(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 111148844 A
(43)申请公布日 2020.05.12

(21)申请号 201880063851.7

(22)申请日 2018.08.31

(30)优先权数据
62/553,676 2017.09.01 US

(85)PCT国际申请进入国家阶段日
2020.03.30

(86)PCT国际申请的申请数据
PCT/US2018/049256 2018.08.31

(87)PCT国际申请的公布数据
WO2019/046814 EN 2019.03.07

(71)申请人 韦恩生物科技股份公司
地址 美国加利福尼亚州

(72)发明人 L.M.A.·丹南-里欧
A.M.E.S.·卡拉斯科
C.R.·贝尔托西 C.B.·拉韦利亚

(74)专利代理机构 北京天达共和律师事务所 11798
代理人 龚建华

(51)Int.Cl.
C12Q 1/37(2006.01)
G01N 33/53(2006.01)
G01N 33/574(2006.01)
G01N 33/68(2006.01)

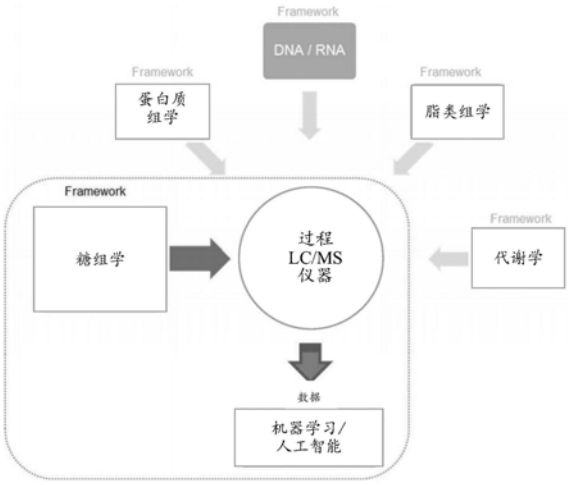
权利要求书1页 说明书10页 附图5页

(54)发明名称

鉴定和使用糖肽作为诊断和治疗监测的生物标记物

(57)摘要

本专利公开了一种利用蛋白质组学、肽组学、代谢学、糖蛋白组学、糖组学、质谱分析及机器学习,来鉴定各种疾病的新的生物标记物的方法。本专利还公开了可以作为各种疾病生物标记物的糖肽,如癌症和自身免疫性疾病。



CN 111148844 A

1.一种鉴定糖基化肽片段作为潜在生物标记物的方法,包括:

从多名受试者分离多种生物样品,将每种生物样品中的糖基化蛋白,用一种或多种蛋白酶进行酶解,以产生糖基化肽片段;

用液相色谱和质谱(LC-MS)法定量分析糖基化肽片段,以提供定量结果;

使用机器学习方法分析定量结果及受试者的分类,以选择可用于预测该分类的糖基化肽片段;和

测定糖基化肽片段的特性。

2.根据权利要求1所述的方法,所述受试者包括患有疾病或症状的受试者和未患有所述疾病或症状的受试者。

3.根据权利要求1或2所述的方法,所述受试者包括接受疾病或症状治疗的受试者和患有该疾病或症状但未接受治疗的受试者。

4.根据权利要求2或3所述的方法,所述疾病是癌症或自身免疫疾病。

5.根据权利要求4所述的方法,所述癌症选自乳腺癌、宫颈癌或卵巢癌。

6.根据权利要求5所述的方法,所述自身免疫疾病是HIV、原发性硬化性胆管炎、原发性胆汁性肝硬化或牛皮癣。

7.根据在先任一权利要求所述的方法,所述糖基化肽片段是N-糖基化的。

8.根据在先任一权利要求所述的方法,所述糖基化肽片段是O-糖基化的。

9.根据在先任一权利要求所述的方法,所述糖基化蛋白是下列中的一种或多种:α-1-酸糖蛋白、α-1-抗胰蛋白酶、α-1B-糖蛋白、α-2-HS-糖蛋白、α-2巨球蛋白、抗凝血酶-III、载脂蛋白B-100、载脂蛋白D、载脂蛋白F、β-2-糖蛋白1、铜蓝蛋白、胎球蛋白、纤维蛋白原、免疫球蛋白(Ig)A、IgG、IgM、触珠蛋白、血红蛋白、富含组氨酸的糖蛋白、激肽原1、血清转铁蛋白、转铁蛋白、玻连蛋白和锌-α-2-糖蛋白。

10.根据权利要求9所述的方法,所述糖基化蛋白是α-1-酸糖蛋白、免疫球蛋白(Ig)A、IgG或IgM。

11.根据在先任一权利要求所述的方法,所述糖基化肽片段的平均长度为约5至约50个氨基酸残基。

12.根据在先任一权利要求所述的方法,所述一种或多种蛋白酶至少包括两种蛋白酶。

13.根据在先任一权利要求所述的方法,所述质谱法包括多反应监测质谱法。

14.根据在先任一权利要求所述的方法,所述生物样品是身体组织、唾液、眼泪、痰、脊髓液、尿液、滑液、全血、血清或血浆。

15.根据权利要求14所述的方法,所述生物样品是全血、血清或血浆。

16.根据在先任一权利要求所述的方法,所述受试者是哺乳动物。

17.根据权利要求16所述的方法,所述受试者是人类。

18.根据在先任一权利要求所述的方法,其中所述机器学习方法是深度学习、神经网络、线性判别分析、二次判别分析、支持向量机、随机森林、最近邻以及它们的组合。

19.根据在先任一权利要求所述的方法,其中所述机器学习方法是深度学习、神经网络及它们的组合。

20.根据在先任一权利要求所述的方法,其中所述分析还包括基因组数据、蛋白质组学、代谢学、脂类组学数据及其组合。

鉴定和使用糖肽作为诊断和治疗监测的生物标记物

技术领域

[0001] 本专利总体上涉及多组学领域，尤其是糖组学和糖蛋白组学、高级仪器大数据、机器学习和人工智能领域，以鉴定用于疾病诊断和治疗监测的生物标记物。

背景技术

[0002] 蛋白糖基化和其他翻译后修饰在人类生长发育的各个方面起着重要的结构和功能作用。蛋白质糖基化缺陷会引发多种疾病。在疾病的早期阶段鉴定出发生改变的糖基化，为受疾病影响的受试者提供了进行早期检测、干预及获得更大存活几率的机会。目前，有一些方法可以鉴定早期癌症的生物标记物，并将某种类型的癌症与其他疾病区分开。这些方法包括采用质谱法(MS)的蛋白质组学、肽组学、代谢学、糖蛋白组学和糖组学。

[0003] 尽管蛋白质糖基化提供了有关癌症和其他疾病的有用信息，但是该方法的一个不足之处是，无法追溯到聚糖的初始蛋白质位点。为了获得有关癌症生物学和早期癌症检测的更多知识，重要的不仅是要鉴定聚糖，还要鉴定聚糖在蛋白质中的附着位点。由于多种因素，糖蛋白分析通常具有挑战性。例如，由于存在不同的糖苷键、支链和许多具有相同质量的单糖，肽中的单个聚糖成分可能包含大量的异构体结构。此外，由于具有相同肽链骨架的多种聚糖的存在，导致MS信号分裂为各种糖型，与未糖基化的肽相比，降低了其各自的丰度。因此，开发从串联MS数据识别聚糖及其糖肽的算法一直是一项挑战。因为它们具有不同的酶解效率，同时获取聚糖及其糖肽的混合片段也具有挑战性。

[0004] 因此，有必要提供一种分析特定位点的糖蛋白的方法，以获取蛋白糖基化模式相关的重要且详细的信息，相对于未病变的细胞、组织或生物体液，针对病变细胞、组织或生物体液中糖基化位点的异质性，提供精确定量信息。这种方法会涉及到鉴定疾病特别是癌症等的生物标记物。为节省鉴定新的生物标记物的时间，需要将特定位点的糖蛋白分析数据和深度学习、高级LC/MS仪器相结合，以鉴定和验证新的疾病靶标，如用于治疗癌症等疾病的基于聚糖的药物靶标。

