年 3 月 16 日在日本申请的特愿 2009-063273 号而主张优先权，在此引用其内容。

背景技术

[0002] 随着染色体组分析研究的进展，功能未知的新遗传基因被大量确认，必须弄清其功能，为了得到揭示其功能的信息，使用了与发现条件（表示出发现遗传基因的条件的信息）对应的遗传基因的发现模式。

[0003] 为此，通过 EST、MPSS、SAGE、CAGE 等，进行了以下这样的处理，即网罗地对从疾病患者、病理模型动物的组织或培养细胞等中取得的大量（数万水平）的遗传基因的发现进行分析。

[0004] 即，在根据信使 RNA（以下称为 mRNA）的计数数目进行的遗传基因分析中，通过根据遗传基因的发现模式的特征，利用遗传基因发现曲线分析，来进行对象的全部遗传基因的分类。

[0005] 一般，通过使用由 n 个遗传基因构成的 mRNA，利用从 k 个独立的实验条件所得到的 mRNA 的发现频度的数据，从而 n 个遗传基因分别成为 k 维特征空间中的具有 k 维特征向量的坐标点。

[0006] 因此，n 个遗传基因分别根据各自的特征向量，而成为上述特征空间中的 n 个坐标点的集合。

[0007] 上述发现曲线分析是指：对在上述特征空间中标绘的坐标点，即在特征空间上成为类似的遗传基因之间，进行分组而分类。

[0008] 通过上述分组的处理，例如得到在处于正常状态的健康人中所发现的遗传基因在任意疾病的患者中没有发现、或者发现量增加或者减少了等对于疾病的患者是特异的发现曲线，从而能够检测出健康人所没有而与疾病相关的特有的遗传基因。

[0009] 这样，遗传基因发现曲线成为为了预测功能未知的遗传基因的功能而使用的重要的工具。

[0010] 在遗传基因发现曲线分析中，使用了对遗传基因发现比的指标进行了矩阵化的数据作为分析对象的数据。

[0011] 例如，将在各行中排列所评价的遗传基因群，在各列中分别排列采样群（成为标的的表现型），该行和列就是遗传基因发现曲线。另外，采样更具体地是指通过不同的多个调查个体或同一个体中的 Time Course 实验而检测到的表现型等。例如，在用 50 个个体检测 100 种遗传基因的发现量时，矩阵 A 的要素 Aij(i 行 j 列的值、$1 \leqslant i \leqslant 100$、$1 \leqslant j \leqslant 50$) 表示与第 i 个遗传基因有关的第 j 个个体所显示出的发现量。

[0012] 在从遗传基因发现曲线分析中的巨大量的采样所得到的结果的分析中，需要一种用于更有效地分析其结果而迅速地发现目标遗传基因的信息处理技术。以前，作为这样的

技术，例如进行了分类（clustering）分析、主成分分析等特别多变量的分析、系统的分析（例如参考非专利文献 1、非专利文献 2）。

[0013] 另外，对异常基因发现量（发现比）进行对数变换来进行遗传基因发现曲线分析。具体地说，对数变换将发现水平的比（发现比、ratio）作为对数变换的指标（例如 log2(ratio) 等），主要在通过微配列（microarray）实验，在采样之间比某遗传基因的发现水平的情况下被使用。作为进行该对数变换的理由，例如如果是 log2(ratio) 变换，可以列举出能够将 -2、-1、0、1、2、1 倍为中心将 1/4 倍、1/2 倍、1 倍（等发现）、2 倍、4 倍这样的发现比变换为等尺度，研究者容易理解、在进行统计分析的基础上是妥当的等。但是，有以下这样的跨学科的问题，即由研究机构、研究者对该对数的底使用 2、e、10 等是没有统一性的，无法对在 Web 上等公开的数据之间进行直接比较。

[0014] 另外，在分类分析中，可以根据多维的特征向量将具有类似的遗传基因发现曲线的遗传基因群或采样群分割为同一类（cluster）。为此，在分类分析中，通过被广泛利用的阶段性分类（Ewing 等，1999，Genome Res.9 :950-959 的研究等），由于计算量的增加而难以通过通用的计算机等进行分析。另外，一般根据现在的巨大量的 EST 数据来预测数千到数万个发现遗传基因。作为对遗传基因模式的分类分析结果的代表性的表现方法的树状图是用于从视觉捕捉遗传基因之间的发现模式的类似性的有用的表现方法（后述的图 8，“van`t Veer，L.J.，Dai，H.，van de Vijver，M.J.，He，Y.D.，Hart，A.A.，Mao，M.，Peterse，H.L.，van der Kooy，K.，Marton，M.J.，Witteveen，A.T.，et al，(2002)Gene Expression profiling predicts clinical outcome of breast cancer，Nature，415，530-536”中的图 1），但遗传基因个数为数千个以上的情况下，难以将树状图全体输出到计算机监视器或印刷纸上，为了根据大规模的树状图解释结果就需要大量的劳力。

[0015] 即，阶段性分类也具有以下这样的缺点等，即伴随着遗传基因个数的增加而计算量增加，另外依存于所给出的数据组而树状图的拓扑逻辑容易发生变化，随着矩阵大小的增加而分析时间急剧加长，需要计算机的 CPU 和存储器。

[0016] 另外，在 Kmeans 法、SOM(Self Organizing Maps) 法中，与阶段性分类相比，能够通过少的计算机资源进行分析。但是，在进行分析时，需要预先决定类别（cluster）数，是主观的方法。

[0017] 另外，在作为多变量分析的一个的主成分分析方法中，虽然能够高速地执行计算，但由于并不是对曲线的分析方法，所以无法根据所得到的分数对发现曲线进行比较。

[0018] 另外，还有以下这样的问题点，即难以视觉地掌握通过上述各方法得到的巨大量（万的级别）采样、遗传基因的类别。因此，现在主要进行以下的操作，即根据 Pearson 的相关系数等从大规模类别中只取出成为目标的类别。

[0019] 但是，对于研究者来说，所得到的类别的观察器并不容易理解（参考图 8）。

[0020] 上述图 8 所示的被称为 two-dimensional-display 的观察器是纵横（或相反）地将各遗传基因和各采样排列起来。另外，进行了视觉化使得各单元（cell）的颜色、该颜色的浓淡表示所对应的采样～遗传基因发现的强弱。

[0021] 另外，主成分分析是直接对遗传基因发现曲线的数值的大小进行比较的统计方法，能够进行更高速的分析。

[0022] 但是，在主成分分析中，进行高速分析的结果是输出了与调查对象的表现型无关

的家政(housekeeping)遗传基因相对于各主轴输出了不同的分数(score)(坐标那样的数据),因此在绘制为散布图的情况下,也难以进行检测。

[0023] 非专利文献1:了解!使用!DNA微配列数据分析入门,羊土社,Steen Knudsen(著),盐岛聪(翻译),迁本豪三(翻译),松本治(翻译)

[0024] 非专利文献2:数据必出DNA微配列实战手册-基本原理,从芯片制造技术到生物信息学,羊土社,岡琦康司(编集),林琦良英

[0025] 如上所述,在分析法中存在各种问题,但分析时间(处理时间)长,另外对微量的遗传基因发现比检测力低(与量化性状相关的遗传基因的检测力低)的问题很大。

[0026] 具体地说,在遗传基因发现曲线分析中,对超过103的巨大量的数据进行处理来进行分析。

[0027] 但是,难以使用通常的计算机迅速地计算这样巨大量的数据。其结果是分析时间长。

[0028] 另外,在以前主要使用的阶段性分类方法中,为了缩短计算时间/简化,主观地关注采样之间的发现比是数倍以上或数倍以下的遗传基因群。这基于以下这样的期望,即越是发现量2～3倍地有很多变化的遗传基因越是明显地对采样之间的表现型的差异产生影响。

[0029] 但是,在该阶段性分类方法中,即使发现比人为地不同,但差异小的遗传基因也从分析对象中被排除了。

[0030] 其结果是例如极其难以检测出与量化性状相关的遗传基因。即,在该方法中,在要检测的表现型不是定性的而是定量的情况下,无法检测出与该表现型相关的遗传基因中的遗传基因发现量的比极少变化的遗传基因。即,在现有的方法中,无法全部检测出与目标的表现型有关的遗传基因。

[0031] 如上所述,在现有分析的状况下,由于不存在网罗地发现发现比极少变化的遗传基因这样的视点,所以在现有的分析方法(对数变换)中,不存在对微量的遗传基因发现比的检测力低这样的问题本身。

[0032] 另外,在以前检测不出来的与量化性状相关的遗传基因中,包含重要的新遗传基因的可能性高。因此,开发出有效地发现与量化性状相关的遗传基因的新大规模分析工具是必要不可缺的。

