



(12)发明专利申请

(10)申请公布号 CN 107064372 A

(43)申请公布日 2017.08.18

(21)申请号 201710313338.6

(22)申请日 2017.05.05

(71)申请人 北京骥骥生物技术有限公司

地址 100000 北京市海淀区树村厢黄旗甲  
100号都宜宾馆118

(72)发明人 王文军 陈显扬 萨日娜 马占青  
任素玲 段晓波

(74)专利代理机构 北京远大卓悦知识产权代理  
事务所(普通合伙) 11369

代理人 史霞

(51) Int. Cl.

G01N 30/02(2006.01)

G01N 30/86(2006.01)

G01N 30/72(2006.01)

G06F 19/12(2011.01)

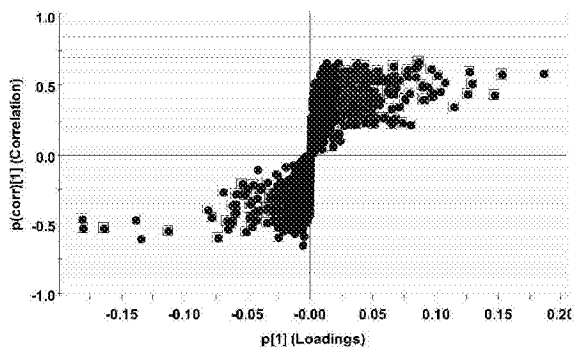
权利要求书1页 说明书8页 附图2页

(54)发明名称

利用脂质生物标志物预测甲状腺癌的方法

(57)摘要

本发明公开了一种利用脂质生物标志物预测甲状腺癌的方法,通过OPLS-DA模型,筛选出正常人组群和甲状腺癌组群之间的差异性化合物R1-R9,即与甲状腺癌相关的脂质生物标志物,并通过构建逻辑回归模型,得到根据这些脂质标志物预测甲状腺癌的方法,快捷方便,准确度达到87.2%。



1. 一种利用脂质生物标志物预测甲状腺癌的方法,其特征在于,主要包括以下步骤:

步骤1、筛选出正常人组群和甲状腺癌组群之间的差异性化合物R1-R9,分别为:

R1:PG (17:0/14:1);

R2:PS (0-20:0/18:1);

R3:PC (16:0/18:2);

R4:bacteriohopane-,32,33,34-triol-35-cyclitolguanine;

R5:PC (16:0/20:4);

R6:TG (16:0/16:1/20:2);

R7:PC (16:0/18:1);

R8:PS (0-18:0/17:0);

R9:PS (P-18:0/22:2);

步骤2、利用逻辑回归模型3进行计算,得到TC值,所述逻辑回归模型3的计算公式为:

$$TC = 1.8002 - 2.2815 * R1 - 2.3474 * R4 + 2.8573 * R9;$$

步骤3、根据所得TC值进行判断,TC=0为否;TC=1为是。

2. 如权利要求1所述的利用脂质生物标志物预测甲状腺癌的方法,其特征在于,步骤1中利用OPLS-DA模型对差异性化合物进行筛选,然后将VIP>1的排名前9位变量的数据提取出来即得R1-R9。

3. 如权利要求2所述的利用脂质生物标志物预测甲状腺癌的方法,其特征在于,步骤1中筛选的具体方法为:

步骤1.1将样品进行超高效液相色谱和质谱分析,得到脂质组学数据,将正常人组群和甲状腺癌组群分别计为CK及JC;

步骤1.2对脂质组学数据进行标准化操作,利用OPLS-DA模型对CK及JC进行S-plot分布得到S形曲线,并进行强制分组,计算影响CK及JC分组的变量重要性,即得VIP值;

步骤1.3按照VIP值大于5的标准得到15个化合物,并将该15个化合物作为与甲状腺癌相关度最高的差异性化合物;

步骤1.4将所得15个化合物按照VIP值大小从高到低排列,取前9位,即得步骤1中所述差异性化合物R1-R9。

4. 如权利要求1所述的利用脂质生物标志物预测甲状腺癌的方法,其特征在于,步骤2中TC值或者用逻辑回归模型2进行计算,所述逻辑回归模型2的计算公式为:

$$TC = 1.6361 - 12.5962 * R + 0.4081 * R2 - 0.962 * R3 - 1.7675 * R4 + 0.7317 * R5 - 7.3848 * R6 + 15.9658 * R7 + 0.494 * R8 + 2.5964 * R9。$$