发明内容

[0005] 本专利涉及鉴定各种疾病的生物标记物的方法。所述生物标记物是通过从生物样品中酶解糖基化蛋白获得的糖基化肽片段。鉴定生物标记物的方法需使用高级质谱技术，用于对糖基化肽片段进行精确的质量测量以及特定位点的糖基化分析。本专利的质谱方法的优势在于可单次分析来自生物样品的大量糖基化蛋白质。

[0006] 在一个实施例中，本专利提供了一种鉴定糖基化肽片段作为潜在生物标记物的方法，包括：

[0007] 从多名受试者分离多种生物样品，将每种生物样品中的糖基化蛋白，用一种或多种蛋白酶进行酶解，以产生糖基化肽片段；

[0008] 用液相色谱和质谱(LC-MS)法定量分析糖基化肽片段，以提供定量结果；

[0009] 使用机器学习方法分析定量结果及受试者的分类，以选择可用于预测该分类的糖

基化肽片段;和

[0010] 测定糖基化肽片段的特性。

[0011] 在另一实施例中,所述方法包括患有疾病或症状的受试者和未患有该疾病或症状的受试者。在另一实施例中,所述受试者包括接受疾病或症状治疗的受试者和患有该疾病或症状但未接受治疗的受试者。

[0012] 在另一实施例中,本发明的方法适用于通过分析受试者生物样品的糖基化肽片段可以检测到的任何疾病或症状。在一个实施例中,所述疾病是癌症。在另一实施例中,所述疾病是自身免疫疾病。在另一实施例中,本发明方法提供了O-糖基化或N-糖基化的糖基化肽片段。在另一实施例中,本发明方法提供了平均长度为5至50个氨基酸残基的糖基化肽片段。

[0013] 在另一实施例中,本专利方法使用的糖基化蛋白是下列中的一种或多种:α-1-酸糖蛋白、α-1-抗胰蛋白酶、α-1B-糖蛋白、α-2-HS-糖蛋白、α-2-巨球蛋白、抗凝血酶-III、载脂蛋白B-100、载脂蛋白D、载脂蛋白F、β-2-糖蛋白1、铜蓝蛋白、胎球蛋白、纤维蛋白原、免疫球蛋白(Ig)A、IgG、IgM、触珠蛋白、血红蛋白、富含组氨酸的糖蛋白、激肽原1、血清转铁蛋白、转铁蛋白、玻连蛋白和锌-α-2-糖蛋白。

[0014] 在另一实施例中,本专利方法包括至少使用两种蛋白酶使糖基化的蛋白质酶解。在另一实施例中,本专利方法包括使用多反应监测质谱法(MRM-MS)的LC-MS技术。

[0015] 在另一实施例中,本专利提供的鉴定糖基化的肽片段作为所述各种疾病的潜在生物标记物的方法中,所述生物样品是来自所述受试者的身体组织、唾液、眼泪、痰、脊髓液、尿液、滑液、全血、血清或血浆。在一个实施例中,所述受试者是哺乳动物。在另一实施例中,所述受试者是人类。

[0016] 在另一实施例中,本专利提供了一种鉴定糖基化肽片段作为潜在生物标记物的方法,包括:

[0017] 从多名受试者分离多种生物样品,将每种生物样品中的糖基化蛋白,用一种或多种蛋白酶进行酶解,以产生糖基化肽片段;

[0018] 用液相色谱和质谱(LC-MS)法定量分析糖基化肽片段,以提供定量结果;

[0019] 使用机器学习方法分析定量结果及受试者的分类,以选择可用于预测该分类的糖基化肽片段;和

[0020] 测定糖基化肽片段的特性,

[0021] 其中所述机器学习方法是深度学习、神经网络、线性判别分析、二次判别分析、支持向量机、随机森林(random forest)、最近邻(nearest neighbor)以及它们的组合。在另一实施例中,所述机器学习方法是深度学习、神经网络或其组合。

[0022] 在另一个实施例中,本专利提供的用于鉴定糖基化肽片段作为所述各种疾病的潜在生物标记物的方法中,所述分析还包括基因组数据、蛋白质组学、代谢学、脂类组学数据或其组合。

附图说明

[0023] 图1是糖组学、LC/MS和机器学习的整合示意图,其还可与蛋白质组学、基因组学、脂类组学、代谢学结合;

[0024] 图2示出了乳腺癌患者血浆样品相对于对照组的免疫球蛋白G(IgG)糖肽比值的变化；

[0025] 图3示出了原发性硬化性胆管炎(PSC)和原发性胆汁性肝硬化(PBC)患者的血浆样品中的IgG糖肽比值相对于健康供体的变化；

[0026] 图4示出了PSC和PBC患者的血浆样品相对于健康供体的IgG、IgA和IgM糖肽的单独判别分析数据；

[0027] 图5示出了PSC和PBC患者的血浆样品相对于健康供体的IgG、IgA和IgM糖肽的组合判别分析数据。

[0028] 详细说明

[0029] 定义

[0030] 在本说明书中，除非上下文另有说明，下列词语和短语通常具有以下含义。

[0031] 应注意的是，说明书及权利要求所使用的单数名词“一”、“一个/种”和“该”、“所述”包括复数，除非上下文另有明确说明。

[0032] 词语“生物样品”是指任何生物体液、细胞、组织、器官或其一部分，包括但不限于通过活检获得的组织切片，或放置在组织培养物中或适应了组织培养的细胞。它还包括但不限于唾液、眼泪、痰、汗液、粘液、粪便、胃液、腹腔液、羊水、囊肿液、腹膜液、脊髓液、尿液、滑液、全血、血清、血浆、胰液、母乳、肺灌洗液、骨髓等。

[0033] 词语“生物标记物”是指某一过程、事件或条件的独特的生物学或来自于生物学的指标。生物标记物还指示某种生物学状态，例如疾病或症状的存在，或疾病或症状的风险。它包括生物分子或生物分子的片段，它的变化或检测与特定的身体状态或健康状态有关。生物标记物的实例包括但不限于核苷酸、氨基酸、脂肪酸、类固醇、抗体、激素、类固醇、肽、蛋白质、碳水化合物等生物分子。其他的实例还包括糖基化肽片段、脂蛋白等。

[0034] 词语“包括”指所述组合物和方法包括所列举的方法，但不排除其他。

[0035] 词语“聚糖”指糖复合物如糖肽、糖蛋白、糖脂或蛋白聚糖中的碳水化合物部分。

[0036] 词语“糖型”是指附着有特定结构聚糖的蛋白质的独特一级、二级、三级和四级结构。

[0037] 词语“糖基化肽片段”是指通过一种或多种蛋白酶对糖基化肽进行酶解而得的糖基化肽(或糖肽)，其氨基酸序列与所述糖基化肽部分相同但不完全相同。

[0038] 词语“多反应监测质谱法(MRM-MS)”是指对生物样品中的蛋白质/肽进行的高灵敏度、高选择性的靶向定量方法。与传统质谱法不同，MRM-MS具有高度选择性(靶向性)，研究人员可以微调仪器以寻找特定目标肽/蛋白质片段。MRM的灵敏度更高、特异性更高、分析速度更快、且定量限更大，可用于检测特定的肽/蛋白质片段，例如潜在的生物标记物。MRM-MS包括三重四极杆(QQQ)质谱仪或四极杆飞行时间(qTOF)质谱仪。

[0039] 词语“蛋白酶”是指能够将蛋白质进行蛋白水解或分解成较小的多肽或氨基酸的酶。蛋白酶的实例包括丝氨酸蛋白酶、苏氨酸蛋白酶、半胱氨酸蛋白酶、天冬氨酸蛋白酶、谷氨酸蛋白酶、金属蛋白酶、天冬酰胺肽裂解酶及其组合。

[0040] 词语“受试者”是指哺乳动物。哺乳动物的非限制性实例包括人、非人灵长类、小鼠、大鼠、狗、猫、马或牛等。在代表动物病时、病前或症前状态的动物模型中，除人类以外的哺乳动物作为受试者比较有利。受试者可以是雄性或雌性。受试者可以是先前已被确定患