发明内容

[0033] 因此,本发明鉴于上述现有的问题,其目的在于:提供一种发现曲线分析系统及其程序,即使在使用了通常的计算机的情况下也能够迅速地分析巨大量的发现曲线数据,同时与现有技术相比,通过对遗传基因的发现模式进行可视化,能够容易地分析出新遗传基因是否具有与任意库的遗传基因接近的功能。

[0034] 本发明的发现曲线分析系统是对遗传基因的发现曲线数据进行分析的发现曲线分析系统,其特征在于包括:

[0035] 存储部件,与评价遗传基因对应地,将遗传基因的多个发现条件的每一个的从评价对象的上述评价遗传基因中发现的mRNA的计数数目存储为发现数据;

[0036] 对应分析处理部件,针对每个上述评价遗传基因,从上述存储部件中读出上述发

现数据，根据发现数据的每个发现条件的计数数目，进行对应分析；

[0037] 坐标变换处理部件，根据通过对应分析所得到的 n(n 是自然数）维的分数，将各评价遗传基因变换为配置为 m(m 是自然数，m ≤ n) 维的坐标值；

[0038] 图像处理部件，描绘为与每个上述遗传基因对应的坐标值而显示在图像显示部件上。

[0039] 理想的是在本发明的发现曲线分析系统中，在对应分析的处理中包含功能已知的已知遗传基因，根据该已知遗传基因和与上述评价遗传基因的上述 n 维坐标的距离，而进行功能与上述已知遗传基因类似的评价遗传基因的抽出处理。

[0040] 理想的是在本发明的发现曲线分析系统中，在对应分析的处理中作为虚拟遗传基因而包含只根据各发现参数发现的上述已知遗传基因，将该虚拟遗传基因的坐标作为表示通过上述 n 维所显示的图形中的只有任意一个发现参数的发现条件的顶点。

[0041] 理想的是在本发明的发现曲线分析系统中，还包括：类似发现条件检索部件，求出配置在上述顶点的上述虚拟遗传基因的坐标与上述评价遗传基因的坐标的距离，针对上述顶点的坐标，抽出位于预定距离内的坐标的评价遗传基因。

[0042] 理想的是在本发明的发现曲线分析系统中，还包括：数据显示部件，通过选择与上述评价遗传基因、上述已知遗传基因对应的坐标，而从上述存储部件中读出被配置在该选择出的遗传基因的图像的坐标位置上的遗传基因相关的信息并显示。

[0043] 理想的是在本发明的发现曲线分析系统中，上述坐标变换处理部件在由对应分析处理部件所求出的各维中，从行分数的贡献率高的维开始对该贡献率进行累计，将累计结果的累计贡献率与预先设置的阈值进行比较，由此通过一维、2 维和 3 维的任意一个来显示由上述顶点构成的图形。

[0044] 理想的是本发明的发现曲线分析程序是对遗传基因的发现曲线数据进行分析的发现曲线分析系统，其特征在于通过计算机执行以下的处理：

[0045] 对应分析处理，分析处理部件针对每个上述评价遗传基因从存储部件中读出发现数据，根据发现数据的每个发现条件的计数数目，进行对应分析，其中，存储部件，与上述评价遗传基因对应地，将遗传基因的多个发现条件的每一个的从评价对象的上述评价遗传基因中发现的 mRNA 的计数数目存储为发现数据；

[0046] 坐标变换处理，坐标变换处理部件根据通过对应分析所得到的 n(n 是自然数）维的分数，将各评价遗传基因变换为配置为 m(m 是自然数，m ≤ n) 维的坐标值；

[0047] 图像处理，图像处理部件描绘为与每个上述遗传基因对应的坐标值而显示在图像显示部件上。

[0048] 如以上说明的那样，根据本发明，通过根据评价对象的评价遗传基因的每个发现条件的 mRNA 数的计数值进行对应分析，通过与各个发现模式对应的坐标值来将各评价遗传基因配置在空间（分析空间）中，通过可显示的维显示在图像显示部件上，因此，能够得到以下这样的效果：用户能够上述图像显示部件的显示画面上容易地抽出由评价遗传基因的每个发现条件的计数数目构成的发现模式的发现曲线接近的形状（一致或类似）的，即功能类似的遗传基因。

[0049] 另外，根据本发明，能够得到以下这样的效果：通过将只按照任意一个发现条件而发现的特异遗传基因的发现模式包含在由分析对象（评价对象）的评价遗传基因构成的评

价遗传基因群中，由此由于各特异遗传基因成为表示各发现条件的标志，所以用户能够容易地在上述图像显示部件的显示画面上确认是否能够明确地发现各分析对象的评价遗传基因将哪个发现条件作为主要因素。

[0050] 另外，根据本发明，通过由用户输入上述空间的任意距离，选择特异遗传基因，类似发现条件检索部件抽出包含在以该特异遗传基因为中心以上述距离为半径的球内的评价遗传基因，因此能够容易地抽出具有由用户设定的距离所对应的类似性的评价遗传基因。

[0051] 另外，根据本发明，能够得到以下这样的效果：由于通过将功能已知的已知遗传基因包含在由评价遗传基因构成的评价遗传基因群中，各已知遗传基因成为表示遗传基因的功能的发现条件的标志，所以用户能够容易地在上述图像显示部件的显示画面上确认各评价遗传基因是否具有与已知遗传基因的功能接近的功能。

[0052] 另外，根据本发明，能够得到以下这样的效果：由于通过选择显示在上述图像显示部件的显示画面上的各遗传基因的显示图像，而将各遗传基因的遗传基因配列 / 测定条件等与遗传基因有关的信息显示在上述图像显示部件的显示画面上，所以能够在显示得很多的信息中容易地确认关注的遗传基因的固有信息。

[0053] 另外，根据本发明，能够得到以下这样的效果：由于根据通过对应分析的结果所得到的多维的累计贡献率，来设定通过 1 维、2 维还是 3 维进行图像显示，所以在图像显示部件的显示画面上识别类似性变得容易（在此，在 2 维的情况下，发现条件在 2 维平面上，被描绘为对于 2 个条件（2 个主轴）将特异地发现的描绘位置的顶点之间连接起来的直线，或者以该描绘位置为顶点而形成的多角形。在该情况下，描绘位置为 2 维坐标）。

附图说明

[0054] 图 1 是表示本发明的一个实施例的发现曲线分析系统的结构例子的框图。

[0055] 图 2 是表示存储在图 1 的存储部件 7 中的发现数据表的结构例子的概念图。

[0056] 图 3 是表示存储在图 1 的存储部件 7 中的分数表的结构例子的概念图。

[0057] 图 4 是表示存储在图 1 的存储部件 7 中的坐标表的结构例子的概念图。。

[0058] 图 5 是表示在 3 维空间中显示出以与 5 个发现条件对应的特异遗传基因的显示图像为顶点的五面体，用线将该五面体的各顶点连接起来，并且在顶点的近旁显示出表示发现条件的字符串的图像的概念图。

[0059] 图 6 是表示在 3 维空间中显示出以与 5 个发现条件对应的特异遗传基因的显示图像为顶点的五面体，用线将该五面体的各顶点连接起来，并且在顶点的近旁显示出表示发现条件的字符串的图像的概念图。

[0060] 图 7 是表示在 3 维空间中显示出以与 5 个发现条件对应的特异遗传基因的显示图像为顶点的六面体，用线将该六面体的各顶点连接起来，并且在顶点的近旁显示出表示发现条件的字符串的图像的概念图。

[0061] 图 8 是表示现有的分析系统中的遗传基因的发现曲线的分析结果的显示工具的显示画面的概念图。

[0062] 附图标号

[0063] 1：对应分析处理部件；2：坐标变换处理部件；3：图像处理部件；4：图像显示部

件 ;5 :类似发现条件检索部件 ;6 :数据显示部件 ;7 :存储部件

具体实施方式

[0064] 以下,参考附图,说明本发明的一个实施例的发现曲线分析系统。本实施例的发现曲线分析系统基于根据从遗传基因的发现曲线数据所得到的每个发现条件的计数值所进行的对应分析(例如,大隅升,L.Lebart,其他著,记述的多变量分析法,1994,日科连出版社记载),推测、确定、预测与预先设定的表现型相关的遗传基因。

[0065] 另外,上述“发现曲线数据”是指各个实验材料,例如在细胞中发现的多个遗传基因的 mRMNA 的发现模式,换一种说法,就是表示由遗传基因的种类、其各自的发现量(或每个发现条件的计数值)构成的数据的集合体。另外,以下,将各个发现曲线数据简单地说明为发现数据、遗传基因发现数据。另外,在作为每个发现条件的计数值的情况下,表示构成发现条件的各条件的计数值,在作为发现条件的发现模式的情况下,表示形成构成发现条件的每个条件的计数值的模式。