5. 如权利要求1所述的利用脂质生物标志物预测甲状腺癌的方法,其特征在于,步骤2中TC值或者用逻辑回归模型1进行计算,所述逻辑回归模型1的计算公式为:

$$TC = 1.6054 - 13.4331 * R1 - 2.4503 * R4 + 0.9397 * R5 + 9.5919 * R7 + 3.3108 * R9。$$

## 利用脂质生物标志物预测甲状腺癌的方法

### 技术领域

[0001] 本发明涉及生物技术领域,尤其涉及一种利用脂质生物标志物预测甲状腺癌的方法,能够准确预测甲状腺癌。

### 背景技术

[0002] 甲状腺癌是常见的内分泌恶性肿瘤,传统上分为甲状腺分化性癌、甲状腺未分化癌和甲状腺髓样癌,分化性癌又分为乳头状癌和滤泡型腺癌,甲状腺未分化癌比较少见且预后极差,甲状腺髓样癌来源于甲状腺滤泡旁细胞(C细胞),属于神经内分泌肿瘤中的一种,在多发性2型内分泌瘤中约25%的患者伴有甲状腺髓样癌。近年来,我国甲状腺癌高发,已成为增长速度最快的恶性肿瘤,发病率10年增长了近5倍。甲状腺癌尤其好发于中青年女性,女性和男性比例为3:1,已成近20年来我国癌症谱中女性恶性肿瘤上升速度最快的肿瘤。如今在一、二线城市的女性群体中,该病发病率基本位列三甲,有的甚至已跃居榜首。

[0003] 目前,确诊甲状腺癌的主要手段主要是甲状腺细胞穿刺、术中病理冰冻及术后石蜡病理切片,甲状腺细胞穿刺虽然能达到确诊的意义,但由于阳性率低而限制了它在临床上的推广,所以临床实际确诊甲状腺癌的方法主要还是依靠术中冰冻或术后石蜡切片。除此之外,在甲状腺癌血清学方面,虽然对甲状腺癌肿瘤标志物的研究进展的很快,但至今还没有一种在临床上得到一致的认可,说明对甲状腺癌肿瘤标志物的研究仍然处于初步阶段,值得进一步发展。

[0004] 研究表明,脂质代谢与组织病变有非常直接的联系,也包括甲状腺机能紊乱。目前,通过核磁共振,MALDI/MS或者GC/MS,已经发现少许与甲状腺癌相关的脂质。其中,对脂肪酸的研究表明,良性甲状腺瘤的患者C14:0,C16:1n7,C18:1n9,C20:1n9,C18:3n3显著降低,C16:0,C20:3n6,C20:4n6,C22:6n3显著升高;恶性肿瘤患者表现出C14:0,C16:0,C18:3n3升高,而C20:3n6降低的趋势。另外,也有研究利用数学模型结合脂质组学,开始寻找肿瘤标志物。他们利用PLS-DA模型,发现了长醇,胆固醇,胆碱以及27种脂肪酸,在甲状腺癌良性和恶性人群中,体现明显的差异。

[0005] 但是,目前为止,还没有出现能够利用这些标志物准确预测甲状腺癌的预测模型,尤其是针对中国人群的甲状腺癌预测模型。

### 发明内容

[0006] 本发明的一个目的是解决至少上述问题,并提供至少后面将说明的优点。

[0007] 本发明还有一个目的是提供一种利用脂质生物标志物预测甲状腺癌的方法,能够快速、准确、高效地预测甲状腺癌。

[0008] 本发明还有一个目的是提供一种针对中国人群的利用脂质生物标志物预测甲状腺癌的方法。

[0009] 本发明还有一个目的是提供一组针对中国人群的与甲状腺癌密切相关的化合物。

[0010] 为了实现根据本发明的这些目的和其它优点,提供了以下技术方案:

