[19] 中华人民共和国国家知识产权局

[51] Int. Cl[7]
G01N 33/48

[12] 发明专利申请公开说明书

[21] 申请号 01822050.9

[43] 公开日 2005 年 6 月 1 日

[11] 公开号 CN 1623091A

[22] 申请日 2001.11.15 [21] 申请号 01822050.9

[30] 优先权

[32] 2000.11.16 [33] US [31] 60/249,835

[32] 2000.12.11 [33] US [31] 60/254,746

[86] 国际申请 PCT/US2001/044843 2001.11.15

[87] 国际公布 WO2002/042733 英 2002.5.30

[85] 进入国家阶段日期 2003.7.15

[71] 申请人 赛弗根生物系统股份有限公司

地址 美国加利福尼亚州

[72] 发明人 C·D·保尔瑟 E·J·加文
L·布拉金斯基 W·E·里奇
冯哲力

[74] 专利代理机构 上海专利商标事务所有限公司

代理人 李家麟

权利要求书 5 页 说明书 23 页 附图 10 页

[54] 发明名称 质谱分析方法

[57] 摘要

本文提出了一种用数字计算机分析质谱的方法。此方法包括将从大量样本中得到的质谱数据集输入数字计算机的方法。每个样本都代表了一个包含两个以上类的类集中的一个类，每个类都对应于一种不同的生物状态。本文接着提出了一个分类模型。此模型可以区分类集中的各个类。

ISSN1008-4274

1.一个用数字计算机分析质谱的方法，其特征在于，此方法包括：

a） 将从来自大量样本的质谱中得到的数据集输入数字计算机，其中每个样本被或将被指定为一个包含两个或更多类的类集中的一个类，每类由一个不同的生物状态表征，而其中每个质谱包含的数据代表信号强度作为飞行时间，质量－电荷比或由飞行时间或质量－电荷比衍生出的值的函数；且

b） 形成一个分类模型，区分类集中的类，其中形成包括通过执行代码分析数据，这些代码实施含有递归划分处理的分类处理。

2．如权利要求 1 所述的方法，其特征在于，所述质谱选自一组包括 MALDI 谱，表面增强激光解吸附/电离谱和电雾电离谱。

3．如权利要求 1 所述的方法，其特征在于，所述的类集恰由两个类组成。

4．如权利要求 1 所述的方法，其特征在于，所述样本包括选自多肽和核酸的生物分子。

5．如权利要求 1 所述的方法，其特征在于，所述的样本来自真核细胞，原核细胞或病毒。

6．如权利要求 1 所述的方法，其特征在于，所述的不同生物状态包括正常状态和病理状态。

7．如权利要求 1 所述的方法，其特征在于，所述的不同生物状态包括无病，低级癌症和高级癌症。

8．如权利要求 1 所述的方法，其特征在于，所述的不同生物状态包括给药状态和未给药状态。

9．如权利要求 1 所述的方法，其特征在于，所述的不同生物状态包括药

物有反应状态和药物无反应状态。

10. 如权利要求 1 所述的方法，其特征在于，所述的不同生物状态包括中毒状态和无毒状态。

11. 如权利要求 10 所述的方法，其特征在于，所述的中毒状态由于暴露于一种药物而发生。

12. 如权利要求 1 所述的方法，其特征在于，所述的数据集是已知数据集，数据集输入数字计算机前，每个样本都指定到一类中。

13. 如权利要求 1 所述的方法，其特征在于，所述的生成分类模型包括使用预先存在的标记物数据生成分类模型。

14. 如权利要求 1 所述的方法，其特征在于，所述的数据集通过以下方法生成：

探测质谱中的信号，每个质谱含有的数据代表信号强度作为质量－电荷比的函数；

把质量－电荷比相近的信号集簇为信号簇；

选择带有至少一个预定信号数的信号簇，其信号强度大于一个预定值；

确定对应于选定信号簇的质量－电荷比；并

用确定的质量－电荷比处的信号强度生成数据集。

15. 如权利要求 1 所述的方法，其特征在于，所述的形成分类模型包括至少一个可以区分不同的生物状态并学习的识别特性。

16. 如权利要求 1 所述的方法，其特征在于，所述的分类处理是二进递归划分处理。

17. 如权利要求 1 所述的方法，其特征在于，进一步包括：

c）询问分类模型以确定一个或多个特性能否区分不同的生物状态。

18. 如权利要求 1 所述的方法，其特征在于，进一步包括：

c）用大量样本重复 a）和 b）。

19. 如权利要求 1 所述的方法，其特征在于，所述的分类处理是一个分类和回归树处理。

20. 如权利要求 1 所述的方法，其特征在于，进一步包括形成数据集，其中形成数据集包括从质谱中获得原始数据，然后处理原始数据形成数据集。

21. 如权利要求 1 所述的方法，其特征在于，所述不同类选自暴露于药物、暴露于一类药物中的一种，和没有暴露于药物或一类药物中的一种。

22. 如权利要求 1 所述的方法，其特征在于，所述的每个质谱包括的数据代表信号强度作为质量－电荷比或由质量－电荷比衍生出的值的函数。

23. 使用数字计算机将一个未知样本分为由一种生物状态表征的类别的方法，其特征在于，此方法包括：

a）将从未知样本质谱得到的数据输入数字计算机；并

b）用权利要求 1 所述的方法所形成的分类模型处理质谱数据，将此未知样本分为由一种生物状态表征的类别。

23. 如权利要求 23 所述的方法，其特征在于，所述的类别由疾病状态表征。

24. 如权利要求 23 所述的方法，其特征在于，所述的不同生物状态包括无病、低级癌症和高级癌症。

25. 如权利要求 23 所述的方法，其特征在于，所述类别由暴露于一类药物中的一种药物表征。

26. 如权利要求 23 所述的方法，其特征在于，类别由对一种药的反应表征。

27. 如权利要求 23 所述的方法，其特征在于，所述类别由毒性状态表征。

28. 用数字计算机估计一个未知样本被准确分类为属于一种生物状态表征的类别的概率的方法，其特征在于，此方法包括：

a) 将从未知样本质谱得到的数据输入数字计算机；并

b) 用权利要求 1 所述的方法，所形成的分类模型处理质谱数据，估计该未知样本被准确分类为由一种生物状态表征的类别的概率。

29. 计算机可读介质，其特征在于，包括：

a) 将从未知样本质谱得到的数据输入数字计算机的代码；及

b) 用权利要求 1 所述的方法，所形成的分类模型处理质谱数据，将未知样本分为一种生物状态表征的类别的代码。

30. 一个系统，其特征在于，包括：

气相离子分光计；

适于处理来自气相离子分光计的数据的数字计算机；以及

权利要求 29 所述的与数字计算机协同运作的计算机可读介质。

31. 如权利要求 30 所述的系统，其特征在于，所述的气相离子分光计适于执行激光解吸附电离处理。

32. 计算机可读介质，其特征在于，包括：

a) 将从未知样本质谱得到的数据输入数字计算机的代码；及

b) 用权利要求 1 所述的方法所形成的分类模型处理质谱数据，估计未知样本被准确划分为由一种生物状态表征的类别的概率的代码。

33. 一种系统，其特征在于，包括：

气相离子分光计；

适于处理来自气相离子分光计的数据的数字计算机；以及

权利要求 32 所述的与数字计算机协同运作的计算机可读介质。

34．如权利要求 33 所述的系统，其特征在于，所述的气相离子分光计适于执行激光解吸附电离处理。

35．计算机可读介质，其特征在于，包括：

a）输入从大量样本质谱得到的数据的代码，其中每个样本被或将被指定为一个包含两个或更多类的类集中的一个类，每类由一个不同的生物状态表征，而其中每个质谱包含的数据代表信号强度作为飞行时间、质量－电荷比，或由飞行时间或质量－电荷比衍生出的值的函数；且

b）用分类处理形成分类模型的代码，此分类处理包括递归划分处理，其中分类模型区分类集中的类。

36．如权利要求 35 所述的计算机可读介质，其特征在于，所述的分类处理是分类和回归树处理。

37．一个系统，其特征在于，包括：

气相离子分光计；

适于处理来自气相离子分光计的数据的数字计算机；以及

权利要求 35 所述的与数字计算机协同运作的计算机可读介质。

38．如权利要求 37 所述系统，其特征在于，所述的气相离子分光计适于执行激光解吸附电离处理。

质谱分析方法

相关申请

本申请声明享有美国临时专利申请 2000 年 11 月 16 日归档的 Nos. 60/249,835 及 2000 年 12 月 11 日归档的 Nos. 60/254,746 的权利。这些 U.S. 临时专利申请也包含在此处，以供各种参考。

技术领域

本发明是关于质谱分析方法的。

背景技术

最近关于基因组的研究已经可以识别与不同疾病相关的许多种基因。但是，尽管基因组研究可以识别出与遗传性的对一种疾病有易患体质的基因，仍然需要对诸如蛋白质等标记物进行表征和识别。“标记物”通常是指一种多肽或其它一些可以区分不同生物状态的分子。在病理状态下，蛋白质和其它标志物是重要的因素。例如，蛋白质可随生物状态的改变（例如生病）而改变。它们也可以传达疾病、中毒或其它刺激的信息。在生病时，某些蛋白质进入休眠状态，而另一些蛋白质则激活。例如，前列腺特异性抗原（PSA）是一种循环性的血清蛋白，其含量增加与前列腺癌有关。如果蛋白质含量的增加能被快速检测到，医生就可以及早诊断疾病并提高疗效。

识别出新的标志物是诊断学药物研发过程中最早最难的步骤之一。辨别一种物质是否是某种疾病的一种方法是，判别相对于没有这种疾病的病人样本，它们在表现出这种疾病的病人的生物样本中，是否被“差别表达”。例如，图 1（a）显示了来自一组 18 个有病患者的样本的大量互相重叠的质谱中的曲线图 100。此患者可能患有例如前列腺癌。图 1（b）所示的另一幅曲线图 102 显示了来自一组 18 个一般病人的样本的互相重叠的质谱。曲线图 100、102 的信号强度都绘成由质量－电荷比的函数。曲线图 100、102 中的信号强度与分子量的标记物浓度成正比，这些标记物与样本中的质量－电荷比 A 相关。如曲线图 100、102 所示，在质量－电荷比 A 处，两幅质谱中都有若干信

号存在。这些信号中的峰值代表了可能的其分子量与质量－电荷比 A 相关的标记物。

将曲线图 100、102 中的信号结合起来看，很明显有病的患者在质量－电荷比 A 处的平均信号强度比一般患者高。称质量－电荷比 A 处的标记物对有病患者“差别表达”，因为平均起来，此标记物的浓度在有病患者样本中高于一般患者样本。

观察图 1（a）、1（b）中的数据，一般可以得出结论，有病患者样本中质量－电荷比 A 的标记物浓度大于一般患者。由于这种标记物在有病患者样本中的浓度高于一般患者，这种标记物可对这种疾病表征为“正相关”。如果有病患者样本中标记物的浓度低于一般患者，此蛋白质可称为“负相关”。

