(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 112151111 B
(45)授权公告日 2022.10.11

(21)申请号 202010881483.6

(22)申请日 2020.08.27

(65)同一申请的已公布的文献号
申请公布号 CN 112151111 A

(43)申请公布日 2020.12.29

(73)专利权人 上海大学
地址 200444 上海市宝山区上大路99号

(72)发明人 赵娟娟 刘秀娟 陆文聪

(74)专利代理机构 上海上大专利事务所(普通合伙) 31205
专利代理师 何文欣

(51)Int.Cl.
G16B 15/00(2019.01)
G16B 40/00(2019.01)

(56)对比文件
CN 110573518 A,2019.12.13
CN 107001374 A,2017.08.01

审查员 马金驹

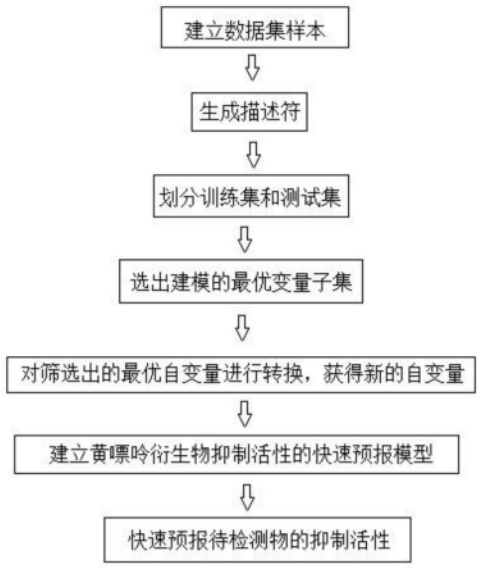
权利要求书1页 说明书8页 附图3页

(54)发明名称

基于多元线性回归快速预测黄嘌呤衍生物抑制活性的QSAR方法

(57)摘要

本发明涉及一种基于多元线性回归快速预测黄嘌呤衍生物抑制活性的QSAR方法，建立数据集样本；生成描述符；随机划分训练集和测试集；利用最大相关最小冗余方法结合多元线性回归留一法验证筛选变量，选出建模的最优变量子集；对筛选出的最优自变量进行转换，获得新的自变量；用多元线性回归建立黄嘌呤衍生物抑制活性的快速预报模型；根据建立的黄嘌呤衍生物的快速预报模型和待检测的黄嘌呤衍生物，快速预报待检测物的抑制活性。本发明基于可靠的文献真实值和建模方法，所建的黄嘌呤衍生物抑制活性的预报模型具有方便快捷，无化学污染等优点。



CN 112151111 B

1.一种基于多元线性回归快速预测黄嘌呤衍生物抑制活性的QSAR方法,其特征在于,包括如下步骤:

1)利用计算机系统,从文献中查找能够抑制DPP-IV活性的黄嘌呤衍生物结构及其对应的IC_{50}值,作为数据集样本;

2)用Chemdraw画出收集到的黄嘌呤结构,再用Dragon软件生成描述符;

3)以IC_{50}值的负对数为目标变量,Dragon生成的描述符为自变量,以互信息阈值为0.45初步筛选数据集;对初步筛选的数据集,随机划分训练集和测试集,测试集的比例占整个数据集的20%;

4)利用最大相关最小冗余方法结合多元线性回归留一法,验证筛选变量,选出建模的最优变量子集;

5)对筛选出的最优自变量进行转换,获得新的自变量;

6)在生成新的描述符的基础上,用多元线性回归方法,建立黄嘌呤衍生物抑制活性的快速预报模型;

7)根据建立的黄嘌呤衍生物的快速预报模型和待检测的黄嘌呤衍生物,快速预报待检测物的抑制活性;

所述步骤5)中获得的新的自变量为:

P(1)=+0.5318X1+0.1015X2+0.01403X3+3.751X4+0.08761X5+0.08958X6+0.5885X7+0.1830X8+0.004620X9+0.9556X10-51.580

P(2)=-0.7803X1-0.2487X2-0.005499X3+4.865X4+0.01781X5+0.06662X6+0.2001X7+0.09473X8-0.006231X9+0.7070X10-16.425

P(3)=-1.326X1+0.2069X2-0.01321X3+5.587X4-0.006104X5+0.05154X6+0.3427X7-0.002190X8-0.008101X9+0.6040X10-8.878

P(4)=-1.085X1+0.04423X2+0.0003891X3+5.462X4-0.1017X5-0.02304X6+0.1587X7-0.2447X8+0.005499X9+1.322X10-10.836

P(5)=-0.1980X1-0.01055X2-0.02453X3+8.016X4-0.09841X5-0.004190X6+0.9428X7-0.3690X8+0.006330X9+0.1503X10-24.193

P(6)=-0.5804X1-0.006145X2+0.0003567X3+8.097X4-0.1426X5+0.04710X6+0.5413X7-0.1006X8+0.007435X9-1.881X10-7.663

P(7)=-0.9359X1-0.05420X2+0.01620X3+2.366X4-0.0156X5-0.04107X6+1.580X7-0.1670X8+0.003524X9-1.594X10+15.090

P(8)=+0.4573X1+0.002638X2+0.02015X3+2.915X4-0.1471X5-0.09858X6+0.7866X7-0.01202X8-0.006955X9-0.1316X10-12.377

P(9)=-0.09285X1-0.005898X2-0.006817X3-3.525X4-0.1834X5+0.01118X6+0.9294X7+0.2454X8+0.0009003X9+0.4599X10+10.845。

基于多元线性回归快速预测黄嘌呤衍生物抑制活性的QSAR方法

技术领域

[0001] 本发明涉及黄嘌呤衍生物抑制活性的预测,特别是一种基于多元线性回归快速预测黄嘌呤衍生物抑制活性的定量构效关系(quantitative structure activity relationship,简称QSAR)方法。

技术背景

[0002] 黄嘌呤衍生物属于二肽基肽酶IV(dipeptidyl peptidase IV,简称DPP-IV)抑制剂的其中一种,有降低糖尿病患者血糖水平的作用。DPP-IV是一种丝氨酸蛋白酶,能够从多肽的N-末端分裂出X-脯氨酸二肽(X为任意氨基酸)。当人们在进餐后,体内血糖升高,葡萄糖依赖性的胰高血糖素样肽1(GLP-1)能刺激胰岛β细胞分泌胰岛素,但其大部分被DPP-IV降解失活。黄嘌呤衍生物作为DPP-IV抑制剂中的一种,能够抑制DPP-IV的活性。因具有降糖作用,黄嘌呤衍生物受到人们的关注。

[0003] 半抑制活性浓度(IC_{50})是指被测量的拮抗剂的半抑制浓度,即某一种药物或者物质(抑制剂)在抑制某些生物程序(或者是包括在此程序中的某些物质,如酶,细胞受体或微生物)的半量。IC_{50}值越低,意味着此抑制剂的抑制活性效果越好。

[0004] 定量构效关系(quantitative structure activity relationship,简称QSAR)作为一种统计模型,是用来分析分子结构与分子的某种活性之间的关系,包含与机器学习方法的结合,已经广泛用于药物发现和先导物优化中。

[0005] 最大相关最小冗余(mRMR)是一种常见的用于自变量筛选的方法。该方法是基于所选择的特征之间的冗余度应最小,与目标变量之间的相关性最大的理论来筛选自变量。

[0006] 多元线性回归(multiple linear regression,简称MLR)是多元数据分析的传统标准方法。该算法通过建立因变量和多个自变量之间的回归模型,从而得到线性方程,最终可以用来预测新的数据。如何应用多元线性回归建模,实现快速预报待检测物的抑制活性,成为亟待解决的技术问题。

发明内容

[0007] 本发明的目的是为了克服现有技术存在的缺陷,提供一种基于多元线性回归快速预测黄嘌呤衍生物的抑制活性QSAR方法,通过计算黄嘌呤衍生物二维结构的描述符,利用最大相关最小冗余筛选变量,并借助多元线性回归算法建模,预测黄嘌呤衍生物的pIC_{50}值,IC_{50}值的负对数。通过这些方法能几分钟就可得到结果,方便快捷,无需实验和繁杂的计算。

[0008] 本发明的目的可通过如下的技术方案实现:

[0009] 一种基于多元线性回归快速预测黄嘌呤衍生物抑制活性的QSAR方法,包括如下步骤:

[0010] 1)利用计算机系统,从文献中查找能够抑制DPP-IV活性的黄嘌呤衍生物结构及其对应的IC50值,作为数据集样本;

[0011] 2)用Chemdraw画出收集到的黄嘌呤结构,再用Dragon软件生成描述符;

[0012] 3)以IC_{50}值的负对数(pIC_{50})为目标变量,Dragon生成的描述符为自变量,以互信息阈值为0.45初步筛选数据集;对初步筛选的数据集,随机划分训练集和测试集,测试集的比例占整个数据集的20%;

[0013] 4)利用最大相关最小冗余方法结合多元线性回归留一法验证筛选变量,选出建模的最优变量子集;

[0014] 5)对筛选出的最优自变量进行转换,获得新的自变量;

[0015] 6)用多元线性回归建立黄嘌呤衍生物抑制活性的快速预报模型;

[0016] 7)根据建立的黄嘌呤衍生物的快速预报模型和待检测的黄嘌呤衍生物,快速预报待检测物的抑制活性。

[0017] 优选地,所述步骤5)中获得的新的自变量为:

[0018] P(1)=+0.5318X1+0.1015X2+0.01403X3+3.751X4+0.08761X5+0.08958X6]+0.5885X7+0.1830X8+0.004620X9+0.9556X10-51.580

[0019] P(2)=-0.7803X1-0.2487X2-0.005499X3+4.865X4+0.01781X5+0.06662X6+0.2001X7+0.09473X8-0.006231X9+0.7070X10-16.425

[0020] P(3)=-1.326X1+0.2069X2-0.01321X3+5.587X4-0.006104X5+0.05154X6+0.3427X7-0.002190X8-0.008101X9+0.6040X10-8.878

[0021] P(4)=-1.085X1+0.04423X2+0.0003891X3+5.462X4-0.1017X5-0.02304X6+0.1587X7-0.2447X8+0.005499X9+1.322X10-10.836

[0022] P(5)=-0.1980X1-0.01055X2-0.02453X3+8.016X4-0.09841X5-0.004190X6+0.9428X7-0.3690X8+0.006330X9+0.1503X10-24.193

[0023] P(6)=-0.5804X1-0.006145X2+0.0003567X3+8.097X4-0.1426X5+0.04710X6+0.5413X7-0.1006X8+0.007435X9-1.881X10-7.663

[0024] P(7)=-0.9359X1-0.05420X2+0.01620X3+2.366X4-0.0156X5-0.04107X6+1.580X7-0.1670X8+0.003524X9-1.594X10+15.090

[0025] P(8)=+0.4573X1+0.002638X2+0.02015X3+2.915X4-0.1471X5-0.09858X6+0.7866X7-0.01202X8-0.006955X9-0.1316X10-12.377

[0026] P(9)=-0.09285X1-0.005898X2-0.006817X3-3.525X4-0.1834X5+0.01118X6+0.9294X7+0.2454X8+0.0009003X9+0.4599X10+10.845。

[0027] 本发明与现有技术比,具有以下显而易见的突出实质性特点和显著的技术进步:

[0028] 1.本发明避免了重复试验,不断试错的过程,利用Dragon软件对画好的黄嘌呤衍生物结构生成描述符,经过变量筛选与多元线性回归建模,可提前预判黄嘌呤衍生物的抑制活性,也能给药物研发人员提供参考,缩短研发时间,降低研发成本;

[0029] 2.本发明是在Dragon软件生成自变量并进行一定的筛选的基础上再对变量进行转换,再以多元线性回归建模,操作过程简单,成本低,仅需一人便可完成;

[0030] 3.本发明整个过程不涉及实验及化学品,不产生环境污染,符合绿色环保理念。

附图说明

[0031] 图1为本发明的程序框图。

[0032] 图2为本发明的黄嘌呤衍生物抑制活性的多元线性回归模型建模效果图。

[0033] 图3为本发明的黄嘌呤衍生物抑制活性的多元线性回归模型留一法交叉验证结果图。

[0034] 图4为本发明的黄嘌呤衍生物抑制活性的多元线性回归模型独立测试集结果图。

具体实施方式

[0035] 以下优选实施例结合附图对本发明进行详细的说明：

[0036] 实施例一：

[0037] 参见图1和图2，一种基于多元线性回归快速预测黄嘌呤衍生物抑制活性的QSAR方法，包括如下步骤：

[0038] 1)利用计算机系统，从文献中查找能够抑制DPP-IV活性的黄嘌呤衍生物结构及其对应的IC_{50}值，作为数据集样本；

[0039] 2)用Chemdraw画出收集到的黄嘌呤结构，再用Dragon软件生成描述符；

[0040] 3)以IC_{50}值的负对数为目标变量，Dragon生成的描述符为自变量，以互信息阈值为0.45初步筛选数据集；对初步筛选的数据集，随机划分训练集和测试集，测试集的比例占整个数据集的20％；

[0041] 4)利用最大相关最小冗余方法结合多元线性回归留一法，验证筛选变量，选出建模的最优变量子集；

[0042] 5)对筛选出的最优自变量进行转换，获得新的自变量；

[0043] 6)用多元线性回归方法，建立黄嘌呤衍生物抑制活性的快速预报模型；

[0044] 7)根据建立的黄嘌呤衍生物的快速预报模型和待检测的黄嘌呤衍生物，快速预报待检测物的抑制活性。

[0045] 本实施例通过计算黄嘌呤衍生物二维结构的描述符，利用最大相关最小冗余筛选变量，并借助多元线性回归算法建模，预测黄嘌呤衍生物的pIC_{50}值，IC_{50}值的负对数。通过这些方法能几分钟就可得到结果，方便快捷，无需实验和繁杂的计算。