有某种疾病或症状，也可以是已经或正在接受某种疾病或症状的治疗干预。作为可供选择的方案，受试者也可以是先前未被诊断出患有疾病或症状。例如，受试者可以表现出一种或多种疾病或症状的危险因素，或者不表现出疾病危险因素，或者是对疾病或症状无表现。受试者也可以正患有疾病或症状或处于患有疾病或症状的风险中。

[0041] 词语“治疗”是指对受试者如哺乳动物的疾病或症状的任何处理，包括：(1)预防所述疾病或症状，即使临床症状不再发展；(2)抑制疾病或症状，即阻止或抑制临床症状的发展；和/或(3)缓解疾病或症状，即使临床症状消退。

[0042] 方法

[0043] 在一些实施例中，本专利涉及使用高级LC/MS仪器的糖蛋白组学，用于发现生物标记物、靶标及进行验证。本专利利用机器学习方法来处理分子数据。该分析还包括利用基因组数据、蛋白质组学、代谢学、脂类组学数据或其组合来发现各种疾病的新生物标记物。本专利方法的总体示意图如图1所示。

[0044] 本专利提供了用于特定位点的糖基化分析的方法，以更高的灵敏度和特异性鉴定新生物标记物。所述方法包括对糖基化肽的定量分析，从而有助于对不同位点的、与特定蛋白质结合的不同糖型的进行差异分析。所述方法提供了有关蛋白质数量和特定位点糖基化情况的信息，从而确认糖基化情况的变化是蛋白质糖基化的变化还是蛋白质浓度的变化导致的。与机器学习方法相结合的特定位点糖基化分析可用于鉴定多种疾病或症状的新的生物标记物。

[0045] 可以利用本专利的定量糖蛋白组学方法来发现各种疾病的生物标记物。该方法基于以下事实：在若干种疾病中，特定糖型升高而其他糖型下降，并且本专利中的LC/MS方法通过分析糖基化的显著变化来区分是否患有疾病。在一个实施例中，特定位点糖基化分析包括鉴定目标糖蛋白、修饰位点、修饰的具体情况、及测量每种修饰的相对丰度。在一些实施例中，所述疾病是癌症。在其他实施例中，所述疾病是自身免疫疾病。

[0046] 使用本专利的方法，将来自数千个受试者的生物样品数字化，生成大数据，经过对所述大数据进行深度地机器学习分析，从而发现各种疾病的新靶标。具体来说，通过深度学习来比较已知肽和未知肽的聚类，以及在疾病状态和对照状态下通过LC/MS所见的糖基化特征。糖基化肽的这种判别分析可以将疾病生物标记物鉴定出来。

[0047] 通过鉴定生物标记物及其对应特征例如它们的表达水平，可用来开发对疾病或症状的诊断测试方法，所述方法至少部分取决于对一种或多种所选的生物标记物的测定，及对结果与疾病或症状的关联关系的分析。所述方法还可用于选择一种或多种疗法，确定治疗方案或监测特定疾病或症状对某种疗法的反应。因此，本专利提供了预防、诊断、治疗、监测和预判某种疾病或症状的方法。在一些实施例中，所述方法可用于区分患有疾病或症状的受试者和健康受试者。在一些实施例中，所述方法可用于区分患有癌症的受试者和健康的受试者。一些实施例中，所述方法帮助癌症诊断或用于癌症监测。

[0048] 靶向及非靶向方法

[0049] 本专利的生物标记物发现方法同时采用靶向和/或非靶向方法。所述方法通常包括三个不同的阶段，即发现阶段、预验证阶段和验证阶段。

[0050] 发现阶段

[0051] 靶向方法包括在受试者的生物样品中鉴定和监测具有已知糖型的已知糖蛋白。

FDA批准过用于各种疾病的已知糖蛋白作为生物标记物，可以使用本专利的方法对所述生物标记物进行监测以确定受试者的分类。通常，生物标记物的糖基化变化具有肿瘤特异性，可用于鉴定疾病或疾病阶段的可能风险。靶向方法关注已知糖蛋白及其糖型；在研究开始阶段进行数据采集之前，就已经对所述糖蛋白及其糖型进行了化学表征，并且能够从生物学的角度说明其所具有的确定的生物学上的重要性。使用内部标准(internal standards)和确切的化学标准进行定量。

[0052] 具体来说，在靶向方法中，特定位点的糖基化分析是以生物样品的对照分析为基础进行的，所述生物样品来自于一些患有疾病或症状的受试者和同等数量的没有疾病或症状的对照受试者。首先，在生物样品中鉴定目标糖蛋白，例如疾病相关的糖蛋白或具有生物活性的糖蛋白。然后，使用LC/MS分析修饰位点、修饰性质、修饰特性和每种修饰的相对丰度，从而鉴定和定量肽片段。所述方法使用三重四极杆(QQQ)质谱仪定量糖基化肽片段，然后分析所述糖基化肽片段与受试者分类的关系。

[0053] 非靶向方法包括学习已知和未知肽片段的糖基化模式，以提供有关糖基化模式变化的更多信息，以便于于确定受试者的分类。非靶向方法是基于“上调或下调”糖蛋白的相对定量技术。具体来说，针对受试者的分类，监测其糖蛋白的上调或下调。例如，相对于与未患有某种疾病或病症的受试者，来监测患有该种疾病或症状的受试者的糖蛋白片段。在该方法中，在获取数据之前，并不知晓各个糖蛋白片段的化学特性。在一个实施例中，非靶向方法使用四极杆飞行时间(qTOF)质谱仪来分析糖基化肽片段。所述方法包括使用仪器准确测量样品中各成分的质量，事先不知道所述成分具体是什么。

[0054] 针对两组(疾病组与无疾病组)之间存在表达差异的候选对象，使用机器学习方法进行进一步选择和评估，以便根据重要临床特征的特征选择技术，进行分类预测。使用训练集来选择特征、构建模型，得到内部交叉验证并通过该验证来评估性能。通过测试集对得到的模型进行评估，测试集不参加模型的构建。采用Benjamin和Hochberg提出的错误发现率(FDR)方法来控制假阳性率。

[0055] 预验证阶段

[0056] 对发现阶段鉴定出来的候选生物标记物，在独立的生物样品测试集中进行测试，以确定候选生物标记物的性能，所述生物样品来自于一些患有某种疾病或症状的受试者及与其对应的未患有疾病或病症的对照组。选定的生物标记物及其级别，以及在发现阶段开发的模型的任何参数估计，都是建模的一部分，并在此独立的预验证阶段进行测试。根据候选生物标记物的信号，通过诊断测试将生物样品分为两组：患有疾病的样品和未患疾病的样品。然后根据阳性预测值、阴性预测值、特异性和灵敏度评估该测试的有效性。此外，使用受试者工作特征(ROC)曲线来评估诊断性能，即评估某种生物标记物或者多种生物标记物的组合在疾病或症状方面的统计学上的诊断测试效果更好。验证有效的各种生物标记物经过审核后，形成一组复合标记物。所述复合标记物通过加权多变量逻辑回归或其他分类算法构建。

[0057] 验证阶段

[0058] 对预验证阶段保留下来的候选生物标记物，通过来自多个受试者的独立盲选生物样品，进行独立验证。本阶段的目的是评估所选生物标记物的诊断精度。

[0059] 在一个实施例中，将所述生物标记物发现方法适用于生物样品，所述生物样品来

自患有癌症的受试者。在一些实施例中，每个组(即患癌组或未患癌组)分析的生物样品至少来自20个、40个、60个、80个或100个受试者。

[0060] 将靶向方法和/或非靶向方法与本专利所述的机器学习方法结合起来，为鉴定各种疾病的可能风险和/或早期检测提供了新的诊断方法。在一个实施例中，本专利提供了生物标记物的鉴定方法，所述方法基于靶向方法和非靶向方法与机器学习方法的结合。根据本专利方法鉴定出来的生物标记物，可用于诊断方法、预后评估方法、监测治疗结果、鉴定可能对特定疗法有反应的受试者、药物筛选等。