标记物一旦发现，就可以用作诊断工具。例如，关于上述例子，一个测试病人的未知样本可以用质谱仪分析并得到一个质谱。可对质谱进行分析并得到测试病人质谱中质量－电荷比 A 处的信号强度。此信号强度可与有病患者和一般患者在质量－电荷比 A 处的平均信号强度相比较。据此可推断此测试病人是否已患或将要患上癌症。例如，如果此未知样本在质量－电荷比 A 处的信号强度更接近有病患者在此处的平均信号强度，则可以推断此患者更可能已经或将要患癌症。

虽然所述差别表达分析有效，但是有许多方面可作改进。例如，分析患者生物样本中的一个标记物如 PSA 的量，对监护病症的发展常常不够可靠。PSA 被认为是目前可用的最好的前列腺癌标记物之一。但是，它并不总是能正确区分良性和恶性前列腺疾病。尽管生物样本中的一种标记物（如 PSA）的浓度分析能在一定程度上判断一个测试病人是否有病，还是需要一个有更大可信度的方法。

而且，当分析大量生物样本的大量质谱时，难以分清究竟哪个信号代表可能区分有病和无病状态的标记物。典型的生物样本质谱包含有大量潜在的标记物信号（例如，大于 200）以及大量噪声，这会使得潜在的重要信号以及平均信号差的识别很困难，因此也就使得识别和量化潜在的标记物很困难。除非潜在的标记物表现出强正相关或强负相关，否则，有病和无病患者的样本的平均信号差将不易区分。例如，通常难以直观看出一组质谱中给定质量处的一簇信号的平均信号强度比另一组质谱中的一簇信号高还是低。另外，很多潜在的重要信号可能强度值较低。谱中的噪声可能会模糊这些潜在的重

要信号。这些信号可能无法被发现并被无意中从差别表达分析中忽略掉了。

能有更好的分析质谱的方法将是很好的事。例如，它可以提供一个更准确地发现潜在有用标记物的方法。它也可以提供一个改进的分类模型，可以用来判断一个未知样本是否与一种特定生物状态有关。

本发明本身就是用于解决这些以及其它一些问题的。

发明内容

本发明是关于质谱分析的方法。在本发明中，利用数字计算机形成一个分类模型，可用于将与不同生物状态关联的样本分类开。此分类模型可用做预诊断工具。也可以用来识别相关于某种生物状态的潜在标记物。进一步，此分类模型可通过诸如回归分离处理这样的处理形成。

本发明的一实施例是关于使用数字计算机进行质谱分析的方法。这个方法包括：a)将从大量样本中得到的质谱数据集输入计算机，其中每个样本都会或将会归于一组包含两个或两个以上类的类集中的一个类，每一类代表一种不同生物状态的特性，其中每个质谱包含代表信号强度的数据，此强度是质量－电荷比或由质量－电荷比得到的一个数值的函数；并且 b) 形成了一个分类模型，可以用来区分类集中的不同类，模型的形成包括通过执行代码来分析数据集，此段代码实施包含回归分离处理的分类处理。

本发明另一实施例包括使用数字计算机进行质谱分析的方法。这个方法包括：a) 将从大量样本中得到的质谱数据集输入数字计算机，其中每个样本都会或将会归于一组包含两个或两个以上类的类集中的一个类，每一类代表一种不同生物状态的特性，其中每个质谱包含代表信号强度的数据，此强度是飞行时间或由飞行时间得到的一个数值的函数；并且 b) 形成了一个分类模型，可以用来区分类集中的不同类，模型的形成包括通过执行代码来分析数据集，此段代码实施回归分离处理。

本发明还包括一个计算机可读的媒介。此媒介包括：a) 将从大量样本中得到的质谱数据输入计算机的代码，其中每个样本都会或将会归于一组包含两个或两个以上类的类集中的一个类，每一类代表一种不同生物状态的特性，其中每个质谱包含代表信号强度的数据，此强度是飞行时间或由飞行时间得到的一个数值的函数，或质量－电荷比或由质量－电荷比得到的一个数值的函数；以及 b)通过回归分离程序生成分类模型的代码，此分类模型可以区分

类集中的类。质谱可使用例如激光解析电离过程生成。

本发明还包括使用数字计算机将一个未知样本分类为代表一类生物状态特性的方法。此方法包括：a）将从未知样本得到的质谱数据输入计算机；以及 b）用分类模型处理质谱数据，从而将未知样本划分为对应于某种生物状态特性的类。分类模型可通过回归分离处理生成。

本发明还包括估计使用计算机将未知样本准确划分为对应于某种生物状态特性的类的可能性的方法。此方法包括：a）将从未知样本得到的质谱数据输入计算机；以及 b）使用分类模型处理质谱，以估计未知样本准确划分为对应于某种生物状态特性的类的可能性。分类模型可通过回归分离处理生成，且由来自归应于两个或更多个不同生物状态的类的样本的质谱数据集生成。

本发明中，所分析的质谱可能是预先存在的，例如，可能在分类模型生成之前早就存在了。或者，质谱数据也可能在分类模型生成的同时产生。

本发明的这些以及其它一些实施例参照图和详细说明来描述。

附图说明

图 1（A）所示为有病患者样本的重叠的质谱。

图 1（B）所示为一般患者样本的重叠的质谱。

图 2 所示为本发明实施例生成质谱的方法的流程图。

图 3 所示为对数归一的强度与识别出的峰值簇的函数图。图中显示了两组不同样本的质谱的信号强度。

图 4 所示为根据本发明实施例质谱预处理的首选的流程图。

图 5 所示为根据本发明实施例质谱预处理及生成分类模型的首选的流程图。

图 6 所示为本发明实施例的一个系统框图。

图 7 所示为本发明实施例的分类和回归树。

图 8 所示为不同预测变量可变重要性的表格。

图 9 所示为癌症患者和一般患者不同样本的凝胶体视图。

图 10 所示为癌症患者和一般患者不同样本的谱图。

具体实施方式

本发明中，从质谱图得到的数据集输入数字计算机并生成分类模型。质

谱图则来自已知特性的生物样本。本发明中，用于生成分类模型的数据集被表征为“已知”数据集，因为在这些数据集被用来生成分类模型之前，与这些生物样本有关的生物状态是已知的。作为比较，“未知”数据集包含的数据来自如下样本的质谱，这些样本若与质谱生成时由分类模型所辨别出的生物状态相关，则是不确定的。未知数据可从要用这个分类模型进行诊断的测试病人的生物样本得出。有时，已知数据集也称为“训练数据”。

作为演示，下述很多例子中都使用已知数据集来生成分类模型。但是，在本发明的某些实施例，用来生成分类模型的数据集可能是未知数据集。例如，在簇分析中，未知生物样本的质谱如果模式相似，则可能被分到一组。对每组样本进行分析以找出它们是否有共同的生物状态。如果有，那么这组样本就可以划分为与这种生物状态相关的一类。例如，生成一组有共同模式的质谱后，可以确定组内的所有谱都来自受辐射的生物样本。则这组样本就可以指定为“受辐射”状态类。其它组的样本同样可以定为具备各组共同生物状态特征的类。这样就可以生成一个分类模型，而未知谱就可以用这个生成好的分类模型进行分类。

本发明中，每个所用样本都被或将被分到含有两个或更多个类的类集中的一个类，每个类对应于一种不同的生物状态特性。例如，第一类样本可能与诸如疾病状态的生物状态有关。第二类样本的质谱可能与无病状态的生物状态有关。第一类和第二类样本就可以形成类集。每一类各自的质谱都含有可以区分开这两类的数据。

在本发明中，所分析的每一个质谱都可以包含信号强度数据，此数据可以是飞行时间，由飞行时间得到的值（例如，质量－电荷比，分子量，等等），质量－电荷比，或由质量－电荷比得到的值（例如，分子量）。如本领域所周知的技术，从飞行时间质谱仪得到的质量－电荷比值由飞行时间值导出。质量－电荷比也可以通过其它方式得到。例如，不用飞行时间谱仪来确定质量－电荷比，使用四极分析器和磁质量分析器的质谱仪也可以用来确定质量－电荷比。

在较佳实施例中，每个质谱都含有信号强度数据，此数据是质量－电荷比的函数。在典型的谱视图型的质谱中，信号强度数据作为质量－电荷比的函数在信号强度图上可能以“峰”的形式存在。每个峰有底部和顶部，峰宽从底部向顶部逐渐变窄。通常与峰相关的质量－电荷比对应于峰的顶部。峰

的强度通常也与峰的顶部相关。

通常，质量－电荷比与潜在标记物的分子量相关。例如，如果一个潜在标记物电荷为+1，则其质量－电荷比就等于此信号所代表的潜在标记物的分子量。这样，一些质谱图可能会显示信号强度为分子量的函数，实际上这里的分子量是由质量－电荷比衍生出来的。

由于这里讨论的本发明的许多特定实施例都是关于质量－电荷比的应用，可以理解，在所特别讨论的作为示例的实施例中的质量－电荷比值，可以用飞行时间值或其它衍生自飞行时间值的值代替。

尽管所分析的质谱中的每一个质谱所包含的信号强度数据都可以作为飞行时间，但是这里使用的质谱的信号强度数据通常是质量－电荷比的函数。离子的飞行时间值会受到仪器的影响，但是质量－电荷比值不会受到影响。例如，在飞行时间质谱测量过程中，对离子得到的飞行时间值会依赖于此特定质谱仪的自由程管道长度。自由程管道长度不同的质谱可能会对同一个离子得到不同的飞行时间值。质量－电荷比就不会这样，因为它只是离子的质量与电荷之比。用质量－电荷比值生成的分类模型也可以不受所使用的特定质谱仪的影响。

数据集可以包含任何适合的数据并可以自动或手动输入计算机。在计算机上运行的分类程序所处理前的数据可能是原始数据也可能是预处理过的数据。例如，质谱中预定质量－电荷比处的信号原始强度可以用作数据集。或者，这些数据在分类模型生成之前也可以先进行处理。例如，有些实施例中，会使用质谱信号强度的对数值（如行为 2）生成数据集。

数据集输入计算机。实施分类处理的计算机代码用数据集生成分类模型。示例的分类处理包括分级分类处理，如分类和回归树处理，多参量统计分析，如簇分析，非线性处理，如神经网络分析。在本实施例中，数据集使用分类和回归树处理进行处理，从而生成分类模型，如分类和回归树。下面将对这些以及其它一些分类处理和分类模型进行更详细的阐述。

生成的分类模型可以是预测性的或描述性的。例如，模型可用来预测一个未知测试生物样本是否与一个特定生物状态相关。或者进一步，分类模型可用来验证那些用于区分所分析的生物状态的数据的特性。一个特性包括可以区分所分析的特定类的质谱数据的任何方面。可识别的适合的特性包括，但并不仅限于，一个或多个质量－电荷比下的信号强度或信号强度范围、信

号形状（例如，峰的形状）、信号区域（例如，峰的区域）、信号宽度（例如，诸如峰底部的宽度）、每个质谱中的信号数目，等等。典型的情况下，分类模型可指明一个特性，如给定质量－电荷比下的特定信号强度，可以区分有病和无病样本。而在另一种情况，分类模型可指明几个不同特性的组合可以将有病样本和无病样本区分开来。例如，不同质量－电荷比下的两个或更多信号的信号强度范围可用来区分有病样本和无病样本。