[0046] 实施例二：

[0047] 本实施例与实施例一基本相同，特别之处如下：

[0048] 所述步骤5)中获得的新的自变量为：

[0049] P(1)＝+0.5318X1+0.1015X2+0.01403X3+3.751X4+0.08761X5+0.08958X6]+0.5885X7+0.1830X8+0.004620X9+0.9556X10-51.580

[0050] P(2)＝-0.7803X1-0.2487X2-0.005499X3+4.865X4+0.01781X5+0.06662X6+0.2001X7+0.09473X8-0.006231X9+0.7070X10-16.425

[0051] P(3)＝-1.326X1+0.2069X2-0.01321X3+5.587X4-0.006104X5+0.05154X6+0.3427X7-0.002190X8-0.008101X9+0.6040X10-8.878

[0052] P(4)＝-1.085X1+0.04423X2+0.0003891X3+5.462X4-0.1017X5-0.02304X6+0.1587X7-0.2447X8+0.005499X9+1.322X10-10.836

[0053] P(5)＝-0.1980X1-0.01055X2-0.02453X3+8.016X4-0.09841X5-0.004190X6+0.9428X7-0.3690X8+0.006330X9+0.1503X10-24.193

[0054] P(6)＝-0.5804X1-0.006145X2+0.0003567X3+8.097X4-0.1426X5+0.04710X6+

0.5413X7-0.1006X8+0.007435X9-1.881X10-7.663

[0055] P(7)＝-0.9359X1-0.05420X2+0.01620X3+2.366X4-0.0156X5-0.04107X6+1.580X7-0.1670X8+0.003524X9-1.594X10+15.090

[0056] P(8)＝+0.4573X1+0.002638X2+0.02015X3+2.915X4-0.1471X5-0.09858X6+0.7866X7-0.01202X8-0.006955X9-0.1316X10-12.377

[0057] P(9)＝-0.09285X1-0.005898X2-0.006817X3-3.525X4-0.1834X5+0.01118X6+0.9294X7+0.2454X8+0.0009003X9+0.4599X10+10.845。

[0058] 本实施例对筛选出的最优自变量进行转换，获得新的自变量，提供丰富的变量条件和变量资源。

[0059] 实施例三：

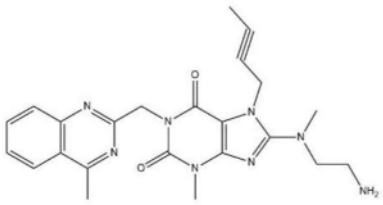
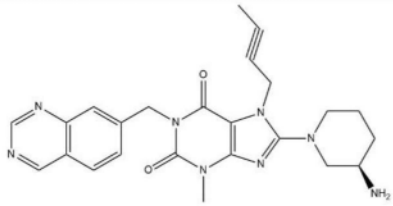
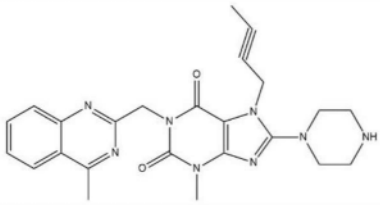
[0060] 本实施例与上述实施例基本相同，特别之处如下：

[0061] 一种基于多元线性回归快速预测黄嘌呤衍生物抑制活性的QSAR方法，包括如下步骤：

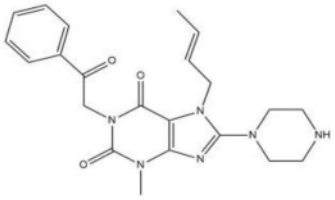
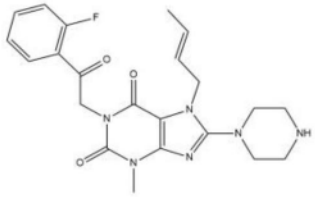
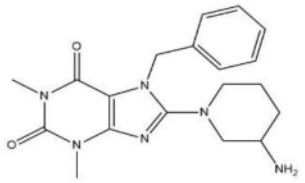
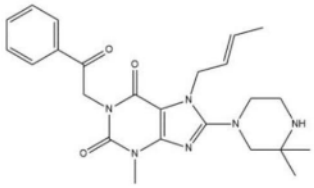
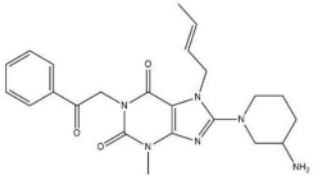
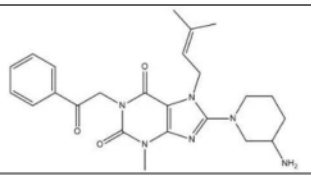
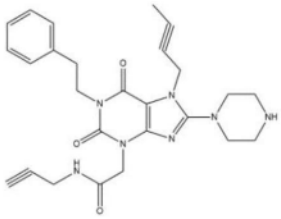
[0062] (1)利用计算机系统，在文献中查找对DPP-IV有抑制活性的黄嘌呤衍生物结构以及对应的IC_{50}值，共找到符合要求的黄嘌呤衍生物51个，部分结构及IC_{50}值如表1所示：

[0063] 表1.部分文献中黄嘌呤衍生物结构及其IC_{50}值

[0064]