[0061] 在一个实施例中，本专利提供了一种鉴定糖基化肽片段作为潜在生物标记物的方法，包括：

[0062] 从多名受试者分离多种生物样品，将每种生物样品中的糖基化蛋白，用一种或多种蛋白酶进行酶解，以产生糖基化肽片段；

[0063] 用液相色谱-质谱联用法(LC-MS)定量糖基化肽片段，得到定量结果；

[0064] 使用机器学习方法分析定量结果及受试者的分类，选择用于预测分类的糖基化肽片段；

[0065] 测定糖基化肽片段的特性。

[0066] 在另一实施例中，本专利提供的如上所述的方法中，其中受试者包括患有疾病或症状的受试者和未患有所述疾病或症状的受试者。在另一实施例中，所述受试者包括接受疾病治疗的受试者和患有疾病但未接受治疗的受试者。

[0067] 本专利方法适用于通过分析来自受试者的生物样品的糖基化肽片段可以检测到的任何疾病或症状。在一个实施例中，所述疾病是癌症。在另一实施例中，所述癌症选自乳腺癌、宫颈癌或卵巢癌。在另一个实施例中，所述疾病是自身免疫疾病。在另一实施例中，所述自身免疫疾病是HIV、原发性硬化性胆管炎、原发性胆汁性肝硬化或牛皮癣。

[0068] 在另一实施例中，本专利提供的如上所述的方法中，所述糖基化蛋白是α-1-酸糖蛋白、α-1-抗胰蛋白酶、α-1B-糖蛋白、α-2-HS-糖蛋白、α-2-巨球蛋白、抗凝血酶III、载脂蛋白B-100、载脂蛋白D、载脂蛋白F、β-2-糖蛋白1、铜蓝蛋白、胎球蛋白、纤维蛋白原、免疫球蛋白(Ig)A、IgG、IgM、触珠蛋白、血红素、富含组氨酸的糖蛋白、激肽原-1、血清转铁蛋白、转铁蛋白、玻连蛋白和锌-α-2-糖蛋白中的一种或多种。在另一实施例中，所述糖基化蛋白是α-1-酸糖蛋白、免疫球蛋白(Ig)A、IgG或IgM中的一种或多种。

[0069] 在另一实施例中，本专利提供的如上所述的方法中，所述糖基化肽片段是N-糖基化的或O-糖基化的。

[0070] 在另一实施例中，本专利提供的如上所述的方法中，所述糖基化的肽片段的平均长度为约5至约50个氨基酸残基。在一些实施例中，糖基化的肽片段的平均长度为约5至约45、或约5至约40、或约5至约35、或约5至约30、或约5至约25、或约5至约20、或约5至约15、或约5至约10，或约10至约50、或约10至约45、或约10至约40、或约10至约35、或约10至约30、或约10至约25、或约10至约20、或约10至约15，或约15至约45、或约15至约40、或约15至约35、或约15至约30、或约15至约25、或约15至约20个氨基酸残基。在一个实施例中，所述糖基化肽片段的平均长度约为15个氨基酸残基。在另一实施例中，所述糖基化的肽片段的平均长度约为10个氨基酸残基。在另一实施例中，所述糖基化肽片段的平均长度约为5个氨基酸残基。

[0071] 在另一实施例中，本专利提供的如上所述的方法中，所述一种或多种蛋白酶包括用于酶解蛋白质的任何蛋白酶。在一个实施例中，所述蛋白酶是丝氨酸蛋白酶、苏氨酸蛋白酶、半胱氨酸蛋白酶、天冬氨酸蛋白酶、谷氨酸蛋白酶、金属蛋白酶、天冬酰胺肽裂解酶或其组合。蛋白酶的一些代表性实例包括但不限于胰蛋白酶、胰凝乳蛋白酶、内切蛋白酶、Asp-N、Arg-C、Glu-C、Lys-C、胃蛋白酶、嗜热菌素、酯酶、木瓜蛋白酶、蛋白酶K、枯草杆菌蛋白酶、梭菌蛋白酶、羧肽酶等。在另一实施例中，本专利提供的如上所述的方法中，所述一种或多种蛋白酶至少包括两种蛋白酶。

[0072] 本专利方法还有其他多种用途。例如，一种或多种生物标记物可用于区分疾病前状态与疾病状态，还是疾病状态与正常状态。还可以确定其他非疾病的特定健康状态。例如，可在不同时段测定生物标记物的变化：在患有疾病的受试者中，监测疾病的进展；在接受治疗的受试者中，监测治疗效果以及在接受治疗后的受试者中，监测可能出现的复发情况。同样，生物标记物定量水平也能决定疾病的治疗过程。例如，可以从正在接受疾病治疗方案的受试者采集生物样品。这种治疗方案可以包括但不限于，对诊断或鉴定为疾病或症状的受试者适用的运动方案、饮食补充、减肥、手术干预、装置植入以及治疗性或预防性的方案。

[0073] 此外，多种糖蛋白中糖肽比例的变化可能与某种疾病状态或未患某种疾病有关。例如，在生物样品中存在多种特定糖肽表明未患疾病，然而，如果生物样品中存在多种其他特定的糖肽则表明患有疾病。因此，各种糖肽谱或糖肽生物标记物组可能与疾病的各种状态有关。

[0074] 例2示出了乳腺癌患者的血浆样品相对于对照组的血浆样品中IgG1、IgG0和IgG2糖肽变化的定量结果。图2说明在该实验研究的乳腺癌的所有阶段中，糖肽A1和A2的水平与对照组相比有所升高，而糖肽A8、A9和A10的水平与对照组相比有所降低。由此说明糖肽A1、A2、A8、A9和A10是乳腺癌的潜在生物标记物。

[0075] 例3示出了PSC患者和PBC患者的血浆样品中IgG、IgM和IgA糖肽变化的定量结果。如图3所示，与健康供体相比，患有PBC和PSC的患者的血浆样品中的糖肽A升高了；而与健康供体相比，患有PBC和PSC的患者的血浆样品中的糖肽H、I和J降低了。因此，糖肽A、H、I和J是PBC和PSC的潜在生物标记物。此外，单独和组合判别分析结果分别如图4和图5所示，表明组合判别分析预测疾病状态的准确率为88％。

[0076] 在一些实施方案中，本专利提供的方法中，检测和分析的生物标记物的数量为1或大于1，例如2、3、4、5、6、7、8、9、10、11，12、13、14、15、16、17、18、19 20、21、22、23、24、25、26、27、28、30或更多。因此，本专利还提供了可用于疾病或症状诊断的生物标记物的组合。

[0077] 质谱法

[0078] 在一个实施例中，本专利提供的如本文所述的方法中，包括通过使用质谱仪定量分析糖基化肽片段。在一个实施例中，所述方法采用一种称为“多反应监测(MRM)”的技术。此技术通常与液相色谱法结合使用(LC/MRM-MS)，在单次LC/MRM-MS分析中可以定量数百个糖基化片段(及其母体蛋白)。本专利的高级质谱技术提供了有效的离子源、更高的分辨率、更快的分离速度和具有更高动态范围的检测器，在保留靶向检测的好处的同时，可以支持宽范围的非靶向检测。

[0079] 本专利所述的质谱法每次可以分析很多个糖基化蛋白质。例如，使用质谱仪每次

至少可以分析大于50、或至少大于60或至少大于70、或至少大于80、或至少大于90、或至少大于100、或至少大于110或120个以上的糖基化蛋白质。

[0080] 在一个实施例中，本专利的质谱法使用QQQ或qTOF质谱仪。在另一实施例中，本专利的质谱法提供的数据具有10ppm或更高、或5ppm或更高、或2ppm或更高、或1ppm或更高、或0.5ppm或更高、或0.2ppm或更高、或0.1ppm或更高的质量精度，分辨力在5,000或更高、或10,000或更高、或25,000或更高、或50,000或更高、或100,000或更高。