- [0011] 一种利用脂质生物标志物预测甲状腺癌的方法,其中,主要包括以下步骤:
- [0012] 步骤1、筛选出正常人组群和甲状腺癌组群之间的差异性化合物R1-R9,分别为:
- [0013] R1:PG (17:0/14:1);
- [0014] R2:PS (0-20:0/18:1);
- [0015] R3:PC (16:0/18:2);
- [0016] R4:bacteriohopane-,32,33,34-triol-35-cyclitolguanidine;
- [0017] R5:PC (16:0/20:4);
- [0018] R6:TG (16:0/16:1/20:2);
- [0019] R7:PC (16:0/18:1);
- [0020] R8:PS (0-18:0/17:0);
- [0021] R9:PS (P-18:0/22:2);
- [0022] 步骤2、利用逻辑回归模型3进行计算,得到TC值,所述逻辑回归模型3的计算公式为:
- [0023]  $TC = 1.8002 - 2.2815 * R1 - 2.3474 * R4 + 2.8573 * R9$ ;
- [0024] 步骤3、根据所得TC值进行判断,TC=0为否;TC=1为是。
- [0025] 优选的是,步骤1中利用OPLS-DA模型对差异性化合物进行筛选,然后将VIP>1的排名前9位变量的数据提取出来即得R1-R9。
- [0026] 优选的是,步骤1中筛选的具体方法为:
- [0027] 步骤1.1将样品进行超高效液相色谱和质谱分析,得到脂质组学数据,将正常人组群和甲状腺癌组群分别计为CK及JC;
- [0028] 步骤1.2对脂质组学数据进行标准化操作,利用OPLS-DA模型对CK及JC进行S-plot分布得到S形曲线,并进行强制分组,计算影响CK及JC分组的变量重要性,即得VIP值;
- [0029] 步骤1.3按照VIP值大于5的标准得到15个化合物,并将该15个化合物作为与甲状腺癌相关度最高的差异性化合物;
- [0030] 步骤1.4将所得15个化合物按照VIP值大小从高到低排列,取前9位,即得步骤1中所述差异性化合物R1-R9。
- [0031] 优选的是,步骤2中TC值或者用逻辑回归模型2进行计算,所述逻辑回归模型2的计算公式为:
- [0032]  $TC = 1.6361 - 12.5962 * R + 0.4081 * R2 - 0.962 * R3 - 1.7675 * R4 + 0.7317 * R5 - 7.3848 * R6 + 15.9658 * R7 + 0.494 * R8 + 2.5964 * R9$ 。
- [0033] 优选的是,步骤2中TC值或者用逻辑回归模型1进行计算,所述逻辑回归模型1的计算公式为:
- [0034]  $TC = 1.6054 - 13.4331 * R1 - 2.4503 * R4 + 0.9397 * R5 + 9.5919 * R7 + 3.3108 * R9$ 。
- [0035] 本发明至少包括以下有益效果:
- [0036] 本发明通过筛选,首次发现与甲状腺癌相关的一组化合物,即与甲状腺癌相关的脂质生物标志物,并通过构建逻辑回归模型,得到根据这些脂质标志物预测甲状腺癌的方法,快捷方便,准确度高。通过AIC值初步判断,并进行ROC曲线绘制,AUC值达到0.872。
- [0037] 本发明的其它优点、目标和特征将部分通过下面的说明体现,部分还将通过对本发明的研究和实践而为本领域的技术人员所理解。

## 附图说明

- [0038] 图1为本发明中所述的OPLS-DA模型的S-plot分布图；  
[0039] 图2为本发明中利用OPLS-DA模型对CK和JC进行强制分组的结果；  
[0040] 图3为本发明中用来考察筛选出的化合物的火山图；  
[0041] 图4为本发明中ROC曲线图。

## 具体实施方式

[0042] 下面结合附图对本发明做进一步的详细说明,以令本领域技术人员参照说明书文字能够据以实施。

[0043] 应当理解,本文所使用的诸如“具有”、“包含”以及“包括”术语并不配出一个或多个其它元件或其组合的存在或添加。

[0044] 一种利用脂质生物标志物预测甲状腺癌的方法,其中,主要包括以下步骤:

[0045] 步骤1、筛选出正常人组群和甲状腺癌组群之间的差异性化合物R1-R9,分别为:

[0046] R1:PG (17:0/14:1);

[0047] R2:PS (0-20:0/18:1);