还有的情况下，可确定为区分不同样本类型的合适特性可能是一类里特定质量－电荷比下的信号发生频率。例如，对有 100 个样本的有病类和有 100 个样本的正常类，一个在质量－电荷比为 X 处强度为 Y 的信号可能在 90 个有病类样本的质谱中出现，而只在 10 个正常类样本中出现。即使此信号在有病和无病样本中的平均强度相同（也就是说，平均强度为 Y），由于此信号在癌症病人类中出现较多，因而此特性可以区分有病类和无病类。诸如这样的频度特性可用此分类模型确定。

任何合适的生物样本都可以用于本发明的实施例。生物样本包括组织（如，活组织切片）、血液、血清、血浆、乳汁、尿、泪、唾液、细胞、软硬组织、器官、精液、粪便、尿等等。生物样本可来自任何适合的有机体，包括真核原核或病毒有机体。

生物样本可含有生物分子，包括大分子，如多肽、蛋白质、核酸、酶、DNA、RNA、多核苷酸、低核苷酸、核酸、碳水化合物、低聚糖、多糖；上述生物大分子片段，如核酸片段、肽片段及蛋白质片段；上述生物大分子的聚合物，如核酸聚合物、蛋白质-DNA 聚合物、受体－配合基聚合物、酶－酶作用物、酶抑制剂、肽聚合物、蛋白质聚合物、碳水聚合物及多糖聚合物；小生物分子，如氨基酸、核苷、核苷、糖、类固醇、脂质、金属离子、药、荷尔蒙、氨基化合物、胺、羧酸、维生素和辅酶、酒精、醛、酮、脂肪酸、卟啉，类胡萝卜素、植物生长调节素、磷酸盐酯及二磷酸核苷糖，合成小分子，如例如药物或疗法有效剂、单体、肽类体、类固醇类体、抑制剂、诱导剂、抗有丝分裂药物、抗生素、离子载体、抗代谢物、氨基酸类体、抗菌素、输运抑制剂、表面激活剂（表面活性剂）、线粒体和叶绿体功能抑制剂、电子释放物、携带物和接收物、合成蛋白酶培养基、磷酸酶培养基、酯酶和脂肪酶和蛋白质修正试剂培养基；以及人工聚合体、低聚体和共聚物。上述物质任何适当的混合或化合也可以包含在生物样本中。

如上面指出的，用于产生数据集的生物样本会分到有两个或更多类的类集中的一个类。每个类表征了一种不同的生物状态。更适宜的情况下，只有两个类和两个生物状态；每类一个生物状态。例如，一类具有病态的生物状态，而另一生物状态具有无病状态。

如此处所用的，样本的“生物状态”是指样本或从中导出样本的有机体或其它来源的生物状态的表征特征。此特征可以是像遗传特性或显性特性这样的生物特性。此特征可以是生理或疾病特性，比如有没有某种特定疾病，包括传染性疾病。此特征也可以是样本所暴露的条件（环境，社会，生理，时间相关，等等）。

遗传性特性包括一种特定基因是否存在，一种基因的多种形态，或基因的组合。基因特性可以表现为显性，或以对其表现易感的形式存在，例如对特定疾病易感（如，对某种癌症或心脏病的倾向）。

显性特性包括，诸如外观，生理特性，物理特性，神经状态，心理状态、反应特性等等，或者对特定药物有无反应。显性特性可包括所谓“正常”和“病理”特性的存在与否，包括疾病特性。另一种状态是一种特定疾病的存在与否。状态也可以是特定人或组所具有的状态，如不同个人、不同家庭、不同年龄状态、不同种族以及不同组织类型。

在本发明的有些实施例中，生理状态可以是，比如下面的一个或更多个的组合形式：疾病状态，正常状态，病理状态，服药状态，未服药状态，药物反应状态，无药物反应状态，以及良性状态。服药状态可包括病人吃过药的状态，而未服药状态可包括病人未吃过药的状态。药物反应状态是生物样本对用药的反应的状态。疾病状态的一些特例包括，如，癌症、心脏病，自身免疫性疾病、毒感染、Alzheimer 病，以及糖尿病。更特殊的癌症状态包括，例如，前列腺癌、膀胱癌、乳癌、结肠癌以及卵巢癌。生物状态还可以包括早期、中期和晚期。例如，不同生物状态可包括某种疾病，例如癌症的早期，中期和晚期。

其它状态可能与样本类所属于的不同环境有关。例如热疗、电磁辐射、锻炼、食谱地理位置等等中的一个或多个。例如一类生物样本（如所有的血液样本）可来自一组受到辐射的病人，而另一类样本可来自一组未受辐射的病人。辐射源可以是预期辐射源，如 X 光机，也可以是未预期辐射源，如手机。又例，一组人可能使用一种特定食谱，而另一组则用不同的食谱。

在本发明的另一些实施例中，不同的生物状态可能对应于与各自不同药物或药物类型相关的样本。例如，生成一幅来自使用已知疗效药物的病人的样本的质谱图。这种已知疗效药物的质谱图就可以代表与之同类的药物。例如，这种已知疗效药物的质谱图可以代表与其相同或相近的特性，结构或同样的基本效果的药物。例如，许多不同的止痛化合物都可以缓解病人的疼痛。这种已知疗效的药物以及与之相同或类似的药物可能在人体内控制同样的生化通道，从而对人体产生同样的效果。这种生物通道（如上下调节蛋白质）的特性可在质谱图中反映出来。

分类模型可通过与已知疗效药物和几种不同药物，或完全不用药的相关质谱生成。一旦分类模型生成，就可以产生与效用未知的候选药物相关的候选样本的质谱。利用此分类模型，就可以对与候选样本相关的质谱进行分类。分类模型可以确定候选样本是与一种已知效用的药物相关还是与另一种不同药物相关。如果，比方说，分类模型认为候选样本与已知效用的药物相关，则候选药物对人体的作用就可能与这种已知效用的药物相同。因此，除了其它应用，本发明可用于发现和/或表征药物。

I.获得质谱

质谱可通过任何适当的处理获得。例如，质谱可通过可访问一个或多个质谱数据库的本地或远端服务器计算机取到（如下载）。这些数据库会包括与不同生物状态相关的不同生理样本的质谱库。另外，质谱可从生物样本生成。不管如何得到，用于生成分类模型的质谱和样本更适于在同样的条件下处理，以确保质谱的任何改变都是由于样本本身，而不是由于处理的差异造成的。质谱可通过特定分类处理生成，也可以不通过特定的数据分类处理产生。

本发明实施例中，应用气相离子分光计生成质谱。气相离子分光计是一种仪器，测量可以转化为样本离子化为气态时形成的离子的质量－电荷比的参量。这包括，例如，质谱仪，离子活性分光计，或总离子流测量设备。

质谱仪可使用任何适当的电离技术。电离技术包括如电子电离，高速原子/离子轰击，阵列辅助激光解吸附/电离（MALDI），表面增强激光解吸附/电离（SELDI），或电雾电离。

在一些实施例中，离子活性分光计可用于探测和表征一个标记物。离子活性分光计的原理是基于不同离子不同的活动性。特别，电离产生的样本离

子因质量、电量或形状的不同而以不同的速率运动，并在电场的作用下通过一个管道。离子（典型地是以电流的形式）在探测器处记录下来，探测器的输出就用来辨别样本里的标记物或其它物质。离子活性分光计的优点在于，它可以在大气压下操作。

在较佳实施例中，使用激光解吸附飞行时间质谱仪产生质谱。激光解吸附分光计特别适用于分析像蛋白质这样的大分子量的物质。例如，MALDI 或表面增强激光解吸附/电离处理的实用质量量程可达 300，000 道尔顿或更多。而且，激光解吸附处理可用于分析复杂混合物且灵敏度高。进一步，在像 MALDI 或表面增强激光解吸附/电离处理这样的激光解吸附处理中，蛋白质破碎的概率要比在其它质谱仪处理低。因此，激光解吸附处理可用于精确表征和量化蛋白质这样的大分子量的物质。

在典型的产生质谱的处理中，对质谱仪的入口系统引入了一个带有标记物的探头。然后将标记物电离。标记物离子产生后，产生的离子被一个离子光学部件收集，然后质量分析器使经过的离子分散并加以分析。离开质量分析器的离子被探测器探测到。在飞行时间质量分析器中，离子在一个短的高压电场中加速并漂移进入一个高度真空室。在高度真空室的远端，加速过的离子在不同时刻撞击一个敏感的探测器表面。由于离子的飞行时间是其质量－电荷比的函数，电离和冲击之间消逝的时间可用于识别特定质量－电荷比的分子存在与否。

然后飞行时间数据就可以转化为质量－电荷比来产生一个谱，表示标记物的信号强度作为质量－电荷比的函数的情况。图 2 显示了将基于飞行时间数据的质谱转化为质量－电荷比数据的一种示例方法的流程图。首先，采集飞行时间谱（第 16 步）。然后，对飞行时间谱进行平滑滤波（第 18 步）。典型地，在最初产生的谱中含有大量高频噪声。应用不同的滤波器以降低噪声而不损坏其中包含的信号。然后，计算一条基线（第 20 步）。这样可以去除特有的向上的漂移，例如 MALDI 或表面增强激光解吸附/电离处理的特征。

“表面增强”解吸附/电离处理是指一种在其上把样本暴露于能源的基底在解吸附/电离处理中起积极作用的处理。在这类方法中，基底例如探针，并不仅仅是样本表达的一个被动场所。有几种表面增强基底可应用于表面增强解吸附/电离处理。一个例子是，表面含有亲和材料，例如阴离子交换组或亲水组（例如，二氧化硅），它们会优先绑定某类分子。此类亲和材料包括如

硅烷醇（亲水）C8 或 C16 烷基（憎水）、固定金属螯化物（并列同价物）、阴离子或阳离子交换（离子的）或抗体（生物特效）。样本暴露于基底绑定吸附物，以根据特定的吸引力基础绑定分析物分子。典型的不绑定分子则被冲洗掉。当分析物是生物分子时，能量吸收材料例如基质，会与绑定样本相关。然后用激光对探测器探测到的分析物进行解吸附和电离。

另一种型式中，基底表面包含有一个能量吸收分子的绑定层，可以避免将样本和基质材料混合，例如在 MALDI 中。表面增强解吸附/电离方法在诸如 U. S. 专利 5,719,060(Hutchens and Yip) 和 WO 98/59360(Hutchens and Yip) (U.S. 专利 6,255,047) 中有描述。当激光将含有能量吸收材料的基质解吸附时，一些基质材料会随所分析的样本材料一道被解吸附。基线计算调整谱图，以补偿解吸附的基质材料呈现的信号。一旦基线计算出来，飞行时间/质量转化就会发生（第 22 步）。此步骤中，飞行时间数据被转化为质量－电荷比。然后计算本地噪声值（第 24 步）。在低质量－电荷比情况下，由于解吸附的基质材料而产生了大量噪声。在电离解吸附处理中，基质材料在高质量-电荷比时比时比低质量－电荷比时更不易解吸附，因此噪声在低质量－电荷比时比高质量－电荷比时更可能存在。可对谱进行调整以修正这个效应。修正之后，谱更新就完成了（第 26 步）。通过图 2 所示方法处理质谱后，质谱的信噪比提高了，能更好地量化和比较潜在的标记物。