[0081] 生物样品

[0082] 本专利方法基于对生物样品的糖基化肽片段的定量分析。在一些实施例中，生物样品是先前收集的一种或多种临床样品，因此节省了用于鉴定新的生物标记物的资源和时间。在一些实施例中，生物样品来自先前的一项或多项研究，时间跨度为1至50年甚至更长。在一些实施例中，所述研究同时包括其他各种临床参数和先前已知的信息，例如受试者的年龄、身高、体重、种族、病史等。这种附加信息可用于将所述受试者与某种疾病或症状关联起来。在一些实施例中，生物样品是从所述受试者预先收集的一种或多种临床样品。

[0083] 在一个实施例中，本专利提供的如本文所述的方法中，从所述受试者分离出的生物样品是下列中的一种或多种：唾液、眼泪、痰、汗液、粘液、粪便、胃液、腹水、羊水、囊肿、液、腹膜液、脊髓液、尿液、滑液、全血、血清、血浆、胰液、母乳、肺灌洗液、骨髓。在另一实施例中，从所述受试者分离出的生物样品是身体组织、唾液、眼泪、痰、脊髓液、尿液、滑液、全血、血清或血浆。在另一实施例中，从所述受试者分离出的生物样品是全血、血清或血浆。在一些实施例中，所述受试者是哺乳动物。在其他实施例中，所述受试者是人类。

[0084] 疾病

[0085] 本专利方法适用于可以通过分析来自受试者的生物样品的糖基化肽片段检测到的任何疾病或症状。在一些实施例中，所述疾病或症状是癌症。在其他实施例中，所述癌症是急性淋巴细胞白血病(ALL)、急性髓细胞性白血病(AML)、肾上腺皮质癌、肛门癌、膀胱癌、血液癌、骨癌、脑瘤、乳腺癌、女性生殖系统癌、男性生殖系统癌、中枢神经系统淋巴瘤、宫颈癌、儿童横纹肌肉瘤、儿童肉瘤、慢性淋巴细胞性白血病(CLL)、慢性髓细胞性白血病(CML)、结肠和直肠癌、结肠癌、子宫内膜癌、子宫内膜肉瘤、食道癌、眼癌、胆囊癌、胃癌、胃肠道癌、毛细胞白血病、头颈癌、肝细胞癌、霍奇金病、下咽癌、卡波济肉瘤、肾癌、喉癌、白血病、肝癌、肺癌、恶性肿瘤纤维组织细胞瘤、恶性胸腺瘤、黑色素瘤、间皮瘤、多发性骨髓瘤、骨髓瘤、鼻腔和鼻旁窦癌、鼻咽癌、神经系统癌、神经母细胞瘤、非霍奇金淋巴瘤、口腔癌、口咽癌、骨肉瘤、卵巢癌、胰腺癌、甲状旁腺癌、阴茎癌、咽喉癌、垂体瘤、浆细胞瘤、原发性中枢神经系统淋巴瘤、前列腺癌、直肠癌、呼吸系统、视网膜母细胞瘤、唾液腺癌、皮肤癌、小肠癌、软组织肉瘤、胃癌、睾丸癌、甲状腺癌、泌尿系统癌、子宫肉瘤、阴道癌、血管系统、巨球蛋白血症、Wilms肿瘤等。在另一实施例中，所述癌症是乳腺癌、宫颈癌或卵巢癌。

[0086] 在另一实施例中，所述疾病是自身免疫疾病。在另一实施例中，所述自身免疫疾病是急性播散性脑脊髓炎、阿迪森氏病、无球蛋白血症、年龄相关性黄斑变性、斑秃、肌萎缩性侧索硬化、强直性脊柱炎、抗磷脂综合征、抗合成酶综合征、特应性过敏、特应性皮炎、自身免疫性自身免疫性疾病心肌病、自身免疫性肠病、自身免疫性溶血性贫血、自身免疫性肝炎、自身免疫性内耳疾病、自身免疫性淋巴增生性综合征、自身免疫性周围神经病、自身免疫性胰腺炎、自身免疫性多内分泌综合征、自身免疫性孕激素性皮炎、自身免疫性血小板减

少性紫癜、自身免疫性荨麻疹同心性硬化症、贝塞特氏病、伯杰氏病、比克斯塔夫氏脑炎、布劳综合征、大疱性天疱疮、癌症、卡斯曼病、腹腔疾病、恰加斯病、慢性炎性脱髓鞘性多发性神经病、慢性反复发作局灶性骨髓炎、慢性阻塞性肺疾病、变应性肉芽肿性血管炎、瘢痕性天疱疮、科干综合征、冷凝集素病、补体成分2缺乏症、接触性皮炎、颅动脉炎、CREST综合征、克罗恩病、库欣综合征、皮肤白细胞性血管炎、德戈氏病皮肤病、疱疹样皮炎、皮肌炎、I型糖尿病、弥漫性皮肤全身性硬化症、德勒综合症、药物性狼疮、盘状红斑狼疮、湿疹、子宫内膜异位、与炎性炎相关的关节炎、嗜酸性筋膜炎、嗜酸性肠炎、嗜酸性粒细胞增多症结节、胎儿成纤维细胞增多症、必要的混合性冷球蛋白血症、伊文氏综合症、进行性骨增生性纤维增生、纤维化性肺炎、胃炎、胃肠道天疱疮、肾小球肾炎、古德帕斯综合征、格雷夫斯病、吉兰-巴雷综合征、桥本脑病、桥本甲状腺炎、过敏性紫癜、艾滋病毒、妊娠天疱疮、化脓性汗腺炎、休斯史托文综合征、低球蛋白血症、特发性炎症性脱髓鞘疾病、特发性肺纤维化、特发性血小板减少性紫癜性慢性、IgA肾病、包涵体肌炎、关节炎、川崎病、兰伯特-伊顿肌无力综合征、白细胞碎裂性血管炎、扁平苔藓、扁平苔藓、线性IgA病、红斑狼疮、Majeed综合征、美尼尔氏病、显微多发性血管炎、混合性结缔组织病、硬斑病、多发性曼氏病、重症肌无力、肌炎、发作性睡病、视神经脊髓炎、神经性肌强直、眼球瘢痕性天疱疮、视神经支配性肌阵挛综合征、奥德甲状腺炎、回旋风湿病、与链球菌相关的小儿自身免疫性神经精神疾病、副交感神经变性、阵发性夜间血红蛋白尿、帕里·罗姆伯格综合征、帕森纳格·特纳综合征、帕尔斯平炎、寻常性天疱疮、恶性贫血、静脉性脑脊髓炎、POEMS综合征、结节性多发性动脉炎、多肌痛风湿病、多发性肌炎、原发性胆汁性胆汁性硬化症、原发性胆囊炎、银屑病关节炎、坏疽性脓皮病、纯红细胞发育不良、拉斯穆森氏脑炎、雷诺综合征、复发性软骨炎、赖特氏综合征、躁动性腿综合征、腹膜后纤维化、类风湿性关节炎、风湿热、结节病、精神分裂症、施密特综合征、施尼茨勒综合征、巩膜炎、硬皮病、血清病、干燥综合征、脊椎关节炎、僵人综合征、亚急性细菌性心内膜炎、Susac综合征、急性发热性嗜中性皮病、交感性眼炎、Takayasu动脉炎、颞动脉炎、血小板减少症、痛性眼肌麻痹综合征、横贯性脊髓炎、溃疡性结肠炎、未分化的结缔组织病、荨麻疹性血管炎、脉管炎、白癜风和韦格纳氏病等。在另一实施例中、所述自身免疫疾病是艾滋病、原发性硬化性胆管炎、原发性胆汁性肝硬化或牛皮癣。

[0087] 机器学习

[0088] 从数千个所述受试者中获得生物样品，将所述生物样品数字化，深度挖掘并验证先前未发现的标志物。在一些实施例中，生物样品是肿瘤样品或血液样品。使用LC/MS仪器将它们数字化，生成大量数据，经过深入的机器学习分析以发现各种疾病的新靶标。在一些实施例中，所述疾病是癌症或自身免疫疾病。