结构	IC_{50} (nM)	pIC_{50}
	2	2.70
	1	3.00
	3	2.52

[0065]

	3	2.52
	3	2.52
	57	1.24
	1	3.00
	2	2.70
	2	2.70
	47	1.33

[0066] (2)用Dragon软件对Chemdraw画出的黄嘌呤二维结构生成描述符共1922个,部分描述符如表2所示:

[0067] 表2.Dragon生成的部分黄嘌呤衍生物描述符

[0068]

UNIP	SMTIV	GMTIV	Ho_D	Me	MW	Sv	Se	Sp	Si
126	27051	57161	40.40	1.06	420.30	30.6924	34.9419	29.11	35.37
133	29400	62581	41.68	1.06	434.31	31.4504	36.1019	29.74	36.66

[0069]

139	30845	64501	41.89	1.05	442.33	32.7356	36.7746	31.28	37.45
115	23106	48965	37.80	1.06	396.28	28.6924	32.9419	27.11	33.37
122	25686	55853	39.08	1.07	415.28	29.3391	34.3964	27.43	34.92
82	15123	31761	32.31	1.06	344.25	24.9776	28.6146	23.66	29.16
127	27022	57077	40.35	1.06	420.30	30.6924	34.9419	29.11	35.37
121	25056	53021	39.09	1.06	408.29	29.6924	33.9419	28.11	34.37
126	27051	57161	40.40	1.06	420.30	30.6924	34.9419	29.11	35.37
145	34146	72277	44.60	1.06	458.33	33.4504	38.1019	31.74	38.66

[0070] (3)以IC_{50}值的负对数为目标变量,Dragon生成的描述符为自变量,以互信息阈值为0.45初步筛选数据集,获得28个描述符;随机划分训练集与测试集,比例为4:1,训练集与测试集的样本量分别为41和10;

[0071] (4)以最大相关最小相关冗余结合多元线性回归筛选描述符,选出了10个最优描述符,分别为X1:SM3_Dz(p);X2:F08[C-O];X3:UNIP;X_4:HyWi_B(v);X5:Ho_D;X6:SpPos_B(v);X7:SpPosLog_D/Dt;X8:Ho_B(p);X9:SpDiam_Dz(p);X10:HyWi_Dz(e),部分样本的10个最优描述符的数据如表3所示:

[0072] 表3. 10个最优描述符的部分数据

[0073]

X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
16.4087	6	139	4.3901	41.8870	44.8475	4.7544	21.7634	353.9451	9.2431
15.8961	6	115	4.2586	37.7979	39.3125	4.1958	19.1507	295.4847	8.8866
16.0969	6	122	4.2824	39.0830	40.4018	4.3923	19.6239	316.0434	8.9829
15.0632	6	82	4.1205	32.3109	34.1970	3.3797	16.6601	206.8854	8.2355
16.2044	8	127	4.3182	40.3495	42.0449	4.5962	20.4301	328.7472	9.0965
16.0510	7	121	4.2889	39.0893	40.6934	4.3958	19.7638	312.4974	8.9970
16.1884	7	126	4.3182	40.3996	41.9668	4.5952	20.3803	323.4780	9.0894
16.7932	11	145	4.4135	44.6039	46.0802	5.1917	22.4280	359.7743	9.3699
16.3220	8	138	4.3670	40.5805	43.7079	4.5706	21.3295	347.0941	9.1879

[0074] (5)基于筛选出来的描述符,根据以下公式进行转换,生成新的描述符,参见表4,公式如下:

[0075] P(1)=+0.5318X1+0.1015X2+0.01403X3+3.751X4+0.08761X5+0.08958X6]+0.5885X7+0.1830X8+0.004620X9+0.9556X10-51.580

[0076] P(2)=-0.7803X1-0.2487X2-0.005499X3+4.865X4+0.01781X5+0.06662X6+0.2001X7+0.09473X8-0.006231X9+0.7070X10-16.425

[0077] P(3)=-1.326X1+0.2069X2-0.01321X3+5.587X4-0.006104X5+0.05154X6+0.3427X7-0.002190X8-0.008101X9+0.6040X10-8.878

[0078] P(4)=-1.085X1+0.04423X2+0.0003891X3+5.462X4-0.1017X5-0.02304X6+0.1587X7-0.2447X8+0.005499X9+1.322X10-10.836

[0079] P(5)＝-0.1980X1-0.01055X2-0.02453X3+8.016X4-0.09841X5-0.004190X6+0.9428X7-0.3690X8+0.006330X9+0.1503X10-24.193

[0080] P(6)＝-0.5804X1-0.006145X2+0.0003567X3+8.097X4-0.1426X5+0.04710X6+0.5413X7-0.1006X8+0.007435X9-1.881X10-7.663

[0081] P(7)＝-0.9359X1-0.05420X2+0.01620X3+2.366X4-0.0156X5-0.04107X6+1.580X7-0.1670X8+0.003524X9-1.594X10+15.090

[0082] P(8)＝+0.4573X1+0.002638X2+0.02015X3+2.915X4-0.1471X5-0.09858X6+0.7866X7-0.01202X8-0.006955X9-0.1316X10-12.377

[0083] P(9)＝-0.09285X1-0.005898X2-0.006817X3-3.525X4-0.1834X5+0.01118X6+0.9294X7+0.2454X8+0.0009003X9+0.4599X10+10.845

[0084] 表4.转换生成的部分新的描述符

[0085]