通过标记物的解吸附和探测所生成的质谱可在生成质谱图前后用数字计算机进行预处理。数据分析可包括确定探测到的标记物的信号强度（例如，信号的高度）和去除“逸出值”（偏离预设的统计分布的数据）的步骤。例如，观察信号可进行归一化。归一化就是计算每个信号相对于某个参考的高度的处理。例如，参考可以是由仪器和化学物品（如，能量吸收分子）所产生的背景噪声，它们在标尺中设为零值。然后，探测到的每个标记物或其它物质的信号强度可以按需要的尺度（如 100）以相对强度的形式显示出。或者，可对样本提出一个标准，从而符合此标准的信号可以用来作为计算观察到的每个标记物或探测到的其它标记物的相对信号强度的参考。

数字计算机可以将结果数据转化为各种格式显示。一种格式称为“谱图或滞留物图”，可以显示标准谱图。谱图描述了在不同分子量下到达探测器的标记物的量。另一种格式称为“峰值图”，只保留了谱图中的峰值高度和质量信息，使图更清晰，使代表具有相近分子量的标记物的信号更易于观察。

还有一种格式称为“凝胶体视图”，将峰值图中的每个质量转化为基于每个峰值高度的灰度图像，得到的图看起来很像电泳凝胶体中的条带。还有一种格式称为“3-D 覆盖图”，将几个谱重叠，以研究相对峰值高度的微小变化。还有一种格式称为“差别视图”，比较两个或多个谱，便于高亮显示代表标记物的信号和代表样本间上下调节的标记物的信号。任何两个样本的标记物形状（谱）可在一幅图上可视地比较。用于生成数据集的数据可从这些及其它一些质谱显示格式中得到。

II. 生成数据集

一旦得到质谱，就可以生成一个数据集，如已知数据集。数据集包含从生物样本类集的质谱中得到的数据。生成数据集的质谱数据可以是原始的，未处理过的数据。例如，质谱中已识别质量值处的原始信号强度值可用来生成数据集。又例如，质谱中的原始信号模式也可以用来生成数据集。

在另一些实施例中，数据在用来生成分类模型之前可以先进行预处理。于是质谱可用任何适当的方法进行处理，然后再生成分类模型。例如，质谱中的数据可作如下处理：取信号强度的对数值，去除逸出值，去除较不可能与潜在标记物有联系的信号，去除低强度信号，等等。

在有些实施例中，数据集可包含原始数据或与每个质谱的特定模式相关的预处理过的模式数据。例如，对一个包含许多信号峰的质谱，信号峰的模式可以包含用来生成此质谱的生物样本的独特特征。分类处理可以根据模式或模式段区分不同的谱，这些模式或模式段在分类模型进行区分的各个不同类内部对谱是一致的。电脑程序例如神经网络程序，可以接收与已知生物状态关联的已知样本的多个质谱。神经网络可用质谱数据进行训练，然后就可以区分属于各自不同类的质谱模式。然后，训练过的神经网络就可以根据质谱模式对与未知样本关联的质谱进行分类。

在其它实施例中，数据集包含关于质谱中信号强度的数据。在这些实施例中，每个质谱中的一些或全部信号都可以用来生成数据集。例如，低于谱图类质谱图中所有信号（如，峰值）的强度可用来生成数据集。在优选实施例中，先确定质量一电荷比，然后用确定的质量一电荷比从质谱中选择信号。所选信号的强度可用来生成数据集。通过只用每个质谱中的部分信号的数据来生成数据集，所需要处理的数据点减少了，因而数据处理进行得更快。代

表可接受的标记物的概率低的信号数据可从数据集中排除出去。

质量－电荷比可以通过任何方式确定。例如，可通过比较不同生理状态的不同类的质谱确定。选择可能可以区分这些类的信号的质量－电荷比。比较可以手工进行（如视觉比较）或由数字计算机自动进行。例如，与样本中不同类关联的质谱可以直观地互相比较，以确定一个样本类里－质谱中一个质量－电荷比处的信号强度是否明显大于或小于另一个样本类里同一个质量－电荷比处的信号强度，由此指出潜在的区分表达。信号有差别处的质量－电荷比就可以选择出来。

例如，图 3 显示了 log（2）归一化的强度对确定的峰簇的曲线。这张图显示了以 2 为底的对数归一的强度值。峰簇里的每个强度值都已经减去了平均强度值，因此零值代表与均值无差别。Y 轴上每单位表示与均值相差两倍。使用图 3 所示这样的图可以确定强上下调节的蛋白质。图 3 显示了对数归一化的强度作为不同信号簇函数的图。图中显示了来自两组不同样本的质谱的信号强度。例如，图 3 中的峰簇 22（x 轴上）表示组 A 和组 B 的数据点之间大范围的变化。这表明与峰簇 22 关联的这个质量－电荷比可确定为一个候选的标记物位置。

另外，或进一步，某些预先定义的标准可用来初步选择某些信号或信号簇。选到的信号簇再用来确定特定的质量－电荷比。例如，可自动选择信号强度或信号强度均值高于或低于信号强度阈值的信号或信号簇。然后就可以确定与所选信号或信号簇关联的质量－电荷比。

包括收集质谱数据、数据预处理以及处理预处理过的数据以生成分类模型的优选方法可参考图 4 和 5 描述。参照图 4，收集与不同生理状态关联的样本质谱（步骤 27）。采集的样本数尽量大。例如，在本发明中，采集的样本数在 100 左右到 1000 左右（或者大于或小于这些值）。较佳地，所有用来生成谱的样本都在同一条件下生成，从而样本间的差别可以在谱中反映出来。

每个谱中相应于潜在标记物存在的信号都确定出来。每个这样的信号都指定了一个质量值。然后，第一组质谱的每个质谱中信噪比大于预设值的信号都被检测出来（步骤 28）。在典型的例子中，信噪比大于值 S 的信号都被检测出来。值 S 可以是绝对或相对值。质谱中这些质量－电荷比下的信号聚到一个簇里（步骤 30）。然后选出符合预设标准的信号簇。例如，在一实施例中，有预设数量信号的信号簇可被选出（步骤 32）。少于预设值的簇则被

丢弃。在典型的例子中，如果一个簇里的信号数少于质谱数的 50%，此信号簇就可以丢弃。在有些实施例中，选择处理得到从少达 20 个到多于 200 个选出的信号簇。一旦信号簇被选出，这些信号簇的质量－电荷比就可以确定（步骤 34）。

一旦质量－电荷比确定下来，此质量－电荷比的“丢失信号”就可以确定。一些质谱可能在所确定的质量－电荷比处不显示信号。这组质谱或与这组质谱关联的样本可进行再分析，以确定在确定的质量－电荷比处信号是否确实存在（步骤 36）。对丢失信号添加估计值（步骤 38）。对于一个簇中找不到任何信号的谱，根据迹线高度或噪声值估计一个强度值。这个估计的强度值是使用者可选的。

参照图 5，一旦质量－电荷比确定下来，所有质谱在确定质量值处对每个信号的强度值也确定（步骤 46）。每个信号的强度值都归一化为 0 到 100，以消除绝对幅度的影响（步骤 48）。然后，对每个归一化的信号强度取对数（如以 2 为底的）（步骤 50）。取信号强度的对数可以去除测量中的歪斜。

对数归一化的数据集接着用分类过程处理（步骤 52），实施该过程的代码由数字计算机执行。代码执行后，分类模型就形成了（步骤 54）。生成分类模型的进一步描述如下。

III. 生成分类模型

由数字计算机执行的代码实施的分类过程可处理数据集。代码可由数字计算机执行以产生分类模型。代码可存储于任何适合的计算机可读介质。计算机可读介质的实例如磁，电，或光盘，磁带，记忆棒，芯片，等等。代码可用任何合适的计算机编程语言书写，包括 C,C++，等等。

数字计算机可以是微机，小型机或大型机，使用标准或专门的操作系统，比如 Windows™系列操作系统。另一部分实施例中，数字计算机可以只是一个或多个微处理器。数字计算机与用来生成质谱的质谱仪可以是分离的。或者，数字计算机可以耦合或物理集成到质谱仪中。质谱数据可手工或自动从质谱仪传送到数字计算机。例如，在一实施例中，可先从大量质谱得到一个已知数据集。这个已知数据集然后被手工输入执行分类过程代码的数字计算机中。另一种实施例中，质谱数据的生成和采集、数据预处理，以及用分类过程处理预处理过的数据，都可以用同一个物理计算设备进行。

有些实施例中，已知数据集可被定为训练集，可为分类模型或提前形成的分类模型“训练”一个前体。分类模型可像生成那样被训练并进行学习。例如，在神经网络中，已知数据集可用来训练神经网络，使之识别输入神经网络的数据类间的差别。初始分类模型生成后，可用更大数量的样本来进一步训练和提炼分类模型，以使它能更准确地区分用来生成分类模型的类。

本发明中，附加的数据可用于生成分类模型。附加数据可以相关于质谱也可以不相关。例如，在有些实施例中，预存在的标记物数据可附加用于已知数据集来形成分类模型。例如，得到一类前列腺癌病人样本的质谱和一类没有前列腺癌病人样本的质谱。这些质谱可形成一组已知数据集。用这组已知数据集和预存在的标记物数据，例如预存在的 PSA 诊断数据（如，PSA 临床化验数据），可生成分类模型。附加的预存在的 PSA 诊断数据可用来协助区分质谱以形成分类模型。例如，可估计每个质谱来观察相应于 PSA 的质量－电荷比处的信号是否与前列腺癌病人的信号强度特性更接近，还是与非前列腺癌病人的特性更接近。该信息有助于对前列腺癌病人或非前列腺癌类指定质谱及其相应的样本。在其它实施例中，被取生物样本的人的非质谱数据，如性别、年龄、等等，也可以用来形成分类模型。例如，如果男性比女性更容易患某种特定疾病，则此信息也可以用来帮助区分样本并形成分类模型。

任何适合的分类处理都可用于本发明。例如，分类处理可以是分级分类处理，如分类和回归树处理或多变量统计分析。多变量统计分析同时观察几个变量间关系的模式。多变量统计分析的例子包括辨别函数分析和簇分析等已知处理。辨别函数分析是基于每组前面的观察指定组的观察的分析方法。簇分析是把数据间的变化用一系列集合表示的统计方法。例如，生物上，这些集通常以分级形式组织并用一个树状图表示，称为系统树图。一些簇分析类型及其它分类处理在 Jain 等人的下面文章里有叙述，“Statistical Pattern Recognition: A Review”，*IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. 22, No. 1, January 2000. 这篇文章在此处完整引用。

或者，分类处理可以使用非线性分类处理，如人工神经网络分析。人工神经网络分析可用已知数据集进行训练。一般来说，人工神经网络可以根据几个可能影响它的其它输入变量预测一个输出变量的值。此预测通过从一个已知模式集中选择一个看起来与一种特定情况最相关的模式得到。顾名思义，

人工神经网络有几个神经元（单元）和它们之间的联系。这些单元根据功能分类到不同的三个层或组。第一组形成输入层，接收输入到系统里的数据。第二组形成输出层，输出代表输出模式的输出数据。第三组包括很多中间层，也称为隐含层，把输入模式转化为输出。

