

核偏最小二乘法及其在 P-glycoprotein 抑制剂设计中应用

李 燕^{*1}, 王永华², 张述伟¹

(1. 大连理工大学 化工学院, 辽宁 大连 116012;

2. 大连水产学院 生命科学与技术学院, 辽宁 大连 116023)

摘要: 介绍了一个新的基于优化推导出的核偏最小二乘(kernel partial least squares, K-PLS)的算法原理和实现步骤, 并且给出了利用 K-PLS 法构建 P-糖蛋白(P-glycoprotein, P-gp)黄酮类抑制剂的定量构效关系(QSAR)模型. 利用该方法结合几个简单的分子拓扑参数, 构建了具有高准确率的预测模型. 该模型将有助于 P-gp 黄酮类抑制剂的虚拟筛选和理性设计. 结果证明 K-PLS 是一个十分稳定可靠的方法, 将会在化学计量学领域得到较好的应用和推广.

关键词: 核偏最小二乘; QSAR; P-glycoprotein; 抑制剂

中图分类号: R914.2 **文献标志码:** A

0 引言

偏最小二乘(partial least squares, PLS)法是一种新型的多元统计数据分析方法, 它于1983年由伍德和阿巴诺等首次提出^[1]. 20余年来, 它在理论、方法和应用方面都得到了迅速的发展. 偏最小二乘法在一个算法下, 可以同时实现回归建模(多元线性回归)、数据结构简化(主成分分析)以及两组变量之间的相关性分析(典型相关分析), 已经在化学计量学中得到广泛和成功的应用. 其特点是快速、容易编程实现, 同时能够较为有效地克服过度拟合等问题. 偏最小二乘回归可能是所有多元校正方法里对变量约束最少的方法, 这种灵活性让它适用于传统的多元校正方法所不适用的许多场合, 例如化学计量学中存在的—个常见现象, 即训练样本点较少但是参数的维数很高. 在这种情形下, PLS 方法却能够胜任.

过高的特征空间维数经常会导致建模准确率下降, 这时可以通过引入核函数, 从而发展成一种新型的非线性 PLS 方法, 即核偏最小二乘(kernel partial least squares, K-PLS)法. PLS 的空间映射性质使得核方法能够很容易和它结合. 简单地说, K-PLS 首先将初始输入点映射到高维特征空间, 然后在高维特征空间中计算主成分. 同 PLS 得到

的主成分具有二阶相关性不同, K-PLS 计算得到的非线性成分经常具有高阶相关性. K-PLS 能够给出和支持向量机(SVM)几乎相同的结果而且更加稳定, 同时程序编写也更加容易^[2]. 作为一种新近发展的方法, K-PLS 在化学计量学及其他领域还没有得到广泛应用. 因此, 探讨并推广 K-PLS 方法在生物制药领域的应用具有很高的理论和实际价值.

P-糖蛋白(P-glycoprotein, P-gp)是一类相对分子质量为 170 kD 的跨膜蛋白, 广泛分布于人体的小肠、肝、肾等许多重要器官中, 它隶属于人体 ABC 转运蛋白家族, 是多药耐药基因产物^[3]. P-gp 是肿瘤细胞产生多药耐药性(multi-drug resistance, MDR)的关键因素之一^[4]. 功能上, P-gp 是一种能量依赖型糖蛋白, 可在细胞内探测出“外来”药物, 并在 ATP 酶水解产生的推动力下, 不断地将进入细胞内的药物泵出细胞, 从而降低细胞内药物的剂量^[3]. 在化疗情况下, 正是 P-gp 这种对药物外排的作用大大地降低了化疗药物的药效, 使得多数肿瘤细胞产生对多种药物的耐药性, 不能被有效杀死^[5]. 因此要有效提高化疗效果, 必须寻找和筛选合适的 P-gp 多药耐药的抑制剂或逆转剂. 但是到目前为止还没有真正进入临床的 P-gp 抑制剂药物.