P(1)	P(2)	P(3)	P(4)	P(5)	P(6)	P(7)	P(8)	P(9)
1.1064	0.9479	-0.3479	-0.0303	0.0468	0.1412	-0.0572	-0.0444	0.0107
-2.2677	0.1517	-0.2735	0.0993	0.1224	-0.1157	-0.0702	0.0471	-0.0033
-1.3741	0.1918	-0.4931	0.0142	0.1259	-0.1385	-0.0028	0.0461	-0.0155
-6.5981	-0.4351	0.3150	0.0428	0.0419	0.1538	-0.0042	0.0003	-0.0135
-0.2166	0.0075	0.0228	0.0720	0.1084	-0.1150	-0.0822	0.0047	0.0397
-1.2354	0.0813	-0.1230	0.1011	0.1301	-0.1186	-0.0799	0.0206	0.0100
-0.3841	0.2927	-0.1160	0.0149	0.1250	-0.1321	-0.0274	0.0123	0.0226
2.8659	-0.1804	0.4527	-0.2944	-0.0914	-0.2760	-0.2302	0.0912	-0.0292
0.6740	0.23771	-0.0260	0.1507	-0.0498	0.1209	-0.2425	0.0537	0.0135
0.5839	0.4707	-0.2593	0.0426	-0.0380	0.3840	-0.0352	0.0611	-0.0226

[0086] (6)在生成新的描述符的基础上，用多元线性回归建立黄嘌呤衍生物抑制活性的快速预报模型；

[0087] (7)根据建立的黄嘌呤衍生物抑制活性的快速预报模型和待检测的黄嘌呤衍生物，快速预报待检测的黄嘌呤衍生物的抑制活性；

[0088] 在本实施例中，基于41个多元线性回归建立的黄嘌呤衍生物的QSAR预报模型的建模效果，如图2所示。利用多元线性回归算法对41个黄嘌呤衍生物样本数据进行回归建模，建立黄嘌呤衍生物抑制活性的多元线性回归定量预报模型，模型预报值与文献真实值的相关系数为0.886，均方根误差为0.5263，p值小于0.0001。

[0089] 在本实施例中，基于41个多元线性回归建立的黄嘌呤衍生物的QSAR预报模型的留一法交叉验证的结果，如图3所示。利用留一法交叉验证对41个样本数据建立的黄嘌呤衍生物的多元线性回归模型进行交叉验证，留一法中黄嘌呤衍生物的模型预报值与文献真实值的相关系数为0.7741，均方根误差为0.7704。

[0090] 在本实施例中，基于41个多元线性回归建立的黄嘌呤衍生物的QSAR预报模型的独立测试集预报结果，如图4所示。通过建立的黄嘌呤衍生物的多元线性回归预报模型对独立测试集中的10个样本进行预报，预报结果较好，黄嘌呤衍生物抑制活性预报值与文献真实值的平均相对误差为30.73％。

[0091] 综上所述，上述实施例基于多元线性回归快速预测黄嘌呤衍生物抑制活性的QSAR

方法，包括以下步骤：(1)利用计算机系统，从文献中查找能够抑制DPP-IV活性的黄嘌呤衍生物结构及其对应的IC_{50}值，作为数据集样本。(2)用Chemdraw画出收集到的黄嘌呤结构，再用Dragon软件生成描述符。(3)以IC_{50}值的负对数为目标变量，Dragon生成的描述符为自变量，以互信息阈值为0.45初步筛选数据集。对初步筛选的数据集，随机划分训练集和测试集，测试集的比例占整个数据集的20％。(4)利用最大相关最小冗余方法结合多元线性回归留一法验证筛选变量，选出建模的最优变量子集。(5)对筛选出的最优自变量进行转换，获得新的自变量。(6)用多元线性回归建立黄嘌呤衍生物抑制活性的快速预报模型。(7)根据建立的黄嘌呤衍生物的快速预报模型和待检测的黄嘌呤衍生物，快速预报待检测物的抑制活性。上述实施例基于可靠的文献真实值和建模方法，所建的黄嘌呤衍生物抑制活性的预报模型具有方便快捷，无化学污染等优点。