作为例子，神经网络可被训练为区分与有病状态和无病状态关联的激光解吸附质谱。然后，用激光解吸附处理生成待测生物样本的质谱并把相关于此质谱的数据输入训练过的神经网络。训练过的神经网络就可以确定测试生物样本与有病状态还是无病状态有关。

本发明中，分类处理较佳地包括分级、回归分类处理，例如分类和回归树处理。在本发明中，分类和回归树处理由数字计算机执行。一个可作为例子的分类和回归树程序是 CART4.0，可从 Salford Systems,Inc. 购买（www.salford-systems.com）。

二进递归分类处理是一个特别有效的分类和回归树处理。此处理是二进的，因为每个父节点总是恰巧分为两个子节点，它是递归的，因为此处理可循环将每个子节点作为下一个父节点。为了给一个已知数据集分类，对已知数据集提问。在本发明中，待分类的数据是相应于生物样本类集的质谱。每个质谱可当作一个待分类的“实例”。可用来对实例进行分类的问题可以是“质量－电荷比 X 处的信号强度比 Y 处大吗？”每个问题将已知数据集分为有更多共性的两组。一旦找到一个最佳划分，分类和回归树处理对每个子节点重复搜索处理，继续递归，直到无法再进行划分或终止。划分只有当一个特定节点只有一例或此节点中所有实例都是同一类型时才会停止。

对数据集提出的问题可由使用者决定或由数字计算机自动决定。在有些实施例中，问题可由数字计算机任意产生，数据划分质量确定此问题是否可接受。例如，可对数据提出一个问题。如果分类结果对此例有统计显著性，此问题就可以保留并用于生成分类和回归树。分类和回归树处理找出分类数据所需问题的最佳数目，补偿每次样本观察时的随机误差影响。

分类和回归树处理观察分析中包含的所有预测变量的所有可能的划分。例如，对有 215 个实例和 19 个预测变量的数据集，处理考虑 215 乘以 19 一共是 4085 种可能的划分。典型地，在形成分类和回归树时，所有这样的划分都要考虑。因此，形成的分类和回归树处理在生成分类模型时考虑许多不同的预测参量。例如，在典型实施例中，形成分类模型时，类集的所有质谱中

超过 100 个质量－电荷比处的信号数据都进行了考虑。作为比较，前面提到过的微分表达分析只考虑一个预测参量。因此，由于在形成分类模型时使用了每个质谱中更多的数据，分类和回归树实施例比其它分类方法准确性更高。

为检查此模型的准确性，分类和回归树处理可以使用计算机增强技术，称为交叉确认。在典型的交叉确认处理中，先生成一棵大树，然后剪除回去。数据集分成 10 个大致相等的部分，每个部分包含所分析的生物状态的相近的分布。数据的前 9 个部分用来建立最大可能树。剩下的一部分用来获得所选子树错误率的初始估计。对数据的其它 9/10 重复同样的过程（生成最大可能树），而用另外 1/10 部分作为测试样本。此过程一直持续到数据的每一部分都作为测试样本保留过一次。10 个最小测试样本的结果接着合起来形成每个可能大小的树的错误率。这些错误率应用于基于整个数据集的树。交叉确认提供了树的独立预测准确性的可靠估计。即使无法得到独立测试样本，也可以预测树分类全新数据（例如，来自大量未知样本的数据）的准确度。

生成的分类和回归树表示了那些预测参量（如果有的话）对样本组间的差异有作用。分类和回归树可用来分类（预测一个实例属于那一组），也可以用来回归（预测一个特定值）。也可以用来识别对区分所分析的类有重要作用的特性。例如，分类模型可能指明特定质量－电荷比下一个或多个信号强度值分别或其组合是区分所分析的类的重要特性。

分类和回归树图示地显示了数据间发现的关系。分类和回归树处理的一个基本输出就是树本身。树可以作为分类模型的一方面，可由用户可视地进行分析。与神经网络分析这样的非线性技术不同，树提供的可视表达使分类分析很易于理解和接受。因此，与训练过的神经网络特性这样的“黑箱”分类模型相比，用户更倾向于相信决定树的结果。这使得分类和回归树成为更愿意接受的分类模型，无论是对不同的健康监护和常规人员（如，食品和药物管理），以及希望对生成分类模型的分析有详细理解的病人，都是如此。树也可以用来发现所分析的数据和生物状态间从前不知道的联系。

分类和回归树处理与神经网络这样的分类处理相比还有其它优点。例如，分类和回归树程序比神经网络更有效率，因为后者常常需要传递大量训练集数据，有时甚至数以千计。而建立决定树所需要的传递量不会多于树的层数。对树的层数没有预定限制，虽然由树的深度和宽度衡量的树的复杂性随预测参量的数目增加而增加。

应用分类和回归树模型，还可识别区分类的特性。数据中识别出的特性可以是所分析的生物状态的特性。例如，分类模型可指出特性的一种组合会与特定的生理状态相联系。例如，模型指明不同质量－电荷比下特定的信号强度将有病状态和无病状态区分开来。与传统的差别分析处理相比，本发明中分析了许多不同的参量。分类模型可以确定用来区分所分析的生物状态一个或多个预测参量。

IV.使用分类模型

分类模型可用来将未知样本分到一个生物状态中。用这种方法，待测样本的质谱与关联于某种特定生理状态的分类模型进行比较，以确定此样本是否适合分到此生物状态。可得到未知样本的质谱，从此未知样本的质谱得到的数据可输入数字计算机。输入的数据可用分类模型处理。然后分类模型可以将未知样本分为某一特定类。此类可能与某一特定生物状态相关联，而患者就可以诊断为具有那种特定生物状态。

此方法特别有临床应用价值。例如，在药物发现过程中，人们希望能确定一个候选分子在生物体系统中是否与一种特定药品或一类药品（例如，一类 seratonin 再起抑制剂）产生的生物效果相同。首先产生分类模型，它可区分暴露于感兴趣的药品或一类药品的生物系统（如，人或实验动物）。然后，生物系统置于实验分子中，产生系统样本的质谱。然后质谱分类为属于或不属于测试的已知药品或一类药品的类。如果候选分子被划为这一类，此信息对决定是否对此药物进行进一步研究很有用。

在其它应用中，开发的分类模型可以区分不同的毒性和无毒生物状态。毒性状态可能由于例如暴露于药物或一类药物引起。也就是说，可以开发分类模型来指明一种药或一类药是否会在生物系统（如，体内或试管内模型系统，包括肝中毒）中引起中毒反应。然后，可对此系统测试研发中或临床试验中的药物是否有毒。此信息在药物开发的毒性研究中有用。

在其它应用中，开发的分类模型可以区分对某种药物是否有反应的人群。然后，在给不知道是否对药物有反应的人服药之前，可以用质谱仪测试此人的样本，并且划分此人为对此药有反应或无反应类。

在其它应用中，开发的分类模型可以区分有没有某种疾病的人。然后一个接受诊断的人可以提交一个样本，用于分类此人为有病还是没病。这样，

此方法在临床诊断中也有用。

一实施例用来分析癌症。病理学家根据组织表现给癌症分级。低级癌症特性有增大的细胞核，其细胞核/细胞质比有不太大的增大，有丝分裂数目少，不太大的细胞异质，以及一般正常结构的保持。高级癌症特性有增大的形状怪异的细胞核，其细胞核/细胞质比很高；有丝分裂数目增加，其中有些看起来非典型；与正常结构类似很少或没有。发展一个分类模型可以区分一个生物样本是没病、低级癌症，还是高级癌症是很有用的，因为这种诊断能确定疗法，还可以预测预后情况。样本可以是怀疑区域的固体组织切片或好的针吸样本。不过，在另一实施例中，样本可从被测个体组的更简单的采集源得到，比如尿、血液或其它体液。这对于会分泌细胞或蛋白质到这些液体中的癌症特别有效，如膀胱癌、前列腺癌和肺癌。当这些状态的分类模型建立后，就可以用于分类待测病人的样本作诊断。在其它应用中，开发的分类模型用于区分个体类别是否有一种非病理的特定物理或生理特性。然后，对此特性未知的样本可通过测试此个体的一个样本并将其谱分为在或不在有此特性的类里而进行分类。

分类模型也可以用来估计一个未知样本准确地分类为属于对应于一个生理状态的类别的概率。例如，在分类和回归树中，可以确定可能的误分类的概率。作为例子，用分类和回归树模型区分一个患者的未知样本是属于有病态还是无病态。此模型可以估计误分类的概率。例如，如果误分类概率低于10%，则可以通知患者有90%的概率患有此病。

V.包含计算机可读介质的系统

本发明的一些实施例是关于包含计算机可读介质的系统。一个包含计算机可读介质和数字计算机的示例系统的方框图在图 6 中显示。系统 70 包括耦合到数字计算机 74 的质谱仪 72。显示器 76 如视频显示器和计算机可读介质 78 可操作性地耦合到数字计算机 74。显示器 76 可用来显示数字计算机 74 产生的输出。计算机可读介质 78 可用来存储数字计算机 74 所执行的指令。

质谱仪可操作性地耦合到数字计算机 74，而不需要物理上或电耦合。例如，可从质谱仪得到数据（如上所述），然后数据可由操作员手工或自动输入数字计算机 74。在其它实施例中，质谱仪 72 可自动把数据传送到可被处理的数字计算机 74 中。例如，质谱仪 72 可从一个或多个生物样本产生原始数

据（如数据飞行时间）。然后数据传送到数字计算机 74，在那里进行预处理或处理。处理数据的指令可从计算机可读介质 78 中得到。来自质谱仪的数据处理过后，输出在显示器 76 上显示。

计算机可读介质 78 可以包含任何适于处理来自质谱仪 72 的数据的指令。例如，计算机可读介质 78 可以包含将来自未知生物样本质谱的数据输入数字计算机 74 的计算机代码。然后数据可以用分类模型处理。分类模型可以估计此未知样本被准确划分为某一生物状态表征的类的概率。

尽管方框图将质谱仪 72、数字计算机 74、显示器 76 和计算机可读介质 78 用分开的框图表示，但可以理解，这些部件中一个或多个可以在同一个或不同的框里表示。例如，在有些实施例里，数字计算机 74 和计算机可读介质 76 可在同一个框里，而质谱仪 72 和显示器 76 在另外的框里。还有的实施例中，所有的部件 72、74、76、78 都可以放在同一个单元里。

例子

从一个生物样本集的一些生物样本中生成大量质谱。此样本集包括来自一般病人的第一类血清样本和来自前列腺癌病人的第二类血清样本。每个病人的血清样本都通过一个表面增强激光解吸附/电离系统，此系统可从 Ciphergen Biosystems, Inc. of Fremont, California 购得。Ciphergen Biosystem 的 ProtienChip®技术也用于本例。关于 ProtienChip®技术的进一步介绍在网站www.ciphergen.com上找到。每个样本的输出结果是信号强度与质量－电荷比的质谱图。离散的峰代表了质谱中的信号。