[0066] 另外,上述“表现型”是指与各遗传基因的性格相关联的任意的性质,包含定性的指标、定量的指标。例如,对于与疾病相关联的指标可以列举疾病的名称、原因、进展状况、预后、残年或病症、再发、转移的可能性等,但并不特别限定于此。

[0067] 另外,本实施例的发现曲线系统是以下这样的系统,即能够高效地迅速地处理通过 EST(Expressed Sequence Tag)、MPSS(Massively Parallel Signature Sequencing)、SAGE(Serial Analysis of Gene Expression) 以及 CAGE(Cap Analysis Gene Expression) 等所得到的巨大量的遗传基因的各发现条件的 mRNA 的发现数,即发现曲线数据。在本实施例中,上述发现条件是指对遗传基因的由来(任意的动物、该动物的任意生物体部分等)、发现时的环境等的发现量进行比较的参数。

[0068] 即,通过发现曲线实验、特别是根据使用大量的发现数据所得到的每个发现条件的计数值所进行的对应分析,能够对与任意的发现型相关的遗传基因进行分析,推测出与该发现型相关的遗传基因。

[0069] 特别地,从遗传基因发现的从 mRNA 利用逆转录酶逆转录反应而合成的从 cDNA 克隆而得到的 cDNA 配列、发现遗传基因断片 EST、另外从下一代高速序列产生器得到的发现遗传基因的配列除了转录产物的数组信息以外,还能够得到遗传基因所发现的生育阶段、器官、组织等的信息。即,这意味着通过针对 1 个以上的生物物种,进行 EST 的配列和由来(生育阶段、器官)的信息收集和调查,能够进行生物物种固有的发现遗传基因的探索,乃至生殖和应力响应、植物的光合作用、从根的养分水分吸收等与各种生物学过程相关联的遗传基因的探索。近年来,通过许多研究者,动植物、微生物的 EST 分析正在发展,登记在国际碱基配列数据库中的 EST 条目数从 2000 年 10 月当时的约 623 万件以指数函数增长到 2008 年 11 月当时的约 5834 万件。

[0070] 另外,近年来,首先广泛使用了利用下一代高速时序产生器的大规模发现分析。这些 EST、从下一代高速时序产生器所得到的信息的积蓄使得遗传基因的发现模式的详细分析和有用的遗传基因的推测成为可能。另一方面,为了从这样的大规模数据中引出有用信息,不开发出通过多数研究者所利用的通用计算机能够处理的统计分析方法和工具,就无法灵活利用所积蓄的基础信息。

[0071] 以下，说明本实施例的发现曲线分析系统。图 1 是表示该实施例的发现曲线分析系统的结构例子的框图。

[0072] 在该图中，发现曲线分析系统具备对应分析处理部件 1、坐标变换处理部件 2、图像处理部件 3、图像显示部件 4、类似发现条件检索部件 5、数据显示部件 6 和存储部件 7。在本实施例中，将各遗传基因的每个发现条件（库）的 mRNA 的发现数目的计数值作为发现数据。因此，被用为该发现条件的每个发现条件的 mRNA 的计数值可以是通过上述 EST、MPSS、SAGE 和 CAGE 的任意一个得到的数值。

[0073] 如图 2 所示，在存储部件 7 中存储有发现数据表，它与所分析的遗传基因名对应地在该遗传基因中表示出多个发现条件的每一个，例如发现条件 A、发现条件 B、发现条件 C、发现条件 D、发现条件 E 的每个的发现了的 mRNA 的计数数目。

[0074] 对应分析处理部件 1 从存储部件 7 中顺序地读入作为各遗传基因的发现数据的每个发现条件的上述 mRNA 的计数值，根据由每个读入的发现条件的发现数据即计数值构成的发现模式，进行对应分析。

[0075] 简单地说明对应分析处理部件 1 中的对应分析。该对应分析与主成分分析一样，是决定用于说明 n 维数据的主轴的分析方法。

[0076] 在本实施例中，对应分析处理部件 1 使用从存储部件 7 的数据表读入的遗传基因的发现数据，求出可以说明表现型（性状等）的不同的 1 个或多个主轴。

[0077] 即，对应分析与简单地缩减进行比较的维数的主成分分析不同，不是将各个数据的量或大小作为分析对象，而是将数据矩阵的曲线（发现条件的发现量，即计数值的模式）作为分析对象，使得不损失发现模式，即作为多维数据的发现数据的本质性信息量（遗传基因的每个发现条件的计数数目的集合即发现模式）。

[0078] 由此，具有类似的活动的遗传基因并不是只根据任意一个发现条件的发现量而检测出的，而是与各发现条件所对应的 mRNA 的计数值的曲线近似这样的具有类似的功能的遗传基因。因此，对应分析对于从每个该发现条件的计数值的曲线即发现曲线抽出具有类似的活动的遗传基因群这样的目的来说，是有用的。

[0079] 其结果是发现模式具有相同的发现曲线的遗传基因被配置（描绘）在空间的同一坐标上（表示发现条件的计数值的分布相同或者类似的程度），能够容易地从巨大量的发现数据中抽出发现的曲线近似的遗传基因或遗传基因群。

[0080] 对于上述的对应分析中的分布的等同性（同样还是类似），可以将后述成为发现模式的分类指标的虚拟遗传基因（例如为了进行功能分类而明确地具有该功能并且具有成为分类的基准的发现模式的已知遗传基因）附加到上述发现数据表中（将后面说明通过附加该虚拟遗传基因而使被描绘曲线的遗传基因群（或遗传基因）的分析的位置具有意义）。

[0081] 依照对应分析的计算方法，对应分析处理部件 1 为了求出各遗传基因的发现数据的发现模式，而进行相对频度的计算。在此，如果设与 q 个遗传基因相关的 p 种发现条件的发现数据 q×p 矩阵的 i 行 j 列的要素为 kij，则作为向相对频度的变换，对应分析处理部件 1 根据以下所示的式 (1) 的第 i 行的列和 ki. 与式 (2) 的第 j 行的行和 k.j 的相乘结果，对各要素 kij 进行除法运算。在此，p 和 q 是 2 以上的自然数。由此，能够与全部行和列相等地对每个发现条件的计数值附加加权，根据不是强度而是由发现曲线的每个发现条件的计

数值的直方图形成的模式形状，能够抽出功能类似的遗传基因。

[0082] 式 (1)

[0083] $$k_{i\bullet}=\sum_{j=1}^{p}k_{ij} \quad \cdots(1)$$

[0084] 式 (2)

[0085] $$k_{\bullet j}=\sum_{i=1}^{q}k_{ij} \quad \cdots(2)$$

[0086] 另外，对应分析处理部件 1 根据由通过相对频度的计算所得的要素构成的相对频度数据矩阵 C 求出倒置矩阵 CT，生成 CT×C 的矩阵，计算出该矩阵的固有值和固有向量，求出说明发现数据的不同的多个主轴。

[0087] 另外，对应分析处理部件 1 使用与 p 种发现条件对应的 n 维（其中 n ≤ p）上的最大 p 角形的空间（以下称为分析空间，另外如果是 1 维，则是直线上）的 n 个固有值，计算出遗传基因配置用的行分数、发现条件（库）配置用的列分数。

[0088] 这时，通过将记载在图 2 的发现数据表中的虚拟遗传基因加入到分析对象的遗传基因群中，对应分析处理部件 1 根据上述虚拟遗传基因计算出作为上述多面体的顶点的发现条件配置的坐标，而作为上述的计算处理的结果。在此，将发现条件配置的坐标作为对于该发现条件特异的虚拟遗传基因的配置位置即行分数。即，将对于各个发现条件特异地发现的虚拟遗传基因配置在各发现条件的配置位置上。其结果是根据是否与被设定为该发现条件的分类指标的虚拟遗传基因的发现模式类似，来推测分析对象的遗传基因与哪个发现条件（或者哪个功能）的发现模式类似。即，在本实施例中，通过在对应分析的处理中包含功能已知的已知遗传基因（包含上述的虚拟遗传基因），而根据该已知遗传基因的坐标与分析对象的遗传基因的坐标的距离，来进行功能与已知遗传基因类似的评价遗传基因的抽出处理。