黄酮类化合物在自然界中广泛存在,主要存在于植物中,是人类食物中的重要成分,现在已知道的有超过 4 000 种不同的黄酮类分子^[6]. 该类化合物不仅具有多种有益的药理学作用,包括抗炎、杀菌、防衰老、抗心血管病等,近年研究还表明黄酮类化合物是最有希望的一种特异性 P-gp 抑制剂^[7],具有副作用小、选择性强等特点. 因此,如何从黄酮类系列化合物中开发出有应用前景的抑制剂是一个重要课题. 本文拟基于核偏最小二乘法原理,结合简单的分子拓扑参数,来构建 P-gp 抑制剂黄酮系列分子的 QSAR 模型,详细地探讨不同高斯核参数及成分数对建模的影响,以利于开发有效的肿瘤耐药逆转剂,及评价研究正常人药物-食物相互作用(drug-food interaction).

1 核偏最小二乘

假设输入变量 $\mathbf{X}$ 经过非线性变换映射到特征空间 Φ 中. 在特征空间中构建非线性 PLS 回归模型,通过映射 $\Phi(\mathbf{X})$ 得到非线性模型. 通过非线性变换,特征空间可以是高维空间. 一个常用的方法是采用高斯核函数,这样特征空间甚至可以构造到无限维. 这就是核偏最小二乘的基本思想. 即,对一个训练样本数据 $(\mathbf{X}, \mathbf{Y})$, 其中 $\mathbf{X}$ 是自变量, $\mathbf{Y}$ 是因变量, $\mathbf{X} \in \mathbf{R}^{m \times n}$, $\mathbf{Y} \in \mathbf{R}^{m \times 1}$, 定义一个映射 Φ 和 $\Phi(\mathbf{X})$.

PLS 是基于最小二乘回归的方法,因此若构造一个基本线性函数 $\mathbf{x}$ 和 $\mathbf{w}$ 的乘积 $f(\mathbf{x}) = \mathbf{x}\mathbf{w}$, 此处 $f(\mathbf{x})$ 就可近似表示 $\mathbf{y}$, 则解 $\mathbf{w}$ 的基本方法就是最小化预测值和实际值的残差平方和函数. 对于 K-PLS,也可采用同样的基本理论. 此时,最小二乘优化问题即为

$$\min \|\Phi(\mathbf{X}) - \mathbf{y}\mathbf{w}'\|^2, \text{ s. t. } \|\mathbf{w}\|^2 = 1 \quad (1)$$

其中 $\mathbf{X}$ 被 $\Phi(\mathbf{X})$ 取代. 本文给出的为单因变量算法步骤,因为这种情况在 QSAR 建模中最为常见.

首先,确定特征空间的克莱姆矩阵 $\mathbf{K}^0 = \Phi(\mathbf{X})\Phi(\mathbf{X})'$, 即 $K_{ij} = K(X_i, X_j)$. $\mathbf{K}^1$ 是 $\mathbf{K}^0$ 的中心化. 令 $\mathbf{Y}' = \mathbf{y}$, 并实行标准化操作,使得其均值为 0, 标准偏差为 1. 潜在变量数设为 M ;

从 $k = 1$ 到 M

$$1. \mathbf{K}^0 = \Phi(\mathbf{X})\Phi(\mathbf{X})', \text{ 即 } K_{ij} = K(X_i, X_j)$$

$$2. \mathbf{u}^m = \mathbf{K}^m \mathbf{K}^{m'} \mathbf{y}^m$$

$$3. \mathbf{u}^m = \mathbf{u}^m / \|\mathbf{u}^m\|$$

$$4. \mathbf{K}^{m+1} = \mathbf{K}^m - \mathbf{u}^m \mathbf{u}^{m'} \mathbf{K}^m$$

$$5. \mathbf{y}^{m+1} = \mathbf{y}^m - \mathbf{u}^m \mathbf{u}^{m'} \mathbf{Y}^m$$

$$6. \mathbf{y}^{m+1} = \mathbf{y}^{m+1} / \|\mathbf{y}^{m+1}\|$$

7. 最终的回归系数由下式计算得到:

$$\mathbf{r} = \mathbf{K}^1 \mathbf{Y} (\mathbf{U}' \mathbf{K}^1 \mathbf{K}^1' \mathbf{Y})^{-1} \mathbf{U}' \mathbf{y}$$

这里 $\mathbf{Y}$ 和 $\mathbf{U}$ 的第 m 列分别为 $\mathbf{y}^m$ 和 $\mathbf{u}^m$.