[0089] 在一个实施例中，本专利提供了一种鉴定糖基化肽片段作为潜在生物标记物的方法，包括：

[0090] 从多名受试者分离多种生物样品，将每种生物样品中的糖基化蛋白，用一种或多种蛋白酶进行酶解，以产生糖基化肽片段；

[0091] 用液相色谱和质谱(LC-MS)定量分析糖基化肽片段，提供定量结果；

[0092] 使用机器学习方法分析定量结果以及受试者的类别，选择可用于预测分类的糖基化肽片段；和

[0093] 测定糖基化肽片段的特性，

[0094] 所述机器学习方法是深度学习、神经网络、线性判别分析、二次判别分析、支持向量机、随机森林、最近邻或它们的组合。在一些实施例中，所述机器学习方法是深度学习、神经网络或其组合。所述分析还包括基因组数据、蛋白质组学、代谢学、脂类组学数据或其组合。图1是糖组学、LC/MS和机器学习的整合示意图，还可进一步与蛋白质组学、基因组、脂类组学和代谢学结合，以鉴定各种疾病的生物标记物。

实施例

[0095] 例1

[0096] 生物标记物的一般发现方法

[0097] 在靶向方法中，首先在生物样品中鉴定目标糖蛋白，然后使用LC/MS分析修饰位点、修饰性质、修饰特性和每种修饰的相对丰度，从而对肽片段进行鉴定和定量。所述方法使用三重四极杆(QQQ)质谱仪对糖基化肽片段进行定量，然后分析其与所述受试者分类的关系。

[0098] 在非靶向方法中，分析所有肽片段(已知和未知)的糖基化模式，以了解各种受试者中糖基化模式的变化信息。具体来说，针对受试者的分类，对糖蛋白的上调或下调与进行监控。例如，针对患有疾病或症状的受试者与未患有疾病或症状的受试者分别监测糖蛋白片段。所述方法使用四极杆飞行时间(qTOF)质谱仪分析糖基化肽片段。

[0099] 例2

[0100] IgG糖肽作为乳腺癌潜在生物标记物的定量分析

[0101] 分析不同阶段乳腺癌患者的血浆样品以及与所述患者年龄对应的对照组的IgG1、IgG0和IgG2糖肽，并比较它们的比例变化。具体来说，在QQQ质谱仪上对处于原位癌阶段的20个样品、EC1阶段的50个样品、EC2阶段的138个样品、EC3阶段的25个样品、EC4阶段的9个样品以及与73个与它们年龄相符的对照样品进行MRM定量分析。定量结果如图2所示，在本实验研究的乳腺癌的各个阶段中，与对照组相比，某些IgG1糖肽的水平升高，某些IgG1糖肽的水平降低。举例来说，在本实验研究的乳腺癌的各个阶段中，监测了IgG1糖肽A1—A11；与对照组相比，发现糖肽A1和A2的水平升高了，糖肽A8、A9和A10的水平降低了。因此，糖肽A1、A2、A8、A9和A10是乳腺癌的潜在生物标记物。

[0102] 例3

[0103] IgG糖肽作为PSC和PBC的潜在生物标记物的定量分析

[0104] 分析来自原发性硬化性胆管炎(PSC)患者、原发性胆汁性肝硬化(PBC)患者的血浆样品以及来自健康供体的血浆样品的IgG1和IgG2糖肽，并比较它们的糖肽比值的变化。具体而言，将100个PBC血浆样品、76个PSC血浆样品和49个健康供体的血浆样品在QQQ质谱仪上进行MRM定量分析。从图3的定量结果可以看出，在PBC和PSC患者的血浆样品中，与健康供体相比，某些IgG1糖肽升高，而某些IgG1糖肽降低。举例来说，与健康供体相比，PBC和PSC患者的糖肽A升高了，糖肽H、I和J降低了。因此，糖肽A、H、I和J是PBC和PSC的潜在生物标记。

[0105] 对PBC患者和PSC患者的血浆样品中的IgA和IgM糖蛋白进行了类似的分析。判别分析结果如图4所示，表明根据IgG、IgM和IgA的单项数据预测的准确率分别为59％、69％和74％。但是，如图5所示，将所有IgG、IgM和IgA的结果进行组合后，判别分析的准确率约为88％。

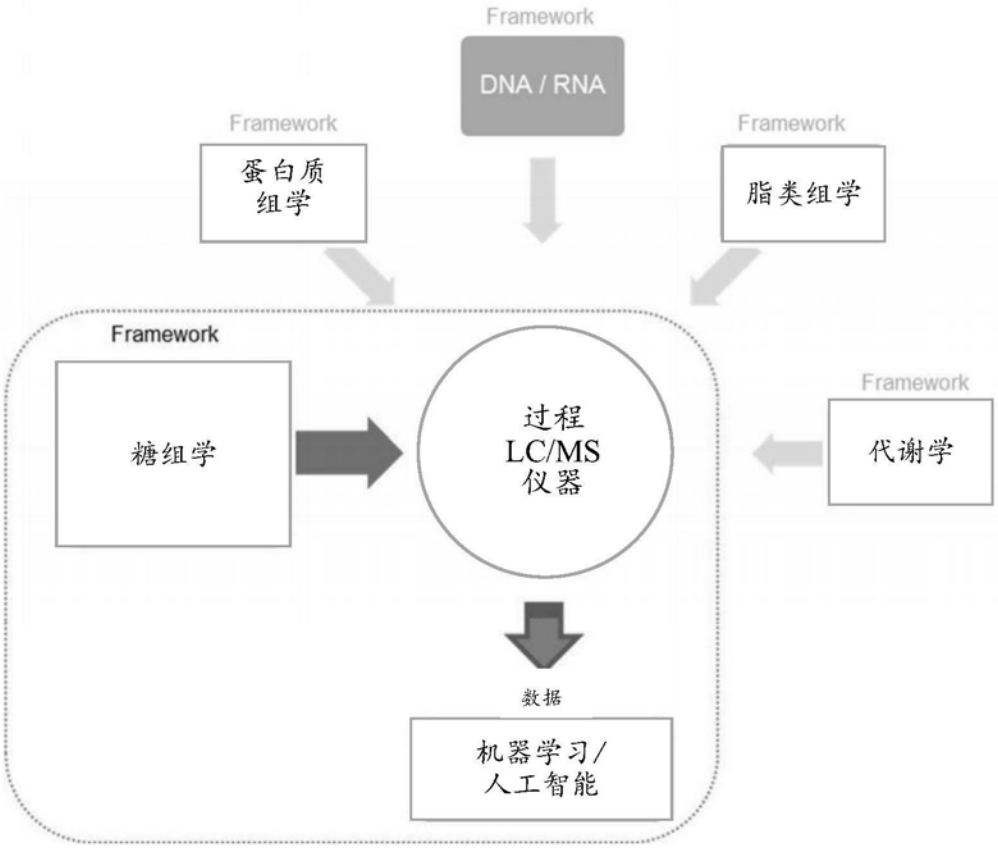

Framework
DNA / RNA
Framework
蛋白质
组学
Framework
脂类组学
Framework
糖组学
过程
LC/MS
仪器
Framework
代谢学
数据
机器学习/
人工智能