[0089] 即，作为对应分析的结果，在 n 维（与权利要求中的 n 维对应，n 是自然数）中，求出与各发现条件的坐标值对应的各遗传基因的坐标而作为分数。这时，对于更类似的发现条件，成为作为更短距离的坐标的分数，对于并不更类似的发现条件，成为作为具有更长距离的坐标的分数。如果只通过 1 个主轴来说明作为表现型的发现数据的不同，则它是在 1 维的线上，该主轴的贡献率是 100％。

[0090] 另外，在通过 1 个主轴无法说明表现型的变化，例如为了说明表现型的不同，需要 2 个主轴（第一主轴和第二主轴）的情况下，则成为通过第一主轴和第二主轴的 2 维平面进行说明的比例（贡献率），例如 70％和 30％等。

[0091] 即，上述“贡献率”对于表现型的变化来说，是指在由各主轴形成的平面上进行说明的比例。另外，将上述贡献率的和设为累计贡献率。这时，第一主轴的贡献率与第二主轴的贡献率席等或者在其以上。同样，随着成为第三、第四主轴，贡献率降低。在能够通过第一和第二主轴说明表现型的不同的情况下，表示分析结果的图可以通过 1 维或 2 维描绘进行绘图。另外，在为了说明表现型的不同，需要第三主轴的情况下，可以通过 3 维图 (3 维空间）的描绘来绘图表示分析结果的图。这样，在对应分析中，直到累计贡献率成为 100％为止，维数（即主轴的数目）增加。

[0092] 另外，根据赋予各主轴的固有值计算出上述贡献率。具体地说，各主轴的固有值相对于全部主轴的固有值之和的比成为该主轴的贡献率。例如，为了通过对应分析来说明表现型的变化，在得到了 10 维的主轴（第一主轴～第 10 主轴）时，向各主轴赋予固有值。另外，各主轴的固有值相对于与该各主轴对应的固有值的总和的比例成为贡献率，从第一主轴到第 10 主轴顺序地求出贡献率之和所得的数据成为累计贡献率。

[0093] 如上所述，在主轴是 1 个的情况下，通过 1 维的线中的坐标来表现表现型的不同，在主轴是 2 个的情况下，通过 2 维平面中的坐标来表现表现型的不同，在主轴是 3 个的情况下，通过 3 维空间中的坐标来表现表现型的不同 ……，在主轴是 p-1 个的情况下，通过 p-1 维的空间中的坐标来表现表现型的不同。

[0094] 但是，在对应分析的结果是计算出直到 3 维的情况下（主轴是 3 个），可以用维图、2 维图、3 维图来表示各遗传基因的坐标。另外，累计贡献率在 3 维的情况下是全体的 100％，能够通过 3 维的描绘图完全地说明表现型的不同。

[0095] 但是，在对应分析的结果是计算出 4 维以上的主轴的情况下，实际上是不可能进行 4 维以上的绘图的（在数学上是可能的，但在计算机处理中不能通过通常的绘图来进行）。

[0096] 因此，无法进行具有全维的视觉化。但是，例如，在计算出 4 维以上的主轴，到第三主轴为止的累计贡献率是 90％，以后的主轴的贡献率是 10％的情况下，通过 3 维图也能够说明全体的 90％。

[0097] 即，能够进行 90％的精度的判断。在该情况下，包含在剩余的 10％中的遗传基因无法用图来说明，因此在将 4 维图缩减到 3 维时，也可能出现若干个将 3 维空间的各顶点连接起来的分析空间以外的坐标上的遗传基因。即，有时也将求出为 n 维的主轴缩减到 m 维（m 是自然数，并且 $m \leqslant n$，在此，$n \leqslant p$）来表现求出为 n 维的主轴。例如，如上所述，由于图像对 4 维进行图像显示，所以将维数减少为由贡献率高的 3 个轴构成的 3 维来进行显示。

[0098] 例如，在如图 2 所示的发现数据表的发现条件那样，使用了发现条件 A、发现条件 B、发现条件 C、发现条件 D、和发现条件 E 的 5 个发现条件的情况下，通过包含对于各发现条件特异地发现的虚拟遗传基因，来通过 5 维以下的主轴说明各遗传基因的表现型。在该例子中，对应分析处理部件 1 求出与 4 维对应的分数 1、分数 2、分数 3、分数 4 的 4 个坐标数据的行分数来作为配置各遗传基因的坐标。

[0099] 在此，对应分析处理部件 1 针对存储部件 7，与各遗传基因对应地，将主轴（主轴 1、主轴 2、主轴 3 和主轴 4）的每个的分数写入到图 3 所示的分数表中。

[0100] 另外，对应分析处理部件 1 无法显示 4 维以上，因此，将显示的维数设为最大 3 维，按照贡献率高的主轴的顺序，将分数 1、分数 2 和分数 3 与各维的贡献率一起输出到坐标变换处理部件 2 中。

[0101] 坐标变换处理部件 2 如果与上述贡献率一起输入了对应分析结果的直到 3 维为止的分数 1、分数 2、分数 3，则将 1 维的贡献率、将 1 维和 2 维的贡献率相加了的累计贡献率、将 1 维、2 维和 3 维的贡献率相加了的了即贡献率分别与预先设定的设定贡献率进行比较，将超过该设定贡献率的维的组作为显示的空间的维。在此，1 维的贡献率最高，成为 2 维、3 维时，贡献率降低。

[0102] 例如，坐标变换处理部件 2 在只有 1 维的累计贡献率，即贡献率超过了上述设定贡

献率的情况下，按照 1 维的分数，在将描绘在 2 维平面上的 2 个顶点连接起来的直线上，配置遗传基因。在该直线上，通过是否与任意的顶点更接近，来表示是否强烈地因各个发现条件而发现。在该情况下，在发现条件的种类超过 2(3 以上）的情况下，多个发现条件的坐标重叠。

[0103] 另外，坐标变换处理部件 2 在 1 维和 2 维的贡献率相加了的了的累计贡献率超过了上述设定贡献率的情况下，通过 1 维和 2 维的分数，在 2 维空间上进行遗传基因的配置。在该情况下，在 2 维平面中，形成由作为各发现条件（特异遗传基因）的配置坐标的顶点构成的多角形的分析平面。在该多角形内的 2 维平面上，根据是否更接近多角形的任意顶点，来表示是否强烈地因各个发现条件而发现的。在此，将 1 维的分数用为 x 坐标的坐标值，将 2 维分数用作 y 坐标的坐标值。

[0104] 另外，坐标变换处理部件 2 在 1 维和 2 维的贡献率的相加结果即累计贡献率没有超过预先设定的设定贡献率的情况下，按照 1 维、2 维和 3 维的分数，在 3 维空间中进行遗传基因的配置。在该情况下，在 3 维空间中，形成由作为各发现条件（特异遗传基因）的配置坐标的顶点构成的多面体的分析空间。在该多面体内的 3 维空间内，根据是否更接近多面体的任意顶点，来表示是否强烈地因各个发现条件而发现的。在此，将 1 维的分数用作 x 坐标的坐标值，将 2 维分数用作 y 坐标的坐标值，将 3 维分数用作 z 坐标的坐标值。

[0105] 另外，坐标变换处理部件 2 例如在 3 维空间中显示各遗传基因的显示图像的情况下，如果输入了图 3 所示的各维的数据，即分数 1、分数 2 和分数 3，则对各分数的每个遗传基因，计算出 + 的分数和 - 的分数各自的绝对值，检测出任意一个的最大值，根据该最大值，对各遗传基因的分数进行除法计算，作为坐标值。由此，在 x 轴、y 轴和 z 轴中的坐标空间中，在实数的范围内，配置各遗传基因的显示图像和顶点。

[0106] 这样，由于通过用各遗传基因的分数除以各个维的分数的最大值来进行正规化，进行使各维的加权相同的处理，而显示在图像显示部件 4 的显示空间中，所以在发现条件的 n 个顶点，在此有 4 个顶点的情况下，将以任意的顶点的坐标作为原点的 4 个顶点的坐标与被该顶点围住的分析空间中的各遗传基因的坐标的相对坐标的坐标值变换为图像显示部件 4 的上述显示空间中的绝对坐标的坐标值。

[0107] 另外，坐标变换处理部件 2 针对存储部件 7，通过图 4 所示的表的结构，与各遗传基因对应地，写入配置遗传基因的坐标（例如坐标 1 :x 轴，坐标 2 :y 轴，坐标 3 :z 轴）的值。另外，坐标变换处理部件 2 与各发现条件对应地，将各发现条件被配置为顶点的坐标值写入到存储部件 7 中。