特定质量－电荷比处的信号强度对应于带有此质量－电荷比的蛋白质数量。例如，高信号强度表明高的蛋白质浓度。每个质谱里的信号都进行了定位、定量和选择。在此例中，如果质谱段的强度值至少是背景噪声的两倍，则认为是可接受的。所有质谱中大约在同样的质量－电荷比下的信号被簇集起来。分簇以后，大约共确定了 250 个信号簇，并标记为 P1 到 P250。P1 到 P250 中的每个信号簇都对应于一个特定的质量－电荷比并被定性为“预测参量”。

每个质谱在确定质量－电荷比下的信号强度组成了已知数据集。这些信号强度被输入分类和回归树程序 CART4.0，可从 Salford Systems, Inc. (www.salford-systems.com)购得。程序由数字计算机执行。数字计算机生成一个分类和回归树。利用这些数据，每个样本都分类为正常或癌症。

质谱数据输入后，数字计算机产生一个图 6 那样的树。此例中，类 0 是正常类而类 1 是癌症类。每个质谱都可以定性为一个“实例”，在树中进行分类。

树中的每个方框代表一个“节点”。最顶部的节点 1 称为根节点。决定树从根节点开始生长，每层将数据分开以形成新的节点。树枝连接新节点。不再分裂的节点称为终端节点。图 6 的树中的终端节点标记为终端节点 1 到 7。如下面将详细解释的，终端节点 1 到 7 可用来对未知样本分类并因此用于预测。

在每个节点中，大部分设定了整个节点的分类。例如，终端节点 1 有四个患者。这四个患者都有癌症。因此终端节点 1 就定性为一个癌症节点。由于所有的实例都有同样的值（癌症），此节点就定性为“纯的”并不再分裂。如果终端节点 1 含有三个癌症患者和一个一般患者，由于其中大部分患者是癌症患者，此节点仍会被定性为癌症节点。此时，这一个一般患者就认为是被误分类了。

图 6 中，每个节点都含有关于此节点的实例数和生物状态癌症的分布的信息。根结点（节点 1）·的实例是质谱数据集中所有的实例。节点 1 含有 194 个实例，其中 96 个正常，98 个癌症。节点 1 分裂为两个新节点，节点 2 和节点 5。数据分裂通过确定簇 P127 的平均信号强度是否小于或等于 3.2946 确定。平均信号强度以及值 3.2946 都是相对尺度。如果此问题的答案是对，则相应的实例就放到节点 2。如果答案是不，则相应的实例就放到节点 5。此例中，85 个癌症患者和 11 个一般患者的质谱在与预测参量 P127 关联的质量一电荷比处的信号强度小于或等于 3.2946，因此被置于节点 2。85 个一般患者和 13 个癌症患者的质谱在与预测参量 P127 关联的质量一电荷比处的信号强度大于 3.2946，置于节点 5。其它节点处类似地使用不同的分类规则进行操作，从而形成树。

分类和回归树的预测效果可以参看表 1 和表 2。

表 1—学习数据误分类

分类	N 样本	N 误分类	差错百分比
0（正常）	96	0	0
1（癌症）	98	0	0

表 2—测试数据误分类			
分类	N 样本	N 误分类	差错百分比
0（正常）	96	9	9.38
1（癌症）	98	11	11.22

分类和回归树程序将已知数据集分成两组。大约 90%的数据用作学习集，大约 10%的数据用作测试集。分类和回归树先用学习集数据生成。树生成后，用剩下的 10%数据进行测试，以观察分类和回归树分类数据的准确性。参看表 1，用生成的分类和回归树，所有的学习集数据都被正确分类。参看表 2，正常情况和癌症情况测试数据的百分数分类错误率分别为 9.38%和 11.22%。反过来，正常情况和癌症情况的分类成功率分别为 90.62%和 88.78%。

像这样的分类成功率表明分类和回归树对分类未知生物样本是高度准确的模型。在分类处理中，分类机制里考虑了多个预测参量。质谱里可以用比前述差别分析法多得多的数据来对与质谱关联的样本进行分类，前述方法只使用了单一质量一电荷比处的平均信号强度来分类一个测试患者。因而，此分类模型在分类测试患者时比许多传统模型更为准确。

树一旦生成，就可以对未知样本进行分类，从树根（顶）开始，沿树枝路径前进，直到遇到终端节点。路径通过对未知样本质谱中的预测参量值应用分裂规则确定。例如，如果一个测试病人未知血清样本的质谱在预测参量 P127、P193 和 P187 的质量一电荷比处的信号强度分别为 1.0、0.05、0.9，则测试病人将被分类到节点 1、节点 2、节点 3，然后最终达到终端节点 1。终端节点 1 是癌症节点，于是病人将被分类到癌症病人。

图 7 显示了一些预测参量（如信号簇）中每个变量重要性的表格。变量重要性表格通过预测参量在建构分类和回归树中的重要性对它们分级。如果一个特定预测参量可以强烈区分质谱数据，则它在构建分类树中是重要的。为了计算一个参量的重要性分值，CART 观察了每个参量作为初级分类代替物的改进测量属性。这些改进的值在每个节点处加起来，然后汇总，然后相对于效果最佳的参量标定。改进最大的参量得分 100，所有其它参量都得到一个较低的分数，直到 0 分。

图 7 中，分类模型表明在生成分类和回归树时，预测参量 P36,P127 和 P90 比其它参量更重要。因此它们在区分癌症和无癌症类时也比其它参量更重要。与这些预测参量相关联的质量－电荷比也与可以将患有前列腺癌和没有前列腺癌的样本区分开的潜在标记物相关。因此，分类模型可用来确定一个或多个可以用来区分所分析的类的标记物。

树模型的有效性可由图 8 和图 9 得到确定。图 8 中的视图是凝胶视图，而图 9 中的视图是轨迹视图。谱放缩到 P127 在质量－电荷比为 5075 道尔顿（电量为+1）处代表的信号。图 8 和 9 表明六个前列腺癌患者和六个无前列腺癌的患者样本中的标记物在相应于预测参量 P127 的质量值为 5075 道尔顿处区分表达了。如图 6 所示，预测参量 P127 是树中的第一个节点。又如图 7 所示，预测参量 P127 能比其它预测参量更有效地区分前列腺癌患者和无前列腺癌的患者。

前述都是关于本发明的某些优选实施例，本发明的其它及进一步的实施例仍可从本发明的基本范围中作出。这些改动的实施例也应归于本发明的范围内。本发明一个或多个实施例的特性可与本发明其它实施例的一个或多个特性组合，但仍不离开本发明的范围。

本申请中引用的所有出版物（如网站）和专利文件都作为参考整个合并进来，以期达到每个出版物或专利文件都分别引用至同等程度。通过将它们引用到此申请里并不表示任何特定参考是他们发明的“原有技术”。