[0048] R3:PC (16:0/18:2);

[0049] R4:bacteriohopane-,32,33,34-triol-35-cyclitolguanine;

[0050] R5:PC (16:0/20:4);

[0051] R6:TG (16:0/16:1/20:2);

[0052] R7:PC (16:0/18:1);

[0053] R8:PS (0-18:0/17:0);

[0054] R9:PS (P-18:0/22:2)。

[0055] 步骤2、利用逻辑回归模型1进行计算,得到TC值,所述逻辑回归模型3的计算公式为:

[0056]  $TC = 1.8002 - 2.2815 * R1 - 2.3474 * R4 + 2.8573 * R9;$

[0057] 经过ROC曲线绘制,模型3的AUC值为0.872,准确度高达87.2%。

[0058] 步骤3、根据所得TC值进行判断,TC=0为否;TC=1为是。R1-R9均为检测样品中该脂质的含量值。

[0059] 步骤1中利用OPLS-DA模型对差异性化合物进行筛选,然后将VIP>1的排名前9位变量的数据提取出来即得R1-R9。

[0060] 步骤1中筛选的具体方法为:

[0061] 步骤1.1将样品进行超高效液相色谱和质谱分析,得到脂质组学数据,将正常人组群和甲状腺癌组群分别计为CK及JC。

[0062] 步骤1.2对脂质组学数据进行标准化操作,利用OPLS-DA模型对CK及JC进行S-plot分布得到S形曲线,并进行强制分组,计算影响CK及JC分组的变量重要性,即得VIP值。

[0063] 步骤1.3按照VIP值大于5的标准得到15个化合物,并将该15个化合物作为与甲状腺癌相关度最高的差异性化合物。

[0064] 步骤1.4将所得15个化合物按照VIP值大小从高到低排列,取前9位,即得步骤1中

所述差异性化合物R1-R9。

[0065] 步骤2中TC值或者用逻辑回归模型2进行计算,所述逻辑回归模型2的计算公式为:  
 $TC = 1.6361 - 12.5962 * R + 0.4081 * R^2 - 0.962 * R^3 - 1.7675 * R^4 + 0.7317 * R^5 - 7.3848 * R^6 + 15.9658 * R^7 + 0.494 * R^8 + 2.5964 * R^9$ 。

[0066] 经过ROC曲线绘制,模型2的AUC值为0.864,准确度为86.4%。

[0067] 步骤2中TC值或者用逻辑回归模型1进行计算,所述逻辑回归模型1的计算公式为:

[0068]  $TC = 1.6054 - 13.4331 * R_1 - 2.4503 * R_4 + 0.9397 * R_5 + 9.5919 * R_7 + 3.3108 * R_9$ 。

[0069] 经过ROC曲线绘制,模型1的AUC值为0.866,准确度为86.6%。

[0070] 实施例1

[0071] 材料和方法

[0072] 1.实验对象(均选自中国人):16位正常人,64位甲状腺癌患者,抽取静脉血5ml。准确量取100 $\mu$ L的血液,加入0.9mL的提取液(100%异丙醇),转入2mL离心管(必须是进口离心管,塑料不易溶于有机溶剂中;Axygen品牌)中,漩涡振荡10s以上,超声10min,然后在-20度冰箱中冷冻1小时,取出后在室温下漩涡振荡,用冷冻离心机10000rpm离心10min,然后去上清液1mL并过0.22 $\mu$ m的有机相滤膜到玻璃进样品种,保存在冰箱中待测。

[0073] 2.主要仪器

[0074] 2.1.1冷冻离心机:型号D3024R,Scilogex公司,美国

[0075] 2.1.2漩涡振荡器:型号MX-S,Scilogex公司,美国

[0076] 2.1.3高分辨质谱仪:ESI-QTOF/MS;型号:Xevo G2-S Q-TOF;厂家:Waters

[0077] 2.1.4超高效液相色谱:UPLC;型号:ACQUITY UPLC I-Class系统;厂家:Waters

[0078] 2.1.4数据采集软件:MassLynx4.1;厂家:Waters

[0079] 2.1.5分析鉴定软件:Progenesis QI;厂家:Waters

[0080] 2.1.6作图软件:EZinfo;HemI;Simca-P

[0081] 3.主要试剂

[0082] 甲醇、乙腈、甲酸、甲酸铵、亮氨酸脑啡肽、甲酸钠。厂家均为Fisher。

[0083] 4.实验设置

[0084] 以提取液(100%异丙醇)作为空白对照(Blank)样品;从每个样品的进样品种取出100 $\mu$ L混合到新的进样品种作为质控(QC)样品;正式样品按照每组样品间隔进样,例如先是Blank1,然后是QC1,然后是W1,然后是W2这个顺序,然后再进行下一轮进样。