8. 训练数据预测的结果即为

$$f(\mathbf{x}) = \sum_{i=1}^k K(x_i, \mathbf{x}) r_i$$

验证集必须中心化, 中心化公式采用:

$$\mathbf{K}_{\text{center}}^{\text{train}} = \left(\mathbf{I} - \frac{1}{\kappa} \mathbf{1}\mathbf{1}' \right) \mathbf{K}^{\text{train}} \left(\mathbf{I} - \frac{1}{\kappa} \mathbf{1}\mathbf{1}' \right)$$

$$\mathbf{K}_{\text{center}}^{\text{test}} = \left(\mathbf{K}^{\text{test}} - \frac{1}{\kappa} \mathbf{1}\mathbf{1}' \mathbf{K}^{\text{train}} \right) \left(\mathbf{I} - \frac{1}{\kappa} \mathbf{1}\mathbf{1}' \right)$$

这里 $\mathbf{1}$ 为元素为 1 的向量, $\mathbf{I}$ 为单位阵. 以上算法可以很容易通过编程实现. 本研究的程序利用 C 语言编写.

2 数据

本研究所用数据来自文献^[7]. 共计 57 个不同的分子(表 1), 其中图 1 显示了表 1 中部分取代基对应的结构. 采用 MolconnZ 程序 (Tripos Associates, St Louis, MO) 计算分子参数. 分子参数的说明在表 2 中给出. 建模采用 K-PLS 方法.

黄酮类分子可以作为 P-gp 的抑制剂, 是因为黄酮类分子能够同 P-gp 的 C-末端的核苷结合区域(C-NBD)特异性结合. 这种结合会导致 P-gp 功能的下降或丧失. 结合力强的分子表现为具有强的 P-gp 抑制能力. 因此, 定量地预测结合力的大小就可以有效预测黄酮类分子对 P-gp 的抑制活性. 实验中结合力用 K_d 表示, 在建模中均转换成 $\log K_d$, 数据如表 1 所示. 在 Pietro 等的研究中, 黄酮类分子的活性实验主要包括 4 步^[7]:

I. 同纯化的重组鼠 C-NBD 或者全长 P-gp 直接结合;

II. 抑制 ATP 水解;

III. 通过流式细胞仪检测 P-gp 的转运活性;

IV. 测定细胞生长的化学感光性变化.

在建模过程中随机选择总样本的 1/4 作为检验集, 剩下的作为训练集. 训练集用来构建模型, 检验集的样本没有用于模型训练中.

3 结果和讨论

采用高斯核函数 $\exp(-(|\mathbf{x} - \mathbf{y}|^2)/\omega)$. 表 3 给出了主成分数为 3 时在不同 ω 参数下的模拟结果. 从表中可见, ω 取不同值时, 训练集和检验集的拟合和预测结果基本吻合, 没有很大差异. 而且从表中可见随 ω 的增大, 训练样本和检验样本的误差和相关系数有接近的趋势, 但是变化幅度很

表 1 P-gp 黄酮类抑制剂分子和其 $\log K_d$ 值
Tab. 1 Flavonoid inhibitors of P-gp with the binding affinity of $\log K_d$ values