[0108] 图像处理部件 3 从存储部件 7 读入与各发现条件对应的顶点的坐标值，针对图像显示部件 4 的上述显示空间，如图 5 所示那样显示出将发现条件 A、发现条件 B、发现条件 C、发现条件 D、和发现条件 E 的 5 个顶点、将各顶点连接起来的线。这时，图像处理部件 3 在各顶点近旁，显示出表示各个顶点所对应的发现条件的字符串。例如，图像处理部件 3 在与发现条件 A 对应的顶点近旁，显示表示发现条件的“A”的字符串。该字符串与各发现条件对应地被存储在存储部件 7 中，在图像处理部件 3 描绘以图 5 的发现条件为顶点的图形时，与各发现条件对应地被从存储部件 7 读出，并显示在对应的发现条件的顶点近旁。

[0109] 另外，图像处理部件 3 从图 4 的坐标表中顺序地读入各遗传基因的坐标值，在以各发现条件为顶点的图形的多面体中，即在该例子中是在具有最大 5 个顶点的多面体内部的

分析空间中，如图 6 所示那样，将表示各遗传基因的显示图像（例如球状点、立方体状点、或者文字等）显示在与遗传基因对应的坐标值上。在图 5 和图 6 中，使用 3 维的主轴，设为在 3 维空间中显示的多面体，以五面体为例进行了显示。

[0110] 在该显示处理中，如已经说明的那样，将对于对应的发现条件特异地发现的遗传基因或虚拟遗传基因配置在与各发现条件对应的各个顶点上。

[0111] 另外，在存储在存储部件 7 的发现数据表中，与遗传基因名对应地，针对各虚拟遗传基因，存储有表示是虚拟遗传基因的虚拟遗传基因识别信息（在形成发现表时，用户针对各虚拟遗传基因，设定虚拟遗传基因识别信息）。

[0112] 另外，图像处理部件 3 在配置虚拟遗传基因时，用与其他所显示的遗传基因不同的颜色显示虚拟异常及的显示图像。

[0113] 另外，在将各顶点之间连接起来的线上，配置与通过该线连接起来的 2 个顶点的发现条件对应地发现的遗传基因。

[0114] 例如，在将与发现条件 A 和发现条件 C 对应的顶点连接起来的线上，配置在发现条件 A 和发现条件 C 下发现了的遗传基因。遗传基因在各线上的配置位置如下，相对于被配置在与 2 个发现条件对应的顶点上的虚拟遗传基因，将配置在线上的各遗传基因的发现模式，配置在与配置了发现模式类似的虚拟遗传基因的任意一个的顶点更接近的位置的坐标上。

[0115] 在图 7 中，将遗传基因只配置在将上述发现条件 A 和发现条件 C 连接起来的线上、将发现条件 A 和发现条件 D 连接起来的线上、将发现条件 C 和发现条件 D 连接起来的线上。

[0116] 即，在分析了的遗传基因组中，只存在只在发现条件 A 和发现条件 C 下发现的遗传基因、只在发现条件 A 和发现条件 D 下发现的遗传基因、只在发现条件 C 和发现条件 D 下发现的遗传基因。

[0117] 另外，在发现条件 A、发现条件 C 和发现条件 D 的 3 种发现条件下发现了的遗传基因被配置在由 5 个顶点形成的多面体中的与发现条件 A、发现条件 C 和发现条件 D 对应的顶点所形成的面（平面）上，在该平面上，被配置在与发现模式类似的顶点更接近的位置上。

[0118] 进而，在发现条件 A、发现条件 B、发现条件 C 和发现条件 D 的 4 种发现条件下发现了的遗传基因被配置在由 5 个顶点形成的多面体中的与发现条件 A、发现条件 B、发现条件 C 和发现条件 D 对应的顶点所形成的多面体的空间（3 维空间）内，在该 3 维空间上，被配置在与发现模式类似的顶点更接近的位置上。

[0119] 另外，图像处理部件 3 针对各线，将配置在将各顶点连接起来的线上的遗传基因的显示图像的颜色设为不同的颜色。

[0120] 同样，图像处理部件 3 将配置在将各顶点连接起来的面上的遗传基因的显示图像的颜色设为与配置在上述线和多面体的其他面上的显示图像不同的颜色。

[0121] 另外，图像处理部件 3 将配置在将各顶点连接起来的多面体的内部空间中的显示图像的颜色设为与配置在上述线、多面体的面、其他多面体的内部空间中的显示图像不同的颜色。

[0122] 另外，图像显示处理部件 3 与用户的设置对应地，将图像设为任意的方向，例如将 x 轴、y 轴、z 轴作为旋转轴，对显示为设定了的角度的图像进行旋转、左右旋转、上下翻转的处理。

[0123] 数据显示部件 6 在图像显示部件 4 的显示画面上，与用户通过点击鼠标等而选择了的遗传基因的坐标数据对应地，与坐标值对应地从存储在存储部件 7 中的坐标表中读出配置在该坐标上的遗传基因的遗传基因名。

[0124] 另外，数据显示部件 6 从发现数据表中读出与该遗传基因名对应的遗传基因的信息，并显示在该遗传基因名的遗传基因的坐标的近旁。

[0125] 另外，数据显示部件 6 在将多个遗传基因配置在同样的坐标上的情况下，在用户选择了该坐标的遗传基因的情况下，根据坐标从坐标表中读出配置在该坐标上的多个遗传基因名，并在选择的坐标的近旁将遗传基因名显示为列表。

[0126] 另外，在记载在上述列表中的遗传基因名中有希望参考信息的遗传基因，通过由用户选择该遗传基因，数据显示部件 6 与在上述列表中选择出的遗传基因的遗传基因名对应地，从存储部件 7 的发现数据表中读出与该遗传基因名对应的遗传基因的信息，将读出的遗传基因的信息显示在选择的坐标的近旁。

[0127] 类似发现条件检索部件 5 根据通过未图示的鼠标、键盘等由用户输入的距离的值、选择出的遗传基因（例如虚拟遗传基因）的坐标，使以输入的距离为半径的球内所包含的遗传基因的名与其他配置的遗传基因的颜色不同。

[0128] 其结果是用户能够容易地抽出与用户选择出的遗传基因类似的遗传基因，即表示成为目标的发现模式的遗传基因（或遗传基因群）。

[0129] 在此，可以利用 X 平方距离作为虚拟遗传基因、或上述线、上述多面体的面的统计上有意义的位置。通过利用该 X 平方距离，能够计算出有意义水平为 1％等的有意义的距离。

[0130] 例如，可以将在 1 个发现条件下特异地发现的遗传基因或遗传基因群定义为相对于各顶点位于 X 平方距离内的遗传基因（遗传基因群）。

[0131] 另外，在有兴趣的发现模式复杂的情况下，即在各发现条件下的计数值的直方图是具有复杂分布的形状的情况下，通过将具有对应的发现模式的已知遗传基因附加（添加）到发现数据表中，能够容易地检测出有兴趣的发现模式的已知遗传基因的坐标、以及与该已知遗传基因的距离比其他遗传基因更短的类似功能的遗传基因。

[0132] 另外，可以推测显示出与生物学功能是已知的已知遗传基因相类似的发现模式的遗传基因具有同样的功能、或者与遗传基因的发现控制相关联。

[0133] 因此，通过将根据发现模式（根据每个发现条件的计数值）预先检测出功能的已知遗传基因添加到存储在存储部件 7 中的发现数据表中，对应分析处理部件 1 根据每个上述发现条件的计数值的发现模式，与其他遗传基因同样地，针对已知遗传基因的坐标，计算出该已知遗传基因的行分数。

[0134] 另外，类似发现条件检索部件 5 将已知遗传基因的坐标作为基准，抽出位于预先设定的距离，例如以上述的 X 平方距离范围为半径的球状空间的范围内的遗传基因，即表示出与该已知遗传基因类似的发现模式（具有类似的功能）的遗传基因。

[0135] 由此，用户能够容易地检测出具有与已知遗传基因近似的功能的遗传基因。在大分类中，该已知遗传基因也包含已经说明了的将坐标用作顶点的虚拟遗传基因，但在在某发现模式中只在某 1 个发现条件下发现这一点上，该虚拟遗传基因被小分类了。

[0136] 另外，在对应分析中，对于分析结果所得到的行分数和列分数各自描绘，在 1 维的

情况下，可以配置 (biplot) 在同一直线上，在 2 维的情况下，可以配置在同一平面上，另外在 3 维的情况下，可以配置在同一空间上。

[0137] 即，如已经说明了的那样，坐标变换处理部件 2 将表示遗传基因的配置的行分数、表示发现条件的配置的列分数变换为图 4 的坐标表所示的坐标。

[0138] 另外，图像处理部件 3 从上述坐标表中顺序地读出遗传基因和发现条件的显示图像，并显示在图像显示部件 4 的显示画面上。

[0139] 在此，在显示在图像显示部件 4 的显示画面上的图像中，对应分析所得到的坐标接近的发现条件与分析对象的遗传基因的距离越是接近，则越是表示相互的关联性高。