图1

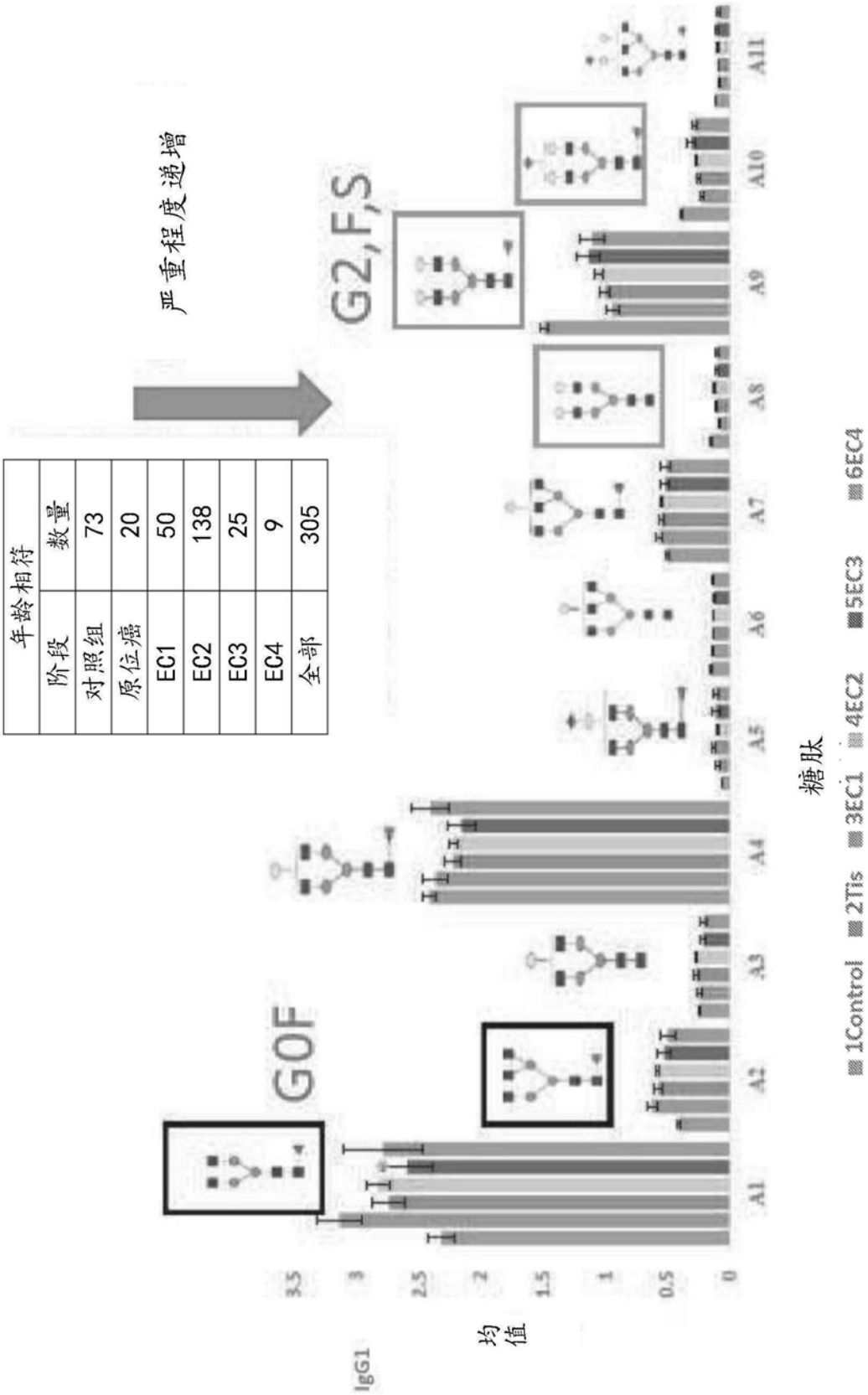

年龄相符
阶段
数量
对照组
73
原位癌
20
EC1
50
EC2
138
EC3
25
EC4
9
全部
305
严重程度递增
G0F
G2,F,S
IgG1
均值
糖肽
A1
A2
A3
A4
A5
A6
A7
A8
A9
A10
A11
1Control
2Tis
3EC1
4EC2
5EC3
6EC4

图2

	P值		
糖肽	HD$PBC	PBC$PSC	HD$PSC
A	0.00191	0.18391	0.00102
B	0.00013	0.01802	0.13285
C	0.35672	0.06573	0.00722
D	0.00004	0.51674	0.00223
E	0.00456	0.00033	0.70878
F	0.00153	0.89509	0.00921
G	0.60349	0.00009	0.01512
H	0.03423	0.90196	0.04887
I	0.00000	0.05185	0.01371
J	0.00001	0.05609	0.04079
K	0.00121	0.83644	0.00835

	P值		
糖肽	HD$PBC	PBC$PSC	HD$PSC
A	0.02350	0.25290	0.28903
B	0.00044	0.00001	0.74745
C	0.42242	0.00220	0.06059
D	0.22823	0.83085	0.24465
E	0.04634	0.53767	0.23297
F	0.53647	0.00803	0.08323
G	0.03045	0.00020	0.22319
H	0.14696	0.12476	0.98701
I	0.86852	0.21465	0.37531

红色表示重要

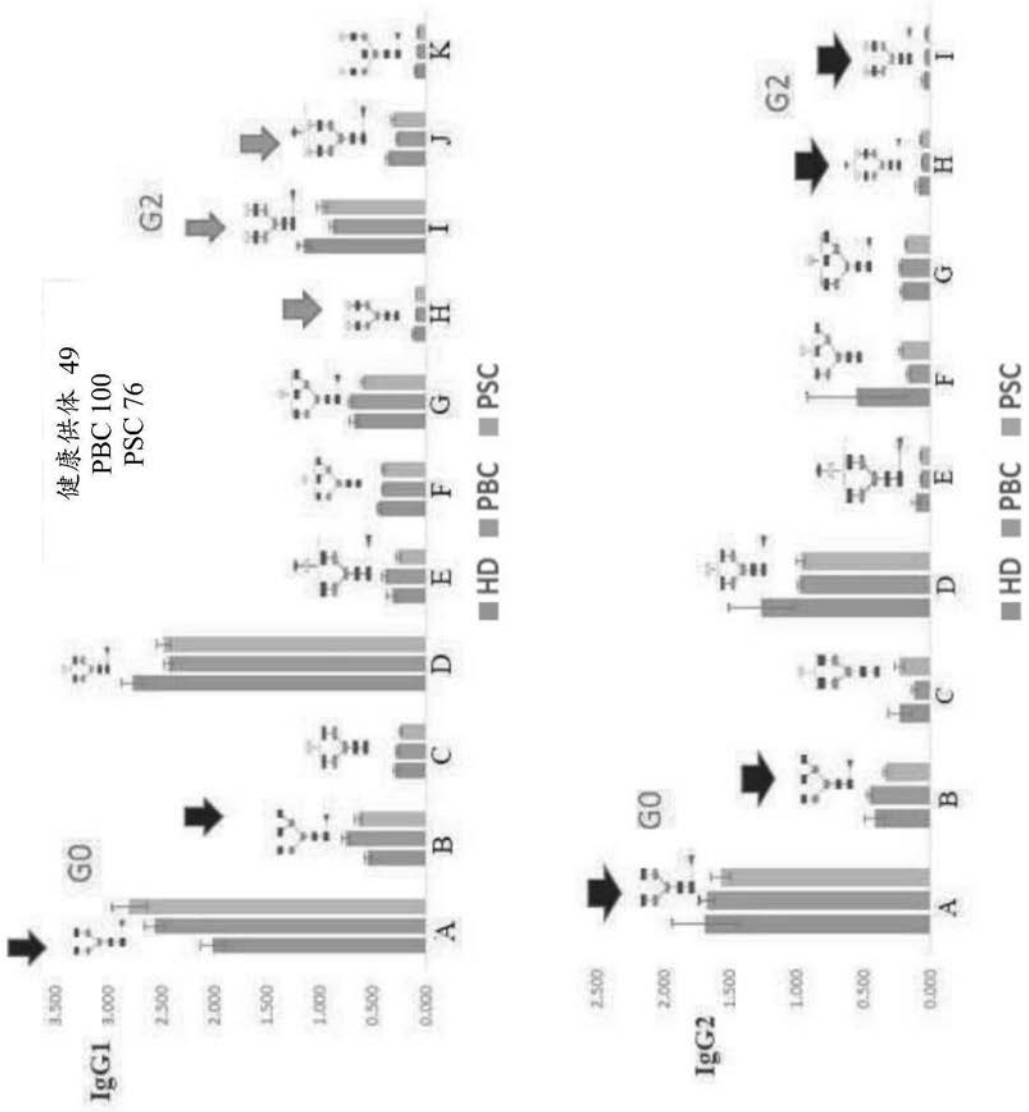
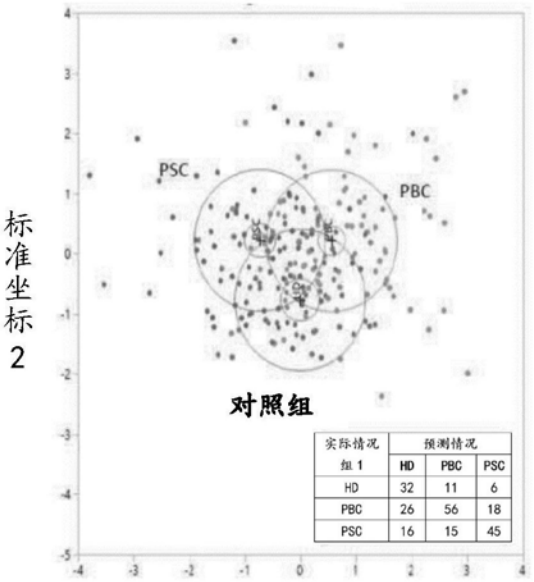