[0085] 液相方法

[0086] 色谱柱:ACQUITY UPLC CSH C18 Column, 130 $\text{\AA}$ , 1.7 $\mu$ m, 1mm X 50mm, 1/pkg [186005292];

[0087] 柱温:55度

[0088] 流速:0.4mL/min

[0089] 流动相:A:ACN/H<sub>2</sub>O(60%/40%),含有10mM甲酸铵和0.1%甲酸

[0090] B:IPA/ACN(90%/10%),含有10mM甲酸铵和0.1%甲酸

[0091] (注:ACN为乙腈,IPA为异丙醇)

[0092] 进样体积:0.2 $\mu$ L

[0093] 洗脱程序:

	Time (min)	% A	% B	Curve
	Initial	60.0	40.0	Initial
[0094]	2.0	57.0	43.0	6 (表示线性增加)
	2.1	50.0	50.0	1 (表示快速增加)
	12.0	46.0	54.0	6
	12.1	30.0	70.0	1
[0095]	18.0	1.0	99.0	6
	18.1	60.0	40.0	6
	20.0	60.0	40.0	1

[0096] 质谱方法

[0097] 数据采集方式:MS<sub>e</sub>;分子量扫描范围:50-1500m/z;分辨率模式(轮廓图)。

[0098] 正负离子模式各采集一次。

[0099] 离子源:电喷雾电离源(ESI)

[0100] 毛细管电压:3KV

[0101] 锥孔电压为:25V

[0102] 碰撞能:15-60V

[0103] 源温度:120度

[0104] 脱溶剂温度:500度

[0105] 锥孔气体速度:50L/h

[0106] 脱溶剂气体速度:500L/h

[0107] 扫描时间:0.2s

[0108] 使用亮氨酸脑啡肽(m/z 556.2771,正离子;554.2615,负离子)进行实时校正。使用甲酸钠进行校正。

[0109] 脂质组学数据分析

[0110] Progenesis QI软件(Waters,Massachusetts,USA)用于结果分析,抽提非靶向脂质分子的特征峰,进行比对和筛选。同时,用QC(quantifying control)和blank(空白)来筛选背景数据。最终的数据,导入Ezinfo 3.0,并进行Principal Component Analysis(PCA)分析,Orthogonal signal correction Partial Least Square Discrimination Analysis(OPLS-DA)建模,Variable Importance in Projection(VIP)的计算,同时得到火山图(coefficients vs.VIP spots),如图3所示。其中,逻辑回归模型以及ROC曲线(如图4)通过R语言进行建设和绘制。

[0111] 结果描述

[0112] 正常人群与甲状腺癌人群脂质差异物质的鉴定。

[0113] 我们首先建立OPLS-DA模型,对正常人(CK)和甲状腺癌(JC)组群进行分类,并研究造成他们出现差异的原因。我们看到在OPLS-DA模型中,利用相关性(correlation)和协方差(covariance)的p值作出的S-plot形成了非常好的S形曲线,如图1所示。利用OPLS-DA将

两组数据进行强制分组,结果如图2所示。计算影响CK和JC分组的变量重要性,即VIP (Variable Importance in Projection) 值。一共筛选到308个化合物,他们的VIP值大于1;我们挑选VIP>5,贡献率大的15个化合物,如表1所示。

[0114] 表1. 变量重要性投影

[0115]