[0092] 上述实施例方法避免了重复试验，不断试错的过程，利用Dragon软件对画好的黄嘌呤衍生物结构生成描述符，经过变量筛选与多元线性回归建模，可提前预判黄嘌呤衍生物的抑制活性，也能给药物研发人员提供参考，缩短研发时间，降低研发成本；上述实施例方法是在Dragon软件生成自变量并进行一定的筛选的基础上再对变量进行转换，再以多元线性回归建模，操作过程简单，成本低，仅需一人便可完成。

[0093] 上面对本发明实施例结合附图进行了说明，但本发明不限于上述实施例，还可以根据本发明的发明创造的目的做出多种变化，凡依据本发明技术方案的精神实质和原理下做的改变、修饰、替代、组合或简化，均应为等效的置换方式，只要符合本发明的发明目的，只要不背离本发明的技术原理和发明构思，都属于本发明的保护范围。

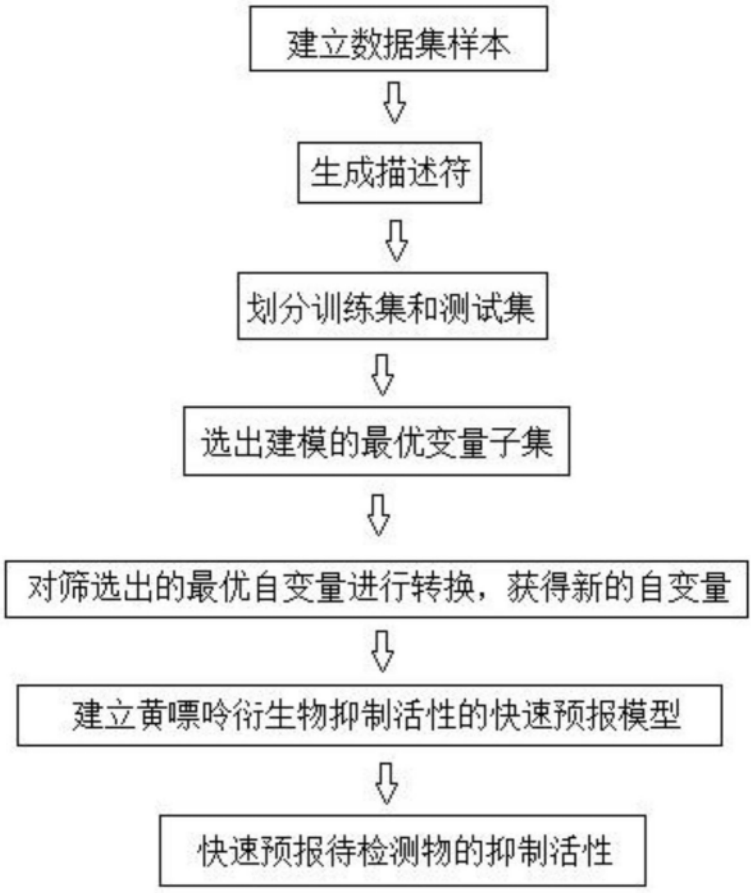

建立数据集样本
生成描述符
划分训练集和测试集
选出建模的最优变量子集
对筛选出的最优自变量进行转换，获得新的自变量
建立黄嘌呤衍生物抑制活性的快速预报模型
快速预报待检测物的抑制活性