序号	黄酮类抑制剂分子	分子参数及活性值							
		$\log K_d$	$C \log P$	Xv0	Xv1	Xv2	Xvp3	Xvp4	Xvp5
1	3-Me-galangin	0.949 4	3.762	10.752	6.125	4.589	3.222	2.237	1.498
2	3-OH-flavone	1.004 3	2.843	9.460	5.574	3.971	2.810	1.917	1.303
3	4-Br-chalcone	-0.244 1	4.487	10.723	6.079	4.531	2.884	1.709	1.245
4	4-Cl-chalcone	0.113 9	4.337	9.893	5.664	4.052	2.607	1.570	1.085
5	4-F-chalcone	0.556 3	3.767	9.137	5.286	3.615	2.355	1.444	0.940
6	4-I-chalcone	-0.602 1	4.747	11.295	6.365	4.861	3.075	1.804	1.355
7	4-MeO-chalcone	0.361 7	3.543	10.167	5.709	3.837	2.601	1.586	1.014
8	4-n-C ₁₀ H ₂₁ -chalcone	-1.221 8	8.884	16.123	10.158	7.030	4.784	3.129	2.114
9	4-n-C ₁₄ H ₂₉ -chalcone	1.152 3	11.000	18.951	12.158	8.444	5.784	3.836	2.614
10	4-n-C ₂ H ₅ -chalcone	0.415 0	4.652	10.466	6.158	4.159	2.873	1.735	1.121
11	4-n-C ₄ H ₉ -chalcone	0	5.710	11.880	7.158	4.909	3.284	2.047	1.409
12	4-n-C ₆ H ₁₃ -chalcone	-0.568 6	6.768	13.295	8.158	5.616	3.784	2.422	1.614
13	4-n-C ₈ H ₁₇ -chalcone	-1.699 0	7.826	14.709	9.158	6.323	4.284	2.775	1.864
14	4-n-cyclohexyl-chalcone	-0.275 7	6.244	12.872	8.202	6.065	4.499	3.186	2.299
15	4-OH-3-prenyl-chalcone	-0.275 7	4.908	12.913	7.288	5.437	3.319	2.057	1.412
16	4-OH-chalcone	0.681 2	2.957	9.206	5.320	3.655	2.379	1.456	0.953
17	6,7-diMe-chrysin	0.113 9	5.009	11.305	6.401	4.881	3.564	2.270	1.532
18	6-PhCH ₂ -chrysin	-0.468 5	5.530	13.846	8.243	6.099	4.357	3.049	2.114
19	6-geranyl-chrysin	-1.346 8	7.495	17.028	9.726	7.512	4.879	3.243	2.110
20	6-Me-chrysin	0.491 4	3.962	10.752	6.125	4.583	3.277	2.133	1.441
21	6-prenyl-chrysin	-0.522 9	5.464	13.537	7.675	5.923	3.847	2.582	1.771
22	6-prenyl-galangin	-0.677 8	4.465	13.907	7.822	6.057	3.966	2.645	1.820
23	7-O-isopropyl-chrysin	0.113 9	4.967	12.368	7.073	5.323	3.213	2.328	1.490
24	7-OH-flavone	1.542 8	3.211	9.460	5.562	4.021	2.771	1.888	1.291
25	8-PhCH ₂ -chrysin	-0.004 4	5.580	13.846	8.243	6.101	4.343	3.069	2.159
26	8-DMA-apigenin	-0.154 9	4.676	13.907	7.706	6.442	4.267	2.795	1.819
27	8-DMA-chrysin	-0.699 0	5.334	13.537	7.572	6.261	4.173	2.772	1.784
28	8-DMA-galangin	-0.346 8	4.335	13.907	7.718	6.395	4.292	2.835	1.831