Primary ID	Primary ID	VIP
PG(17:0/14:1)	磷脂酰甘油 (17:0/14:1)	8.43028
PS(O-20:0/18:1)	烷基磷脂酰丝氨酸 (O-20:0/18:1)	7.35583
PC(16:0/18:2)	磷脂酰胆碱 (16:0/18:2)	7.35516
bacteriohopane-,32,33,34-triol-35-cyclitolg uanine	细菌何帕烷三醇环多醇鸟 嘌呤	6.95001
PC(16:0/20:4)	磷脂酰胆碱 (16:0/20:4)	6.78734
TG(16:0/16:1/20:2)	三酰甘油 (16:0/16:1/20:2)	6.65287
PC(16:0/18:1)	磷脂酰胆碱 (16:0/18:1)	6.08453
PS(O-18:0/17:0)	烷基磷脂酰丝氨酸 (O-18:0/17:0)	5.73065
PS(P-18:0/22:2)	缩醛磷脂酰丝氨酸 (P-18:0/22:2)	5.57504
TG(15:1/17:0/20:3)	三酰甘油 (15:1/17:0/20:3)	5.56667
PA(19:0/18:1)	磷脂酸 (19:0/18:1)	5.31453
PC(16:0/22:6)	磷脂酰胆碱 (16:0/22:6)	5.23476
TG(16:0/16:0/20:3)	三酰甘油 (16:0/16:0/20:3)	5.17578
TG(16:0/18:1/20:2)	三酰甘油 (16:0/18:1/20:2)	5.00514

[0116] 我们通过VIP>1筛选出来的化合物在S-plot上用红色方框标出,发现他们均匀的分布在两侧。同时,我们利用火山图,来考察筛选出的化合物的分布,如图3所示,发现筛选出来的化合物都分布在火山图的外围。这些结果都说明,借由OPLS-DA模型,我们成功筛选出造成CK和JC差异的化合物。

[0117] 建立逻辑回归模型以及ROC曲线。我们对脂质组学的数据进行标准化操作 (Rproject:scale)。然后将VIP>1的排名前9位变量的数据提取出来,建立逻辑回归模型和ROC曲线。

[0118] 公式中,TC:是否患有肿瘤,0为否,1为是

[0119] R1:PG (17:0/14:1)

[0120] R2:PS (O-20:0/18:1)

[0121] R3:PC (16:0/18:2)

[0122] R4:bacteriohopane-,32,33,34-triol-35-cyclitolg  
uanine

[0123] R5:PC (16:0/20:4)



- [0124] R6: TG (16:0/16:1/20:2)  
 [0125] R7: PC (16:0/18:1)  
 [0126] R8: PS (0-18:0/17:0)  
 [0127] R9: PS (P-18:0/22:2)  
 [0128] 模型3:  $TC = 1.8002 - 2.2815 * R1 - 2.3474 * R4 + 2.8573 * R9$   
 [0129] AIC: 57.484

---

系数显著性

---

- |        | (Intercept) | *** |
|--------|-------------|-----|
| [0130] | R1          | *   |
|        | R4          | *   |
|        | R9          | *   |
- 

[0131] Signif. codes: '\*\*\*' 0.001; '\*\*' 0.01; '\*' 0.05; '.' 0.1;

[0132] 模型2:

- [0133]  $TC = 1.6361 - 12.5962 * R + 0.4081 * R2 - 0.962 * R3 - 1.7675 * R4 + 0.7317 * R5 - 7.3848 * R6 + 15.9658 * R7 + 0.494 * R8 + 2.5964 * R9$   
 [0134] AIC: 65.973

---

系数显著性

---

- |        | (Intercept) | * |
|--------|-------------|---|
|        | R1          |   |
|        | R2          |   |
|        | R3          |   |
| [0135] | R4          |   |
|        | R5          |   |
|        | R6          |   |
|        | R7          |   |
|        | R8          |   |
|        | R9          |   |
- 

[0136] Signif. codes: '\*\*\*' 0.001; '\*\*' 0.01; '\*' 0.05; '.' 0.1;

[0137] 模型1:

- [0138]  $TC = 1.6054 - 13.4331 * R1 - 2.4503 * R4 + 0.9397 * R5 + 9.5919 * R7 + 3.3108 * R9$   
 [0139] AIC: 58.782

系数显著性	
(Intercept)	**
R1	
[0140] R4	*
R5	
R7	
R9	*

[0141] Signif. codes: '\*\*\*' 0.001; '\*\*' 0.01; '\*' 0.05; '.' 0.1;