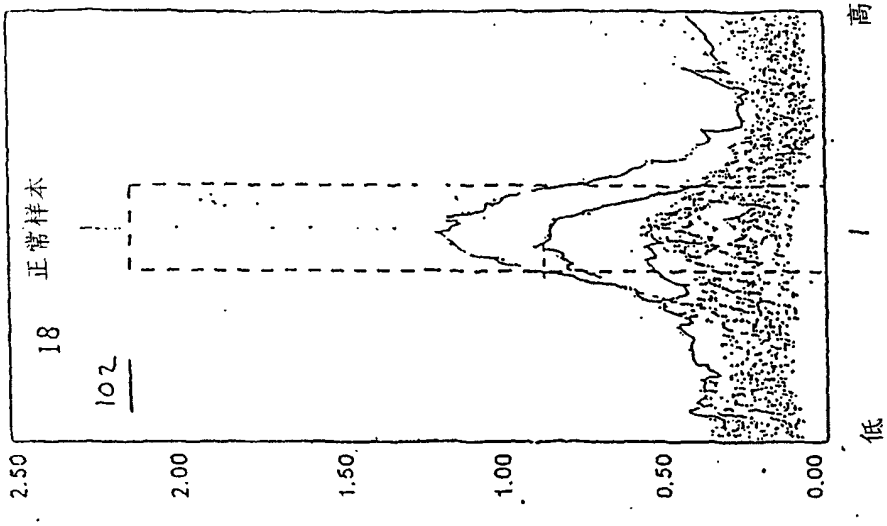

正常样本
18
102
2.50
2.00
1.50
1.00
0.50
0.00
低
高
质量－电荷比

图 1B

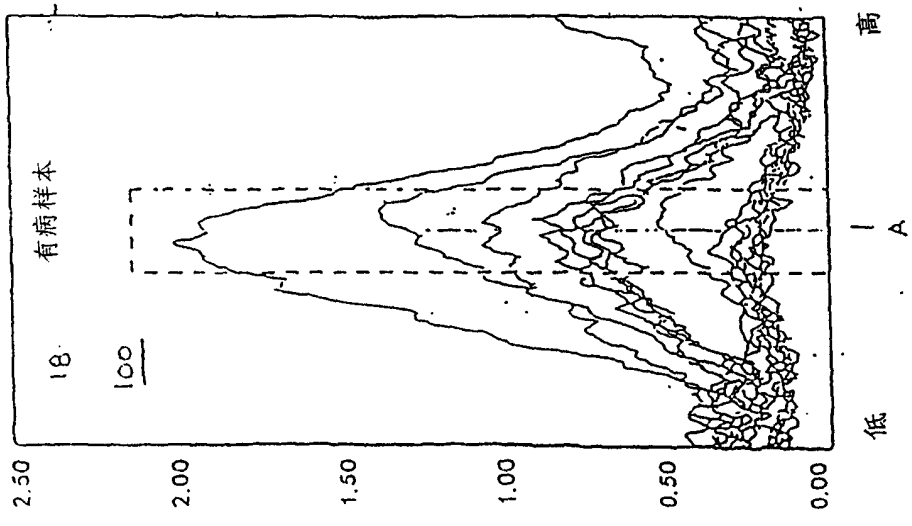

有病样本
18
100
2.50
2.00
1.50
1.00
0.50
0.00
低
高
质量－电荷比

图 1A

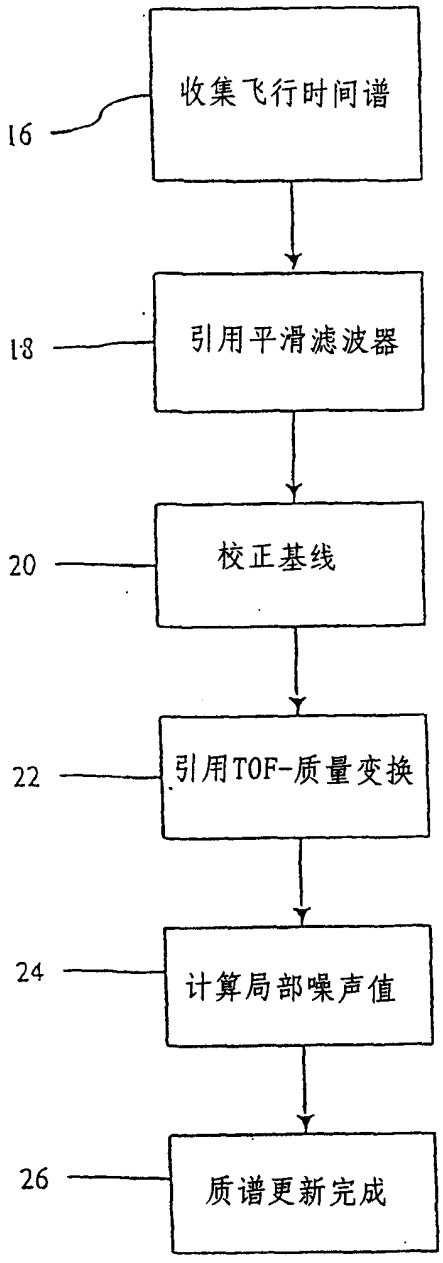

收集飞行时间谱
16
引用平滑滤波器
18
校正基线
20
引用TOF-质量变换
22
计算局部噪声值
24
质谱更新完成
26

图 2

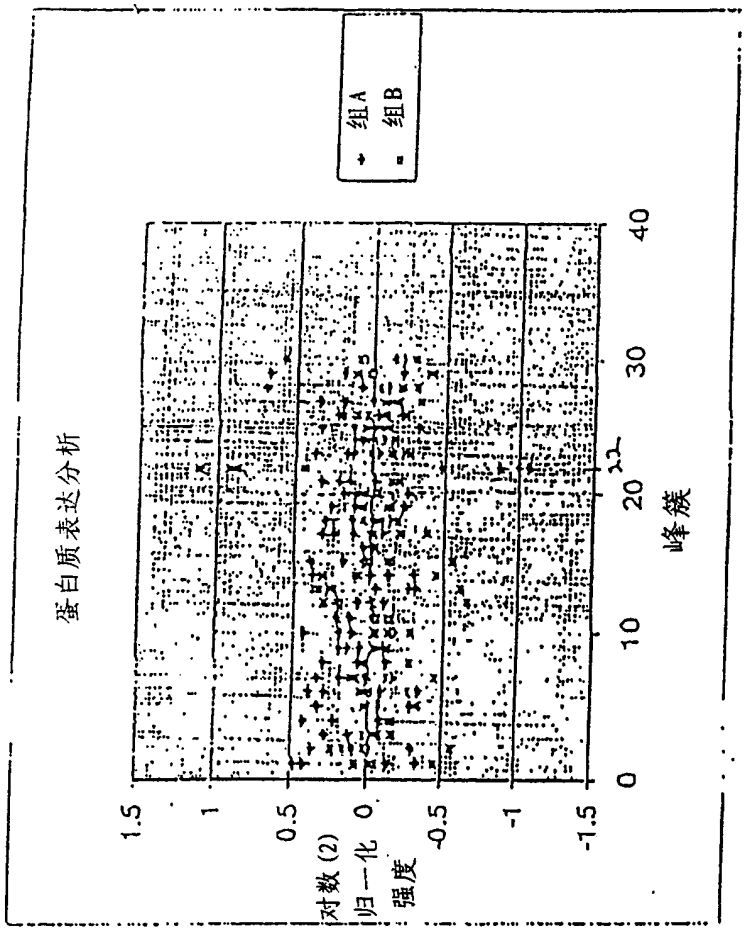


图 3

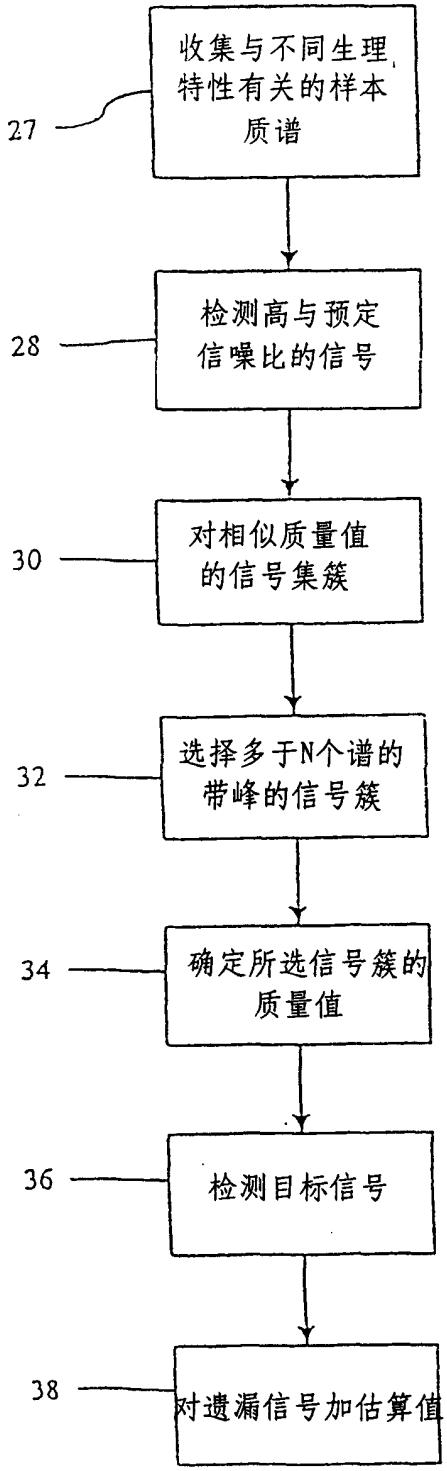

收集与不同生理特性有关的样本质谱
27
检测高与预定信噪比的信号
28
对相似质量值的信号集簇
30
选择多于N个谱的带峰的信号簇
32
确定所选信号簇的质量值
34
检测目标信号
36
对遗漏信号加估算值
38