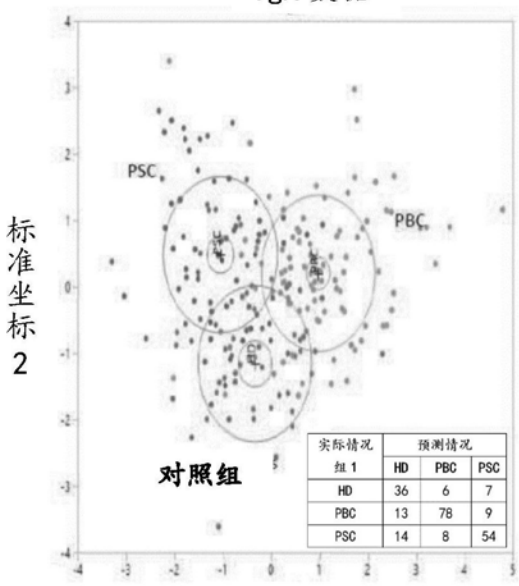
图3

IgG 数据



实际情况	预测情况		
组 1	HD	PBC	PSC
HD	32	11	6
PBC	26	56	18
PSC	16	15	45

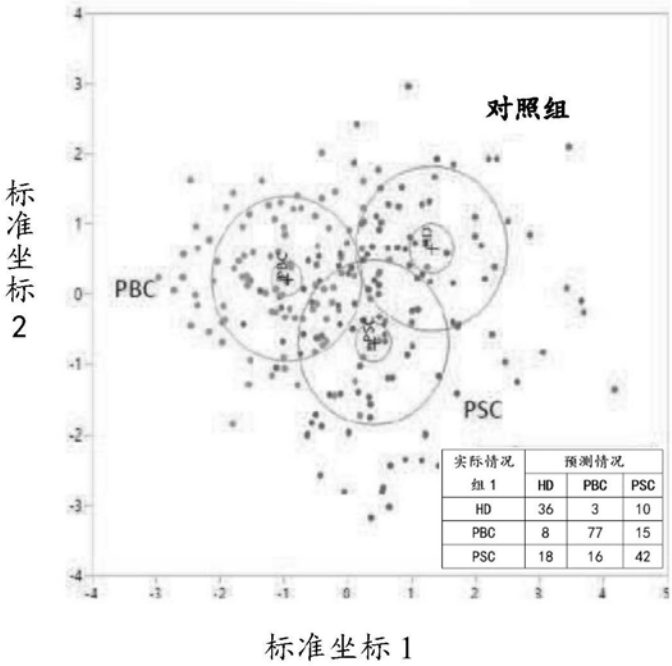
IgA 数据



实际情况	预测情况		
组 1	HD	PBC	PSC
HD	36	6	7
PBC	13	78	9
PSC	14	8	54

蛋白质	准确率%
IgG	59
IgM	69
IgA	74

IgM 数据



实际情况	预测情况		
组 1	HD	PBC	PSC
HD	36	3	10
PBC	8	77	15
PSC	18	16	42

图4

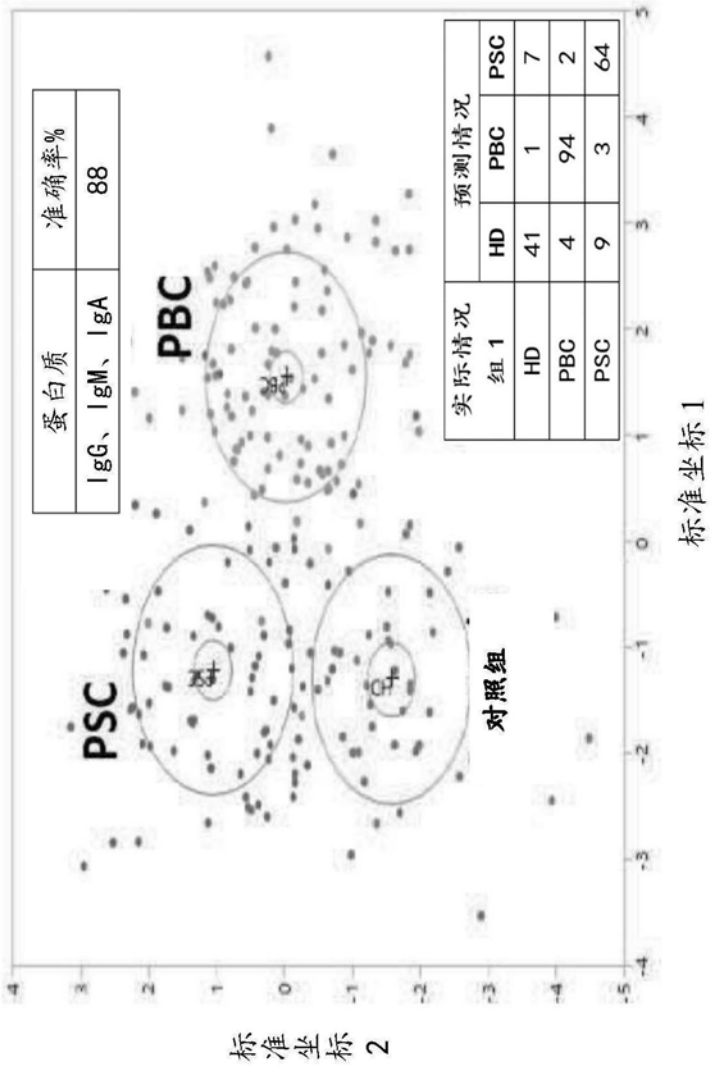

蛋白质
IgG、IgM、IgA
准确率%
88
PSC
PBC
对照组
实际情况
组 1
HD
PBC
PSC
预测情况
HD
PBC
PSC
41
1
7
4
94
2
9
3
64
标准坐标 2
标准坐标 1

图5

专利名称(译)	鉴定和使用糖肽作为诊断和治疗监测的生物标记物		
公开(公告)号	CN111148844A	公开(公告)日	2020-05-12
申请号	CN201880063851.7	申请日	2018-08-31
发明人	L.M.A.·丹南-里欧 A.M.E.S.·卡拉斯科 C.R.·贝尔托西 C.B.·拉韦利亚		
IPC分类号	C12Q1/37 G01N33/53 G01N33/574 G01N33/68		
CPC分类号	G01N33/564 G01N33/57415 G01N33/6842 G01N33/6848 G01N2400/00 G01N2800/08 G16B20/00 G16B40/10 G16B40/20 G01N33/6857 G16B40/00		
代理人(译)	龚建华		
优先权	62/553676 2017-09-01 US		
外部链接	Espacenet SIPO		

摘要(译)

本专利公开了一种利用蛋白质组学、肽组学、代谢学、糖蛋白组学、糖组学、质谱分析及机器学习，来鉴定各种疾病的新的生物标记物的方法。本专利还公开了可以作为各种疾病生物标记物的糖肽，如癌症和自身免疫性疾病。