[0142] 针对以上三个逻辑回归模型,进行ROC曲线绘制,如图4所示,模型1为M1,AUC值达到0.866;模型2为M2,AUC值为0.864;模型3为M3,AUC值达到0.872。我们发现模型3最靠近左上角定点,同时AUC值最高,最终,确定模型3为基于脂质指标来预测糖尿病血脂异常较好的预测模型。也可以根据所测样品数据情况,选择模型1或者模型2进行计算预测,或者优先选择模型3进行计算,同时使用1和2模型进行辅助验证,协同分析预测。

[0143] 尽管本发明的实施方案已公开如上,但其并不仅仅限于说明书和实施方式中所列运用,它完全可以被适用于各种适合本发明的领域,对于熟悉本领域的人员而言,可容易地实现另外的修改,因此在不背离权利要求及等同范围所限定的一般概念下,本发明并不限于特定的细节和这里示出与描述的图例。

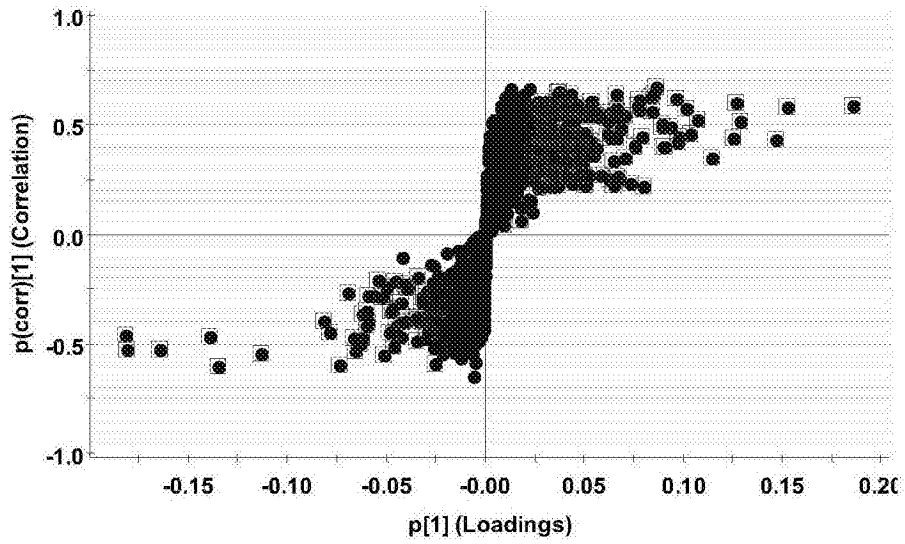


图1

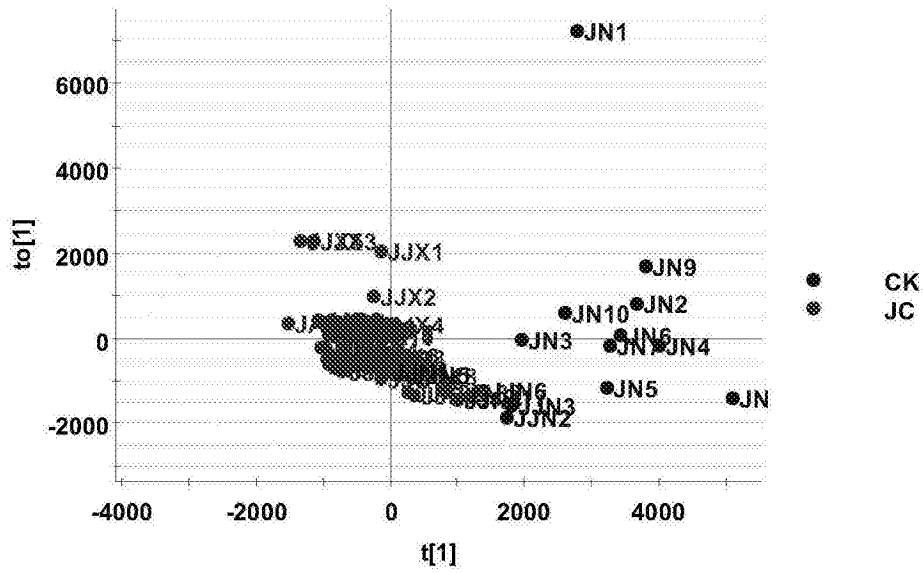


图2

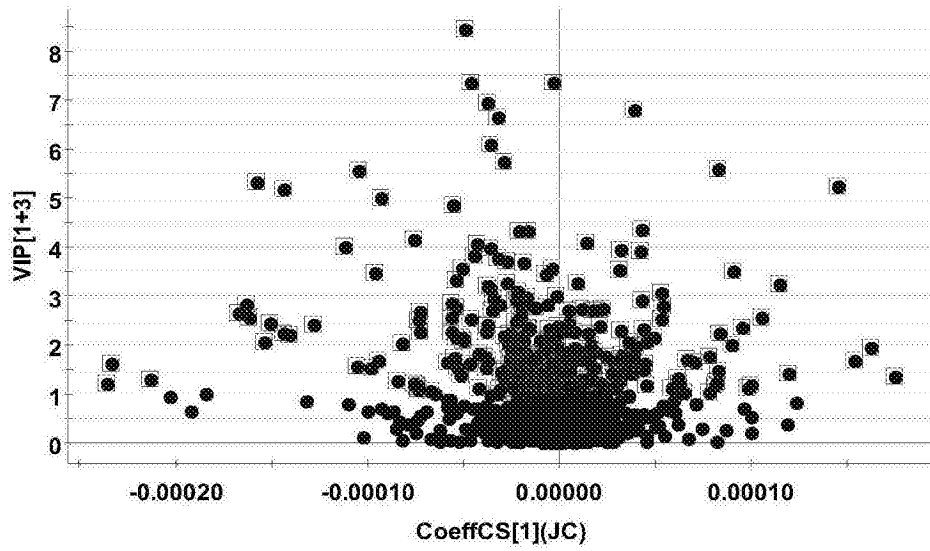


图3

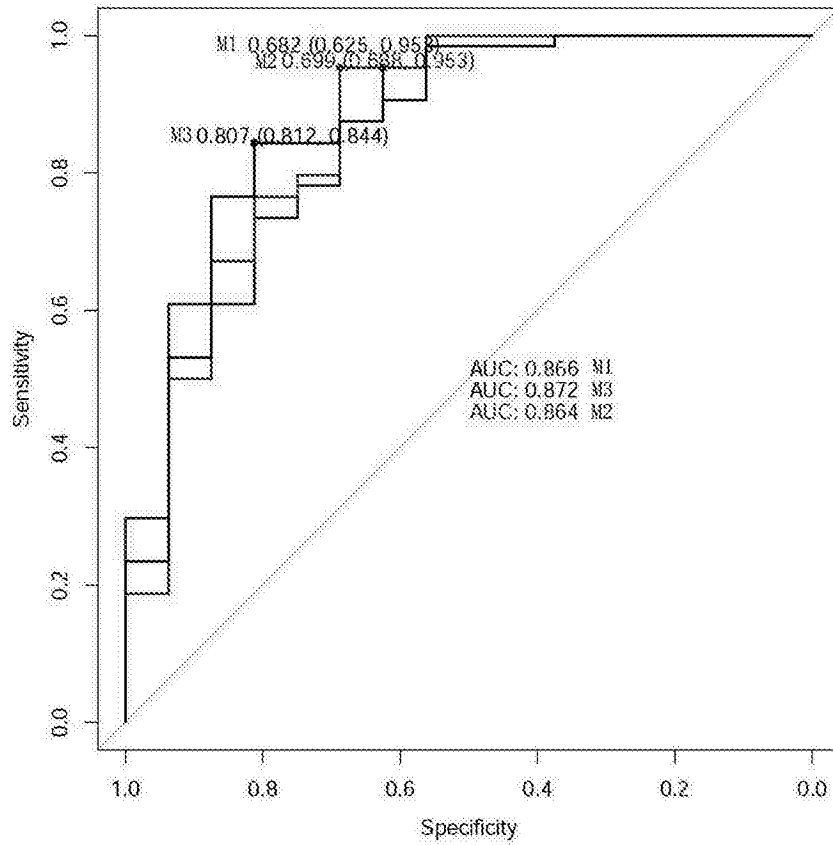


图4