图 4

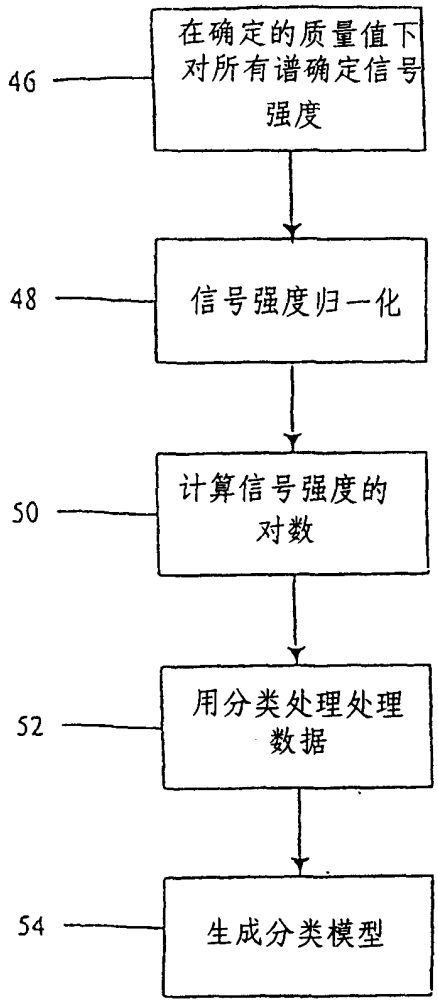

在确定的质量值下对所有谱确定信号强度
46
信号强度归一化
48
计算信号强度的对数
50
用分类处理处理数据
52
生成分类模型
54

图 5

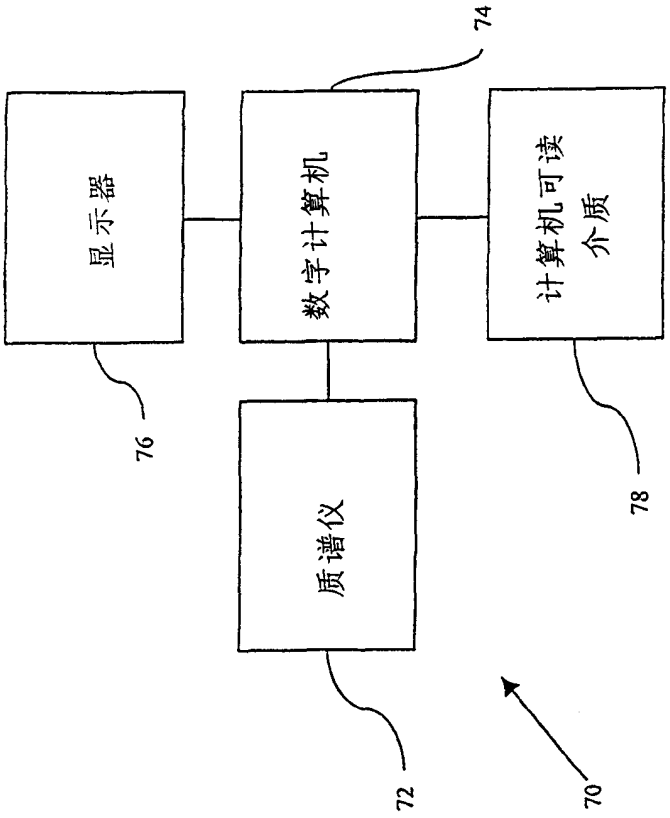


图 6

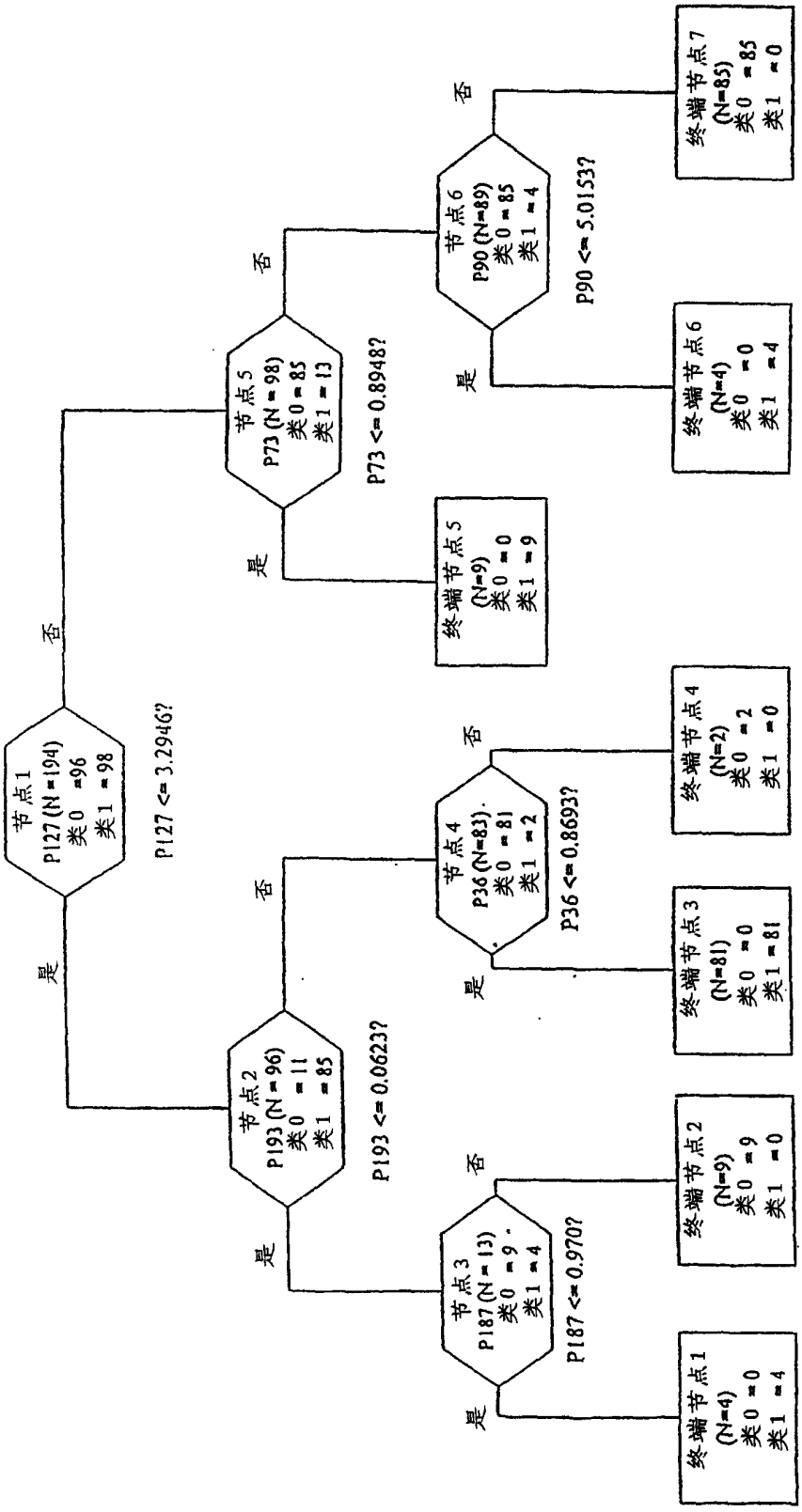
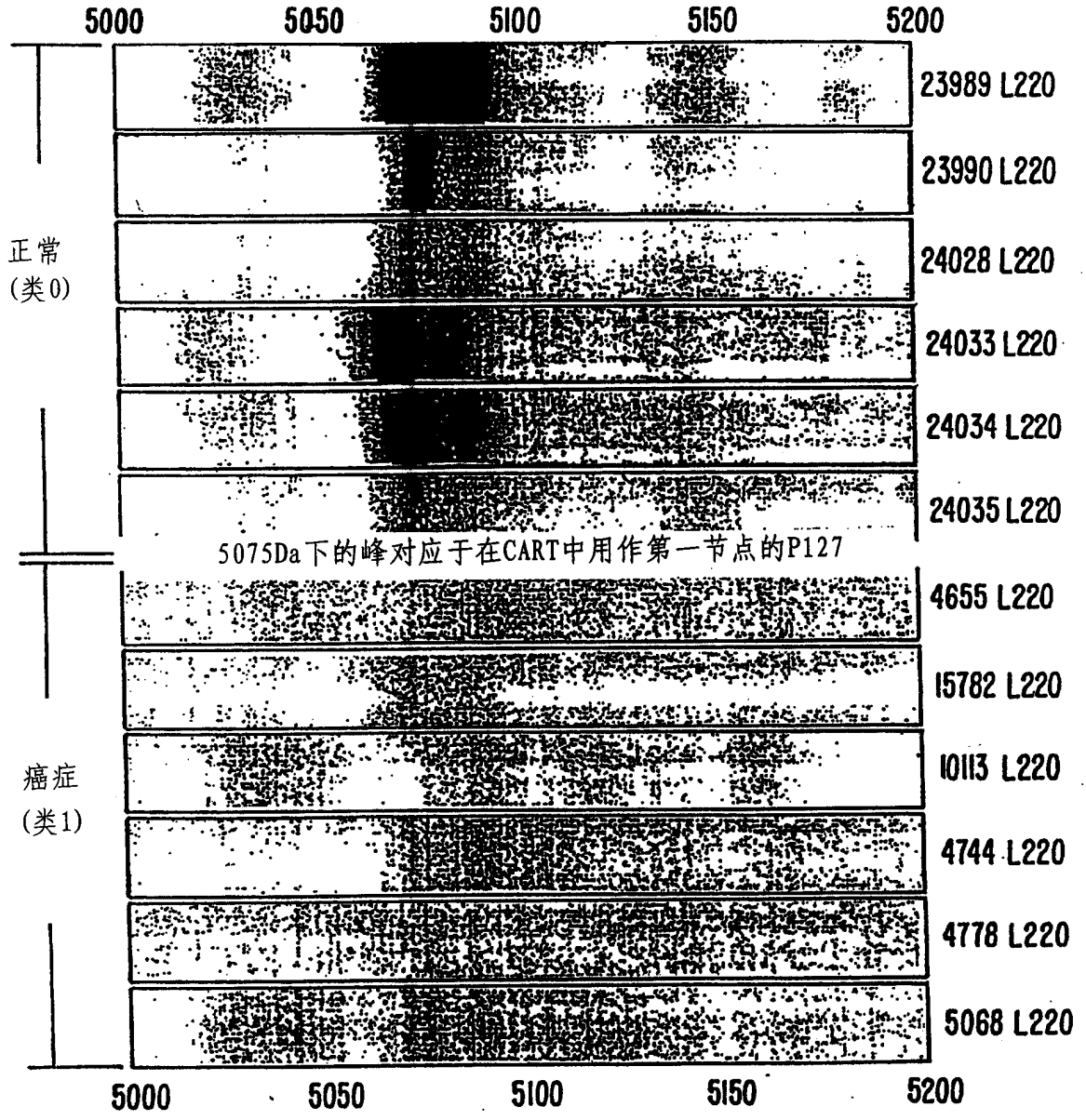

节点1
P127 (N=194)
类0 =96
类1 =98
P127 <= 3.2946?
是
否
节点2
P193 (N=96)
类0 =11
类1 =85
P193 <= 0.0623?
节点5
P73 (N=98)
类0 =85
类1 =13
P73 <= 0.8948?
节点3
P187 (N=13)
类0 =9
类1 =4
P187 <= 0.970?
终端节点1
(N=4)
类0 =0
类1 =4
终端节点2
(N=9)
类0 =9
类1 =0
节点4
P36 (N=83)
类0 =81
类1 =2
P36 <= 0.8693?
终端节点3
(N=81)
类0 =0
类1 =81
终端节点4
(N=2)
类0 =2
类1 =0
终端节点5
(N=9)
类0 =0
类1 =9
节点6
P90 (N=89)
类0 =85
类1 =4
P90 <= 5.0153?
终端节点6
(N=4)
类0 =0
类1 =4
终端节点7
(N=85)
类0 =85
类1 =0
类1 = 癌症
类0 = 正常

图 7

参量至要性

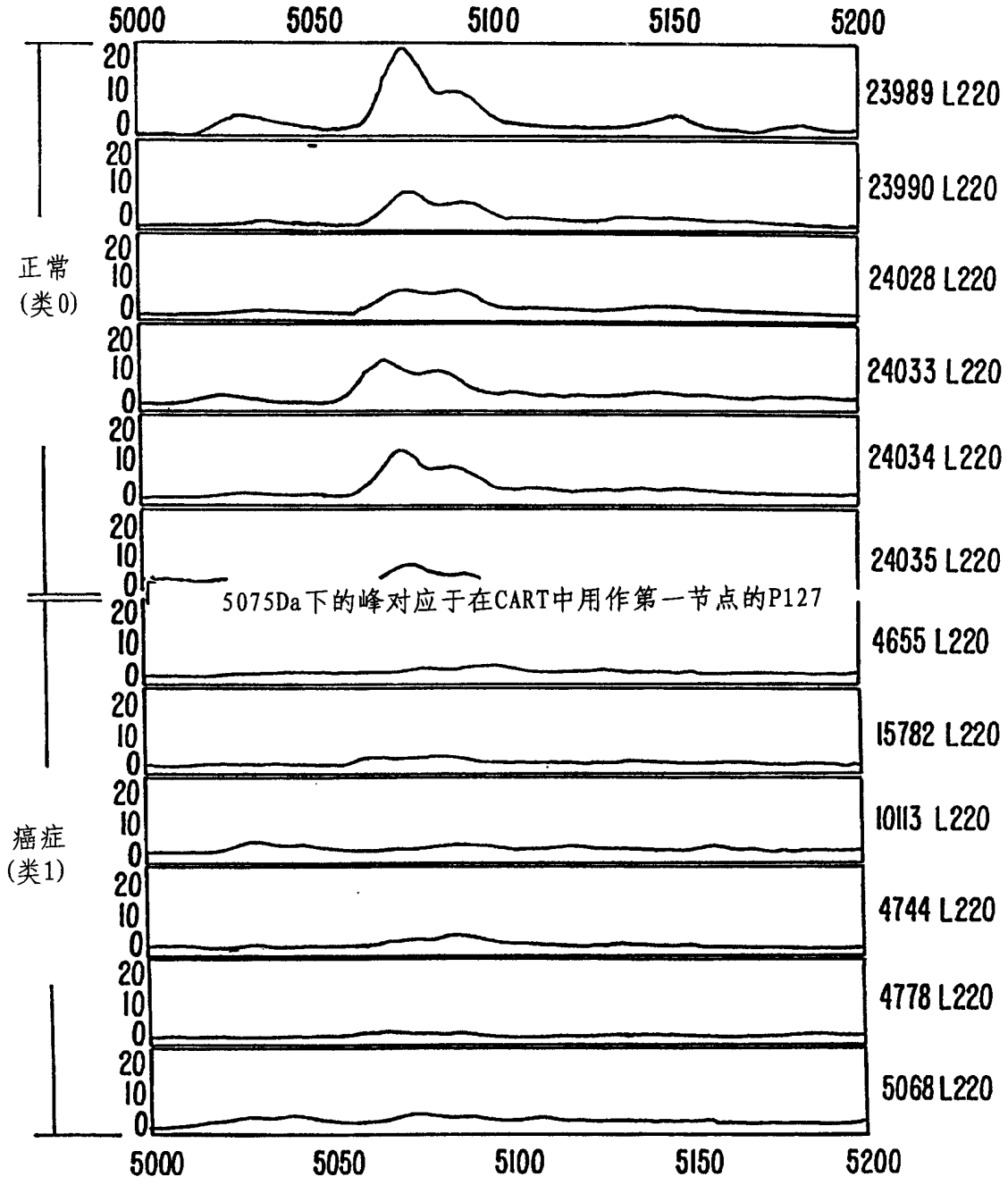
参量	分数	
P36	100.00	
P127	95.89	
P90	93.52	
P185	91.57	
P128	76.95	
P119	63.80	
P73	26.10	
P21	18.18	
P193	17.93	
P50	17.30	
P94	16.84	
P163	13.89	
P158	13.89	
P190	13.89	
P15	13.32	
P164	11.77	
P135	11.61	
P16	10.01	
P55	9.89	
P250	9.89	
P54	9.89	
P11	9.89	
P187	9.72	
P209	9.69	
P46	6.91	
P1	6.58	
P78	6.58	
P18	3.43	
P6	3.43	
P4	3.43	
P19	0.00	
P2	0.00	
P5	0.00	
P32	0.00	
P3	0.00	
P35	0.00	
P23	0.00	

图 8


5000
5050
5100
5150
5200
23989 L220
23990 L220
24028 L220
24033 L220
24034 L220
24035 L220
正常
(类0)
5075Da下的峰对应于在CART中用作第一节点的P127
4655 L220
15782 L220
10113 L220
4744 L220
4778 L220
5068 L220
癌症
(类1)
5000
5050
5100
5150
5200

P127凝胶

图 9



P127轨迹和光谱视图

图 10