[0140] 例如，位于表示癌症患者等的表现型（发现曲线）的发现条件的近旁的遗传基因（群）与该表现型相关的可能性高。在此，类似发现条件检索部件 5 通过利用相对于发现条件的坐标的 X 平方距离范围（前述），来抽出为被推测为关联性高的遗传基因（群）。

[0141] 另外，也可以将用于实现图 1 的发现曲线分析系统的功能的程序记录在计算机可读取的记录介质中，由计算机系统读出记录在该记录介质中的程序并执行，由此来进行发现曲线的分析处理。另外，在此所述的“计算机系统”包含 OS、外围设备等硬件。另外，“计算机系统”也包含具备主页提供环境（或显示环境）的 WWW 系统。另外，“计算机可读取的记录介质”是指软盘、光磁盘、ROM、CD-ROM 等可移动介质、内置于计算机系统中的硬盘等存储装置。进而，“计算机可读取的记录介质”也包含如经由因特网等网络、电话线路等通信线路发送程序的情况下的服务器、客户端的计算机系统内部的易失性存储器 (RAM) 那样在一定时间内保存程序的设备。

[0142] 另外，也可以从将该程序存储在存储装置等着的计算机系统，经由传输介质，或者通过传输介质中的传输波，将上述程序传输到其他计算机系统。在此，传输程序的“传输介质”是指如因特网等网络（通信网路）、电话线路等通信线路（通信线）那样具有传输信息的功能的介质。另外，上述程序也可以是用于实现上述功能的一部分的程序。进而，也可以是通过与已经将上述功能记录在计算机系统中的程序组合来实现的，所谓差分文件（差分程序）。

[0143] 根据本发明，提供一种以下这样的发现曲线分析系统，即通过通常的计算机高速地分析从下一代高速时序产生器、类似的实验方法等得到的大量的发现曲线数据，将遗传基因的发现模式可视化，容易地对新遗传基因是否具有与任意的遗传基因接近的功能进行分析，因此本发明在工业上极其有用。

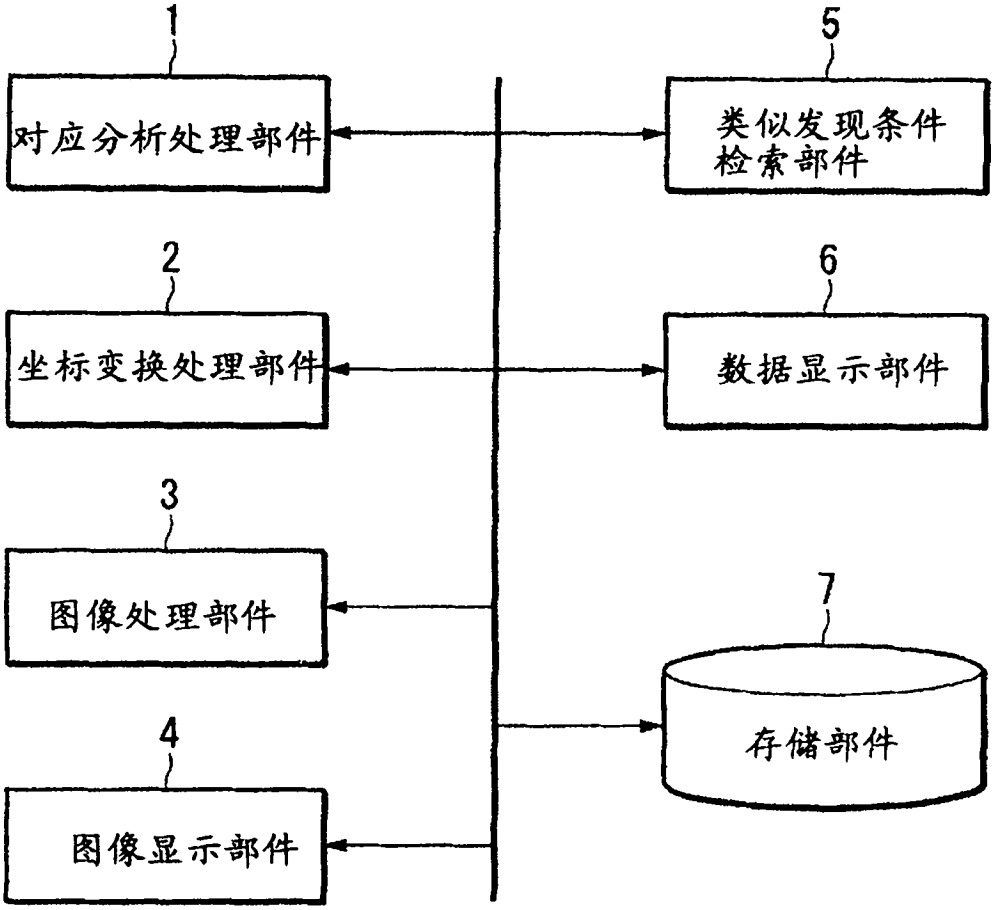
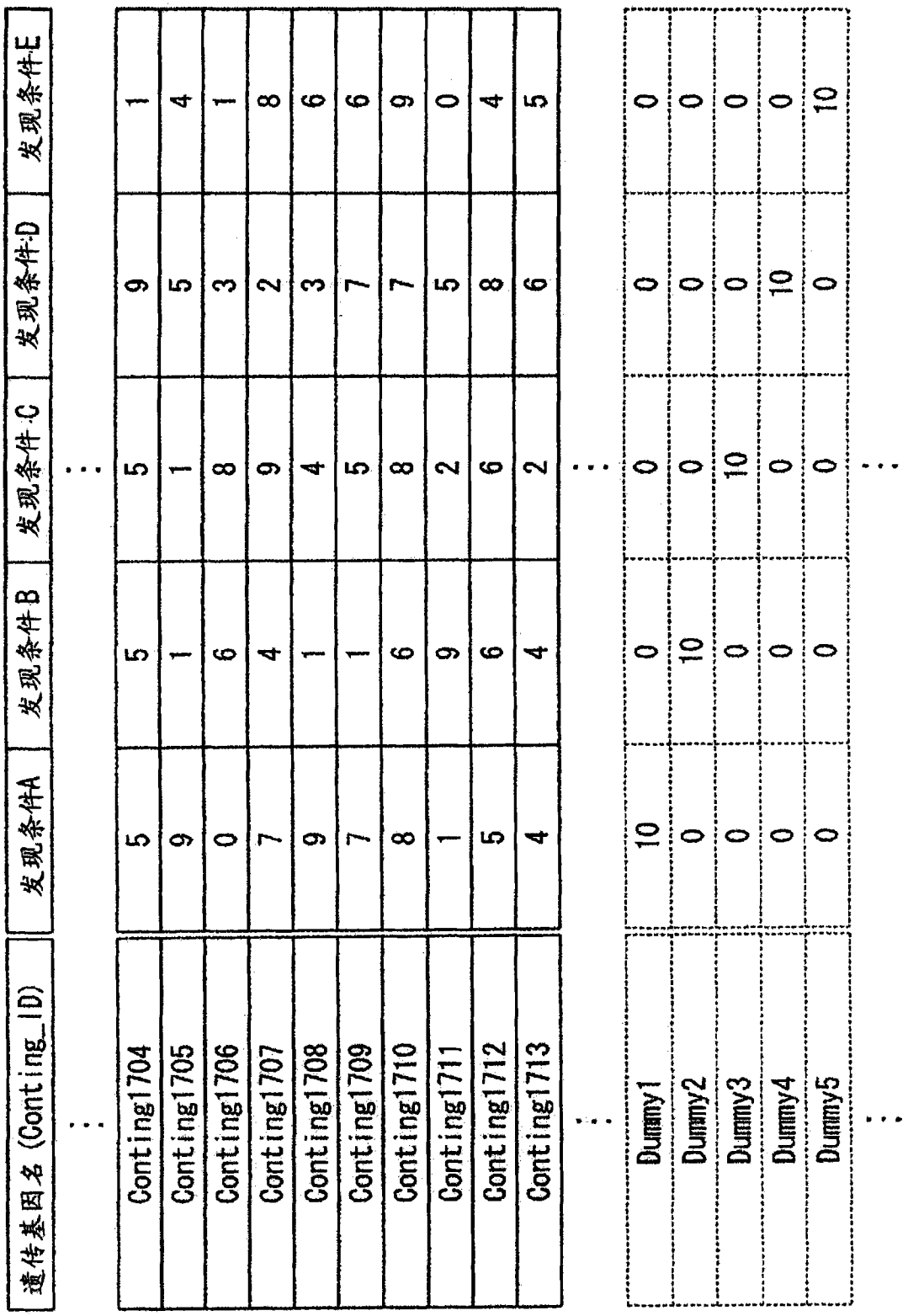

1
对应分析处理部件
2
坐标变换处理部件
3
图像处理部件
4
图像显示部件
5
类似发现条件检索部件
6
数据显示部件
7
存储部件

图 1



遗传基因名 (Conting_ID)	发现条件A	发现条件B	发现条件C	发现条件D	发现条件E
...			...		
Conting1704	5	5	5	9	1
Conting1705	9	1	1	5	4
Conting1706	0	6	8	3	1
Conting1707	7	4	9	2	8
Conting1708	9	1	4	3	6
Conting1709	7	1	5	7	6
Conting1710	8	6	8	7	9
Conting1711	1	9	2	5	0
Conting1712	5	6	6	8	4
Conting1713	4	4	2	6	5
...			...		
Dummy1	10	0	0	0	0
Dummy2	0	10	0	0	0
Dummy3	0	0	10	0	0
Dummy4	0	0	0	10	0
Dummy5	0	0	0	0	10
...			...		