29	8-DMA-kaempferide	-0.699 0	4.268	15.238	8.241	6.757	4.608	2.988	1.928
30	8-geranyl-chrysin	-1.602 1	7.545	17.028	9.726	7.513	4.865	3.263	2.160
31	8-prenyl-chrysin	-0.552 8	5.514	13.537	7.675	5.924	3.833	2.602	1.822
32	8-prenyl-galangin	-0.657 6	4.515	13.907	7.822	6.058	3.952	2.665	1.867
33	apigenin	1.004 3	2.905	10.200	5.837	4.360	2.946	2.022	1.332
34	brousochalcone A	-0.356 5	4.361	13.283	7.422	5.632	3.384	2.086	1.449
35	chalcone	0.662 8	3.624	8.836	5.186	3.474	2.285	1.435	0.910
36	chrysin	0.949 4	3.563	9.830	5.702	4.179	2.853	1.999	1.298
37	3',4'-diF-chrysin	0.799 3	3.779	10.431	5.908	4.442	2.999	2.046	1.336
38	4'-I-chrysin	0.342 4	4.686	12.288	6.881	5.565	3.643	2.370	1.707
39	dehydrosilybin	0.342 4	2.483	18.258	10.476	7.917	5.723	4.034	2.811
40	8-DMA,3,7-diMe-galangin	-0.823 9	6.630	15.012	8.271	6.969	4.761	3.328	2.089
41	galangin	0.724 3	2.564	10.200	5.849	4.313	2.971	2.060	1.352
42	4'-F-galangin	0.832 5	2.728	10.500	5.948	4.454	3.042	2.072	1.375
43	2',4'-diCl-galangin	0.602 1	2.878	10.801	6.054	4.571	3.111	2.131	1.387
44	4'-I-galangin	0.041 4	3.708	12.658	7.027	5.699	3.761	2.432	1.762
45	4'-n-C ₈ H ₁₇ -galangin	-1.221 8	6.766	16.072	9.820	7.162	4.970	3.403	2.294
46	genistein	1.423 2	2.405	10.200	5.830	4.363	3.020	2.059	1.363
47	6-geranyl-dehydrosilybin	-0.744 7	6.415	25.457	14.500	11.250	7.749	5.279	3.618
48	8-geranyl-dehydrosilybin	-0.920 8	6.465	25.457	14.500	11.251	7.735	5.299	3.665
49	kaempferide	0.653 2	2.497	11.530	6.372	4.675	3.288	2.213	1.448
50	kaempferol	0.826 1	1.900	10.569	5.983	4.494	3.065	2.084	1.387
51	naringenin	1.562 3	2.445	10.407	6.091	4.663	3.250	2.300	1.570
52	6-prenyl-dehydrosilybin	-0.431 8	4.384	21.965	12.449	9.661	6.718	4.618	3.280
53	8-prenyl-dehydrosilybin	-0.602 1	4.434	21.965	12.449	9.662	6.704	4.639	3.326
54	quercetin	0.845 1	1.304	10.939	6.123	4.650	3.173	2.135	1.410
55	silybin	0.832 5	1.949	18.413	10.703	8.200	6.000	4.280	3.025
56	taxifolin	1.572 9	0.771	11.094	6.350	4.934	3.450	2.382	1.621
57	tectochrysin	0.799 3	4.149	10.791	6.091	4.361	3.077	2.106	1.400