图1

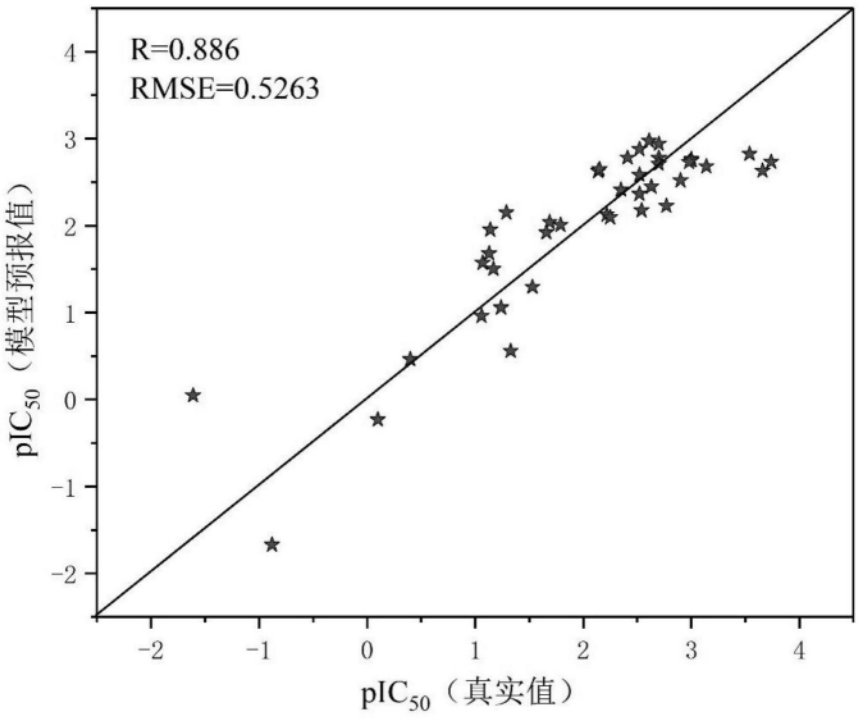

R=0.886
RMSE=0.5263
pIC50（模型预报值）
pIC50（真实值）

图2

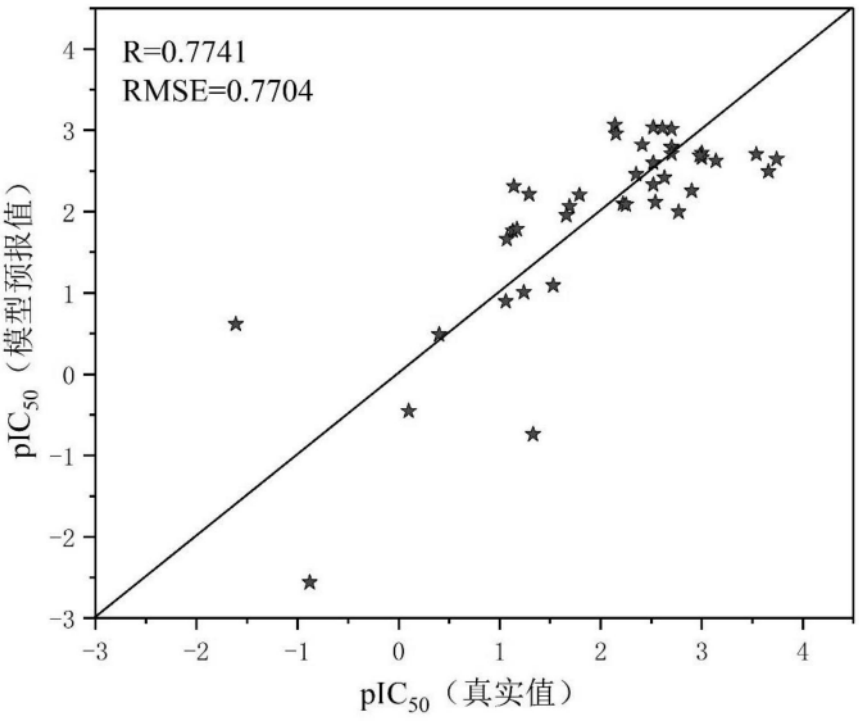

R=0.7741
RMSE=0.7704
pIC50（模型预报值）
pIC50（真实值）

图3

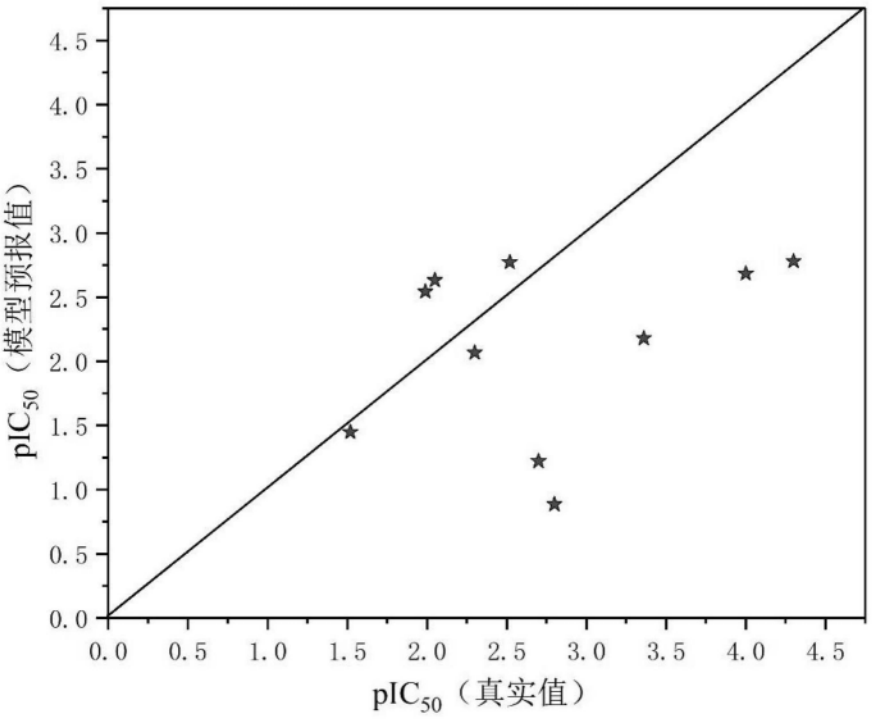

4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
pIC50（模型预报值）
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
pIC50（真实值）

图4