图 2

遗传基因名	分数1	分数 2	分数 3	分数 4
Conting1704	0.1338102	-0.3260156	0.3687407	-0.0065748
Conting1705	-0.2586218	-0.5324871	-0.2553498	-0.3824707
Conting1706	0.1972816	0.500422	0.6076132	0.1790254
Conting1708	-0.2350021	-0.1104955	-0.2887062	-0.4572235
⋮			⋮	

图 3

遗传基因名	坐标1	坐标 2	坐标 3
Conting1704			
Conting1705			
Conting1706			
Conting1708			
⋮		⋮	

图 4

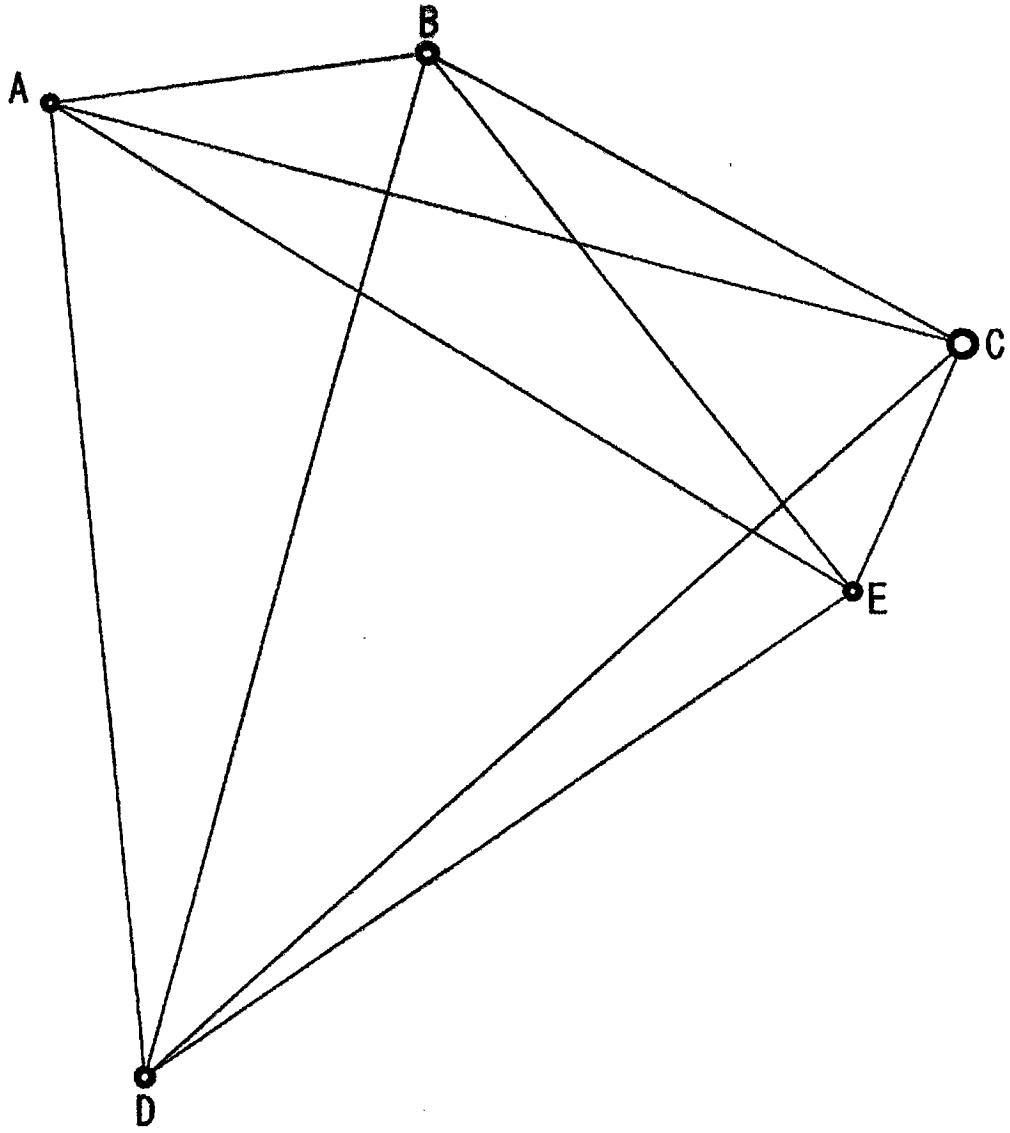


图 5

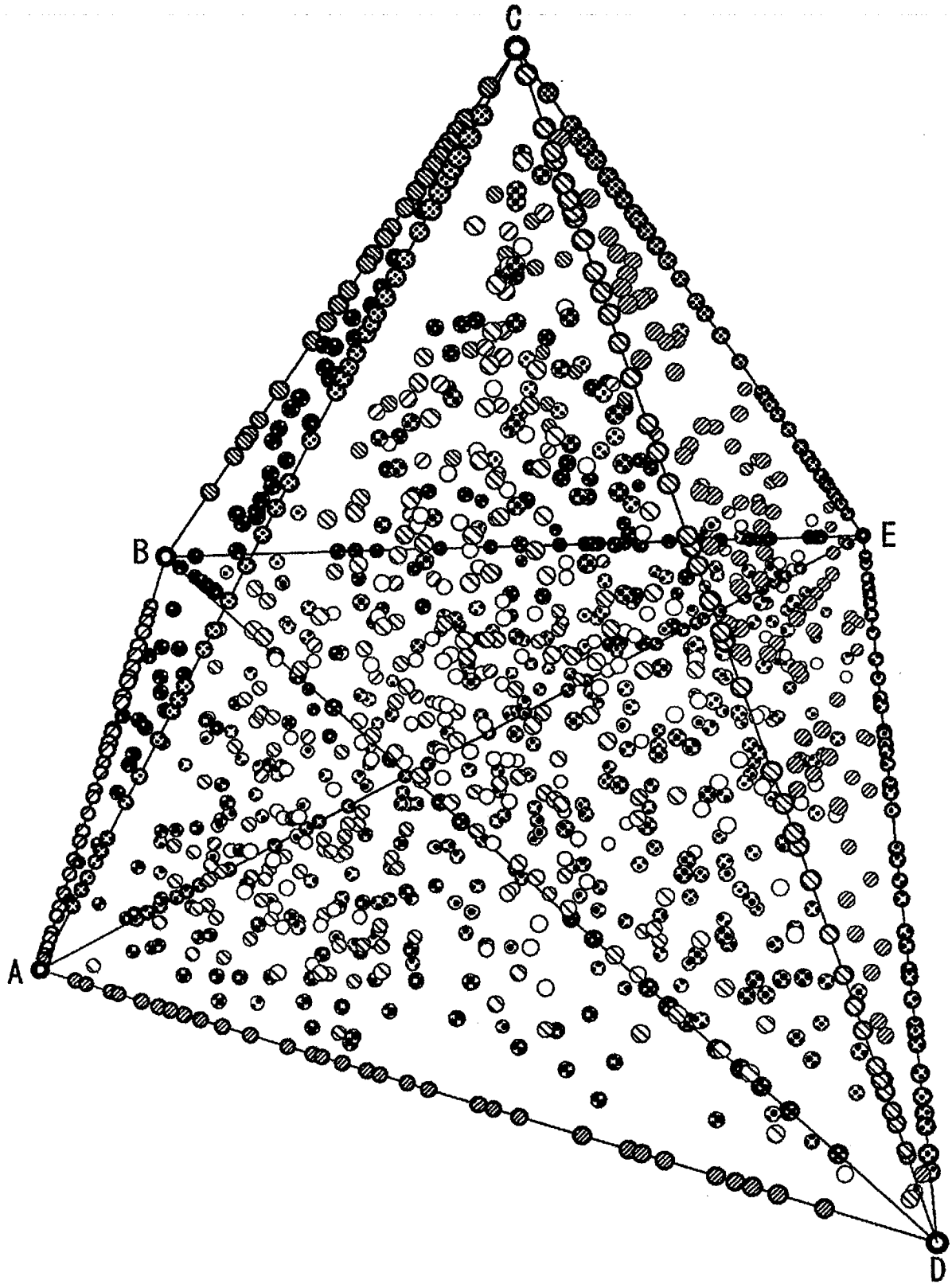


图 6

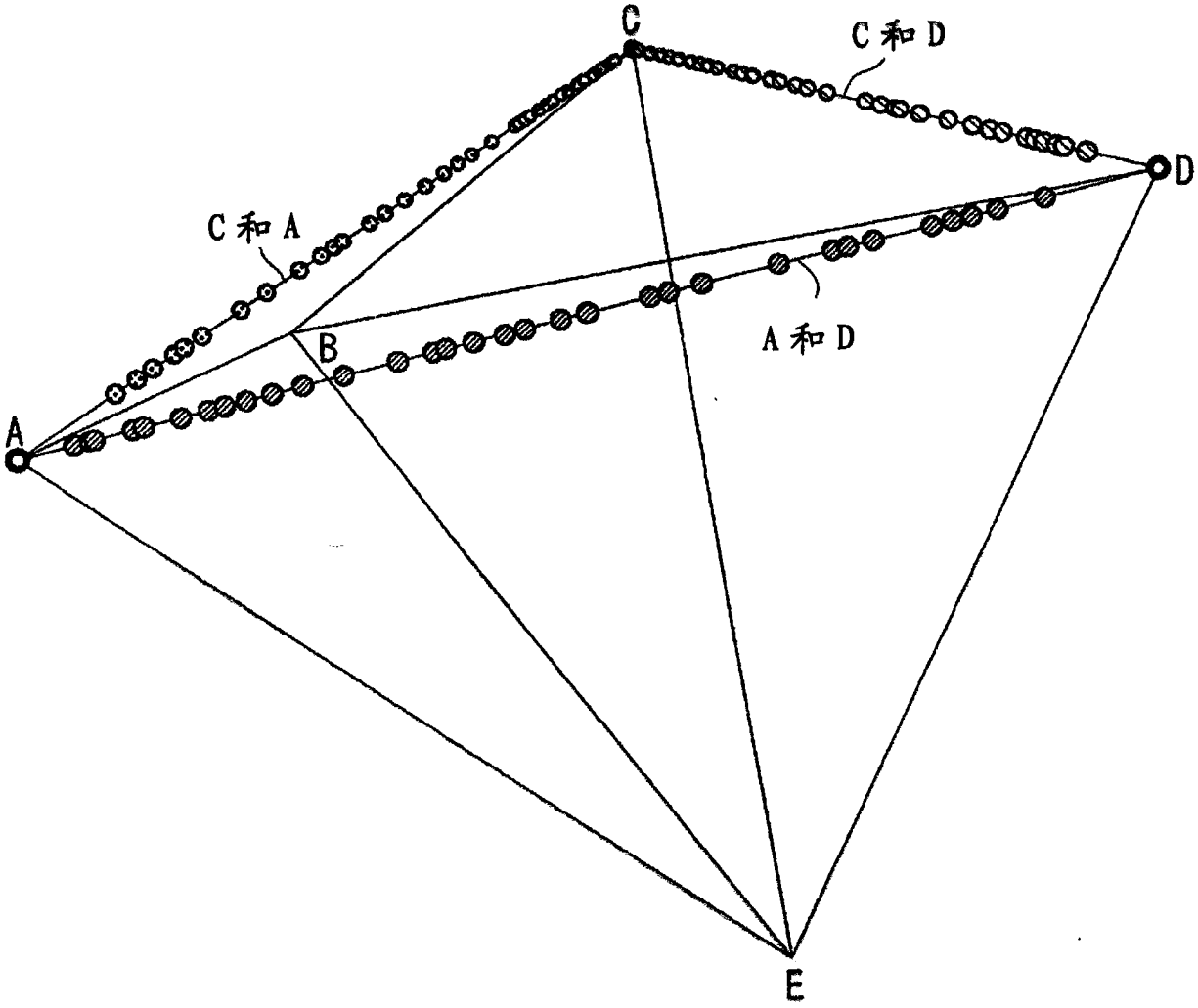


图 7

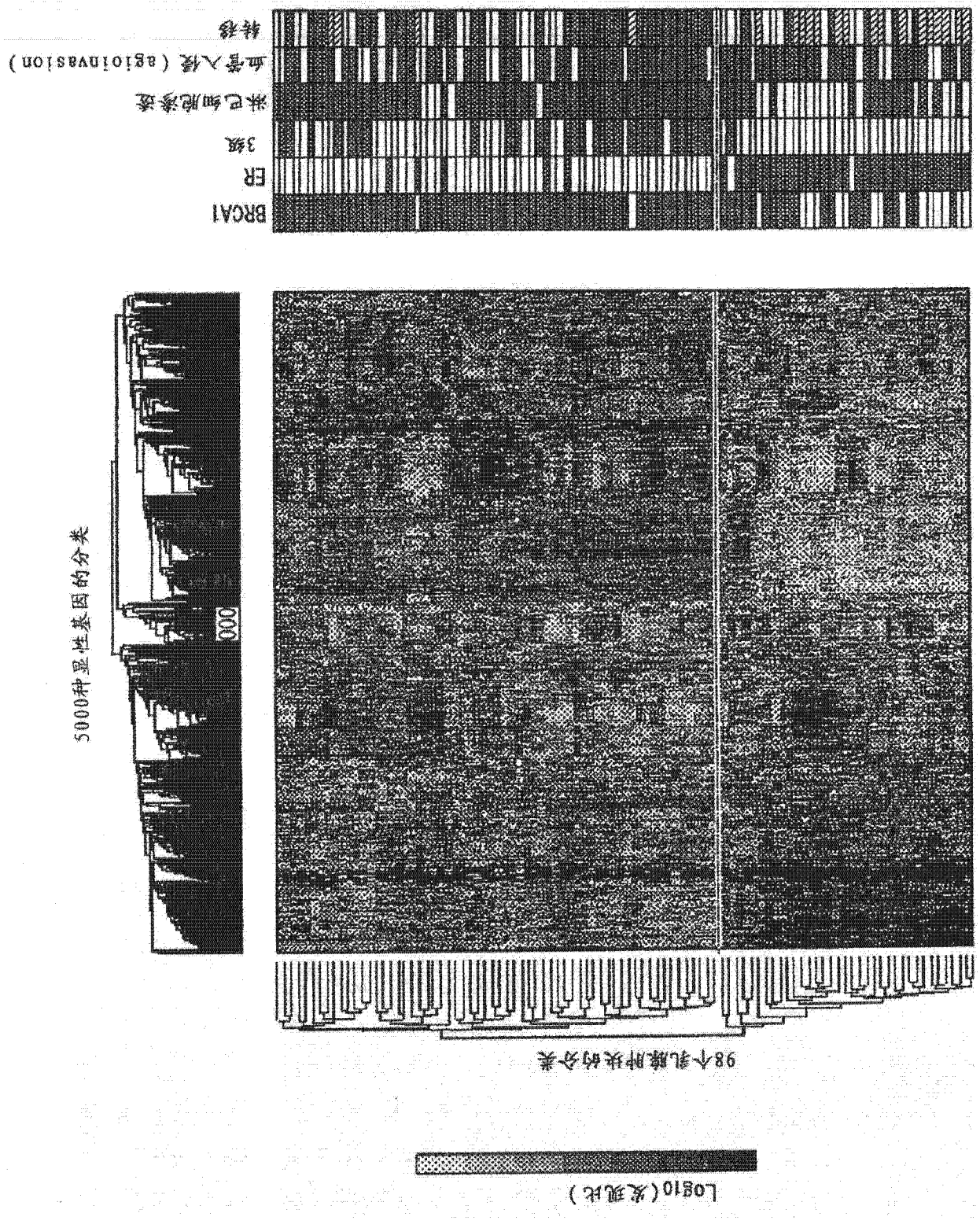

BRCA1
ER
3级
淋巴细胞渗透
血管入侵 (agioinvasion)
转移
5000种显性基因的分类
98个乳腺肿块的分类
Log10(发现比)

图 8