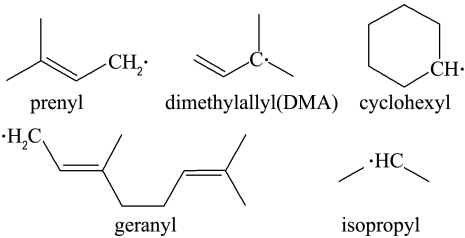


图 1 表 1 中部分取代基对应的结构
Fig. 1 Structures of part substituents in Tab. 1

表 2 建模所用分子参数的意义
Tab. 2 The definition of the molecular descriptors used in this work

分子参数	参数意义
$C \log P$	calculated distribution coefficient in n-octanol/water systems ^[8]
Xv0~Xv2	valence chi indices; the modified indices
Xvp3~Xvp5	include the valence of electron ^[9,10]

表 3 不同 ω 值下训练集和检验集的模拟结果
Tab. 3 The modeling results for the training and test sets with different ω values

ω	训练集 相关系数	预测集 相关系数	训练集 均方差	预测集 均方差
0.5	0.995	0.903	0.067	0.144
1	0.988	0.913	0.017	0.113
1.5	0.983	0.912	0.023	0.112
2	0.979	0.910	0.028	0.114
3	0.970	0.915	0.039	0.113
4	0.963	0.919	0.049	0.112
5	0.957	0.921	0.057	0.111
7	0.952	0.921	0.064	0.110

小. 对于预测集, 不同 ω 值下的预测结果十分接近, 反映了 K-PLS 的稳定性. 经过比较, 选择 $\omega=3$ 的情况, 这时训练样本和检验样本的误差和相关系数十分接近. 图 2 给出了在 $\omega=3$ 的情况下训

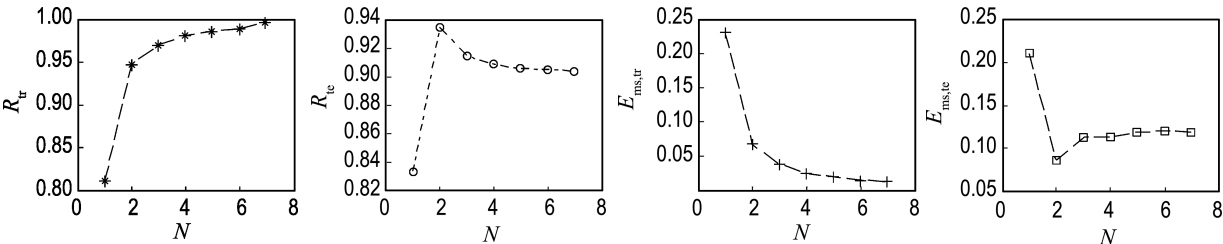


图 3 在 $\omega=3$ 时不同主成分数下训练集和检验集的相关系数和均方差
Fig. 3 The coefficients and residues for the training and test sets when $\omega=3$

在先前的研究中, 利用超过 50 个的分子参数并结合贝叶斯神经网络方法, 构建了一个结果相当的模型^[11]. 但是本研究中提出的基于 K-PLS 的模型更加简单, 大大简化了建模和运算过程. 利用该模型可以对于常见饮食中的大于 4 000 个的

训练集和检验集的模拟结果(活性数据中心化了). 可以看到训练和预测的拟合结果均比较好. 这时训练集和预测集的相关系数分别为 0.970 和 0.915, 拟合偏差分别为 0.039 和 0.113.

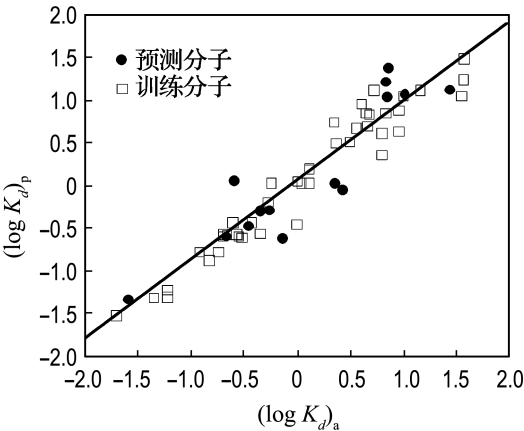


图 2 在 $\omega=3$ 的情况下训练集和检验集的模拟结果
Fig. 2 The modeling results for the training and tests set when $\omega=3$

为考察不同成分数下 K-PLS 模型的拟合能力, 本文将主成分数分别在 1~7 等几种情况做了改变, 得到一系列拟合偏差和回归相关系数的结果, 如图 3 所示.

由图 3 可见, 随着主成分数 N 的增加, 训练样本的相关系数 R 逐渐增大, 拟合残差 E 逐渐减小, 但是变化幅度均非常小. 检验样本的相关系数和预测残差趋势变化不明显, 呈现微小的波动, 然后趋于平衡. K-PLS 对主成分数不敏感, 从而再次显示了 K-PLS 的稳定性.

黄酮类同系物进行虚拟筛选, 判断它们的 P-gp 活性, 从而促进抑制剂的开发, 同时也可以预测由 P-gp 介导的黄酮类分子所引发的食物-药物相互作用.

4 结 论

本文给出了一种新发展的核偏最小二乘法的算法步骤. 同时, 利用该算法结合 7 个简单的分子连接性参数, 构建了一个具有很好的泛化能力的预测模型. 该模型的训练集和预测集的相关系数分别为 0.970 和 0.915, 预测偏差分别为 0.039 和 0.113.

本模型中采用的参数包括 $C \log P$ 等都可以通过程序快捷地计算得到, 因此该模型具有良好的实用性和可重复性. 该模型的成功开发将会促进 P-gp 黄酮类抑制剂的进一步设计和开发. 同时, 该研究将会促进这种新的算法——核偏最小二乘法在化学计量学等领域更大的应用和推广.

参考文献:

- [1] HÖSKULDSSON A. PLS regression methods [J]. *Journal of Chemometrics*, 1988, **2**(3):211-228
- [2] JOHN S T, NELLO C. *Kernel Methods for Pattern Analysis* [M]. London:Cambridge University Press, 2004
- [3] PAJEVA I K, GLOBISCH C, WIESE M. Structure-function relationships of multidrug resistance P-glycoprotein [J]. *Journal of Medical Chemistry*, 2004, **47**(10):2523-2533
- [4] WIESE M, PAJEVA I K. Structure-activity relationships of multidrug resistance reversers [J]. *Current Medicinal Chemistry*, 2001, **8**(6):685-713
- [5] WANG Y H, LI Y, YANG S L, *et al.* Classification of substrates and inhibitors of P-glycoprotein using unsupervised machine learning approach [J]. *Journal of Chemical Information and Modeling*, 2005, **45**(3):750-757
- [6] HARBORNE J B, WILLIAMS C A. Advances in flavonoid research since 1992 [J]. *Phytochemistry*, 2000, **55**(6):481-504
- [7] PIETRO A D, CONSEIL G, PEREZ-VICTORIA J M, *et al.* Modulation by flavonoids of cell multidrug resistance mediated by P-glycoprotein and related ABC transporters [J]. *Cellular & Molecular Life Sciences*, 2002, **59**(2):307-322
- [8] HANSCH A L. *Substituent Constants for Correlation Analysis in Chemistry and Biology* [M]. New York: Wiley, 1979
- [9] HALL L H. *Computational Aspects of Molecular Connectivity and Its Role in Structure-activity Modeling in Computational Chemical Graph Theory* [M]. New York:Nova Press, 1989
- [10] KIER L B, HALL L H. Differential molecular connectivity in data-base fragment searching [J]. *Pharmaceutical Research*, 1989, **6**(6):497-500
- [11] WANG Y H, LI Y, YANG S L, *et al.* An in silico approach for screening flavonoids as P-glycoprotein inhibitors based on Bayesian-regularized neural network [J]. *Journal of Computer Aided Molecular Design*, 2005, **19**(3):137-147

Kernel partial least squares method and its applications for flavonoids as P-glycoprotein inhibitors

LI Yan^{*1}, WANG Yong-hua², ZHANG Shu-wei¹

(1. School of Chemical Engineering, Dalian University of Technology, Dalian 116012, China;

2. College of Life Science and Biotechnology, Dalian Fisheries University, Dalian 116023, China)

Abstract: A novel derivation based on optimization for the kernel partial least squares (K-PLS) algorithm is provided. Using the K-PLS method, a quantitative structure-activity relationship (QSAR) model for flavonoid derivatives as P-glycoprotein (P-gp) inhibitors is built. The model exhibits high-accuracy by means of this method using several simple molecular topological parameters. The QSAR model for flavonoids will be helpful in screening and design of new P-gp inhibitors. The accuracy and efficiency of the QSAR model indicate that the K-PLS method will obtain more popularity and success for chemometrics applications.

Key words: kernel partial least squares; QSAR; P-glycoprotein; inhibitor