文章编号：1674－5566(2011)05－0773－06

基于分子电性距离矢量的CDK4抑制剂Fascaplysin定量构效关系研究

谢天宝，印春生，杨　红

(上海海洋大学 海洋科学学院，上海　201306)

摘　要：海洋生物的多样性、活性物质结构的新颖性以及海洋环境的独特性使得海洋生物成为天然抗肿瘤活性物质的重要来源之一。以分子电性距离矢量(MEDV-13)为结构描述子，应用基于预测的变量筛选(VMSP)方法，对44种海洋Fascaplysin类CDK4抑制剂的活性作用数据进行模拟分析，获得一个6变量的定量结构—活性关系(QSAR)模型，而且对样本内部使用LOO(Leave-One-Out)验证方法、外部通过测试集对模型进行了验证。模型相关系数$r=0.923\ 9$，LOO交互检验相关系数$q=0.878\ 9$，显示模型具有良好的估计能力和对外部样本的预测能力。随后，计算了模型最优子集各变量对生物活性的碎片贡献率，结果显示，影响抑制剂活性的主要分子结构单元为=C-(或-C-)、>C-、>N-、-O-以及二联苯中苯环的位置，苯环在对位时对活性有利。

研究亮点：发展小分子CDK4抑制剂是当前癌症治疗的新研究领域。首次使用QSAR方法对44种海洋Fascaplysin类化合物的生物活性与其分子结构之间的定量关系进行研究，且对模型进行了内部验证和外部验证，计算了基团对活性的贡献大小。

关键词：Fascaplysin；CDK4；分子电性距离矢量；定量结构-活性关系

中图分类号：P 745

文献标志码：A

CDK4是丝氨酸/苏氨酸蛋白激酶家族成员，由催化亚基和调节亚基(周期蛋白)组成，在细胞周期过程中起着重要作用。大量研究表明，一些细胞周期蛋白的失控与一些和肿瘤相关联的CKI(CDKS抑制剂)的缺乏有关，CDK4的失控会启动肿瘤细胞的发生[1]。因此，探索小分子CDK4抑制剂成为当前癌症治疗的新研究领域。

目前天然抗肿瘤药物的研制和开发受到国内外学者们的极大关注，而海洋生物资源在天然药物资源中是保留最完整、来源最丰富、最具有抗肿瘤新药开发潜力的领域[2-3]。

Fascaplysin[4]是1988年从海绵(*Fascaplysinopsis Bergquist* sp.)中分离的一种吲哚生物碱(图1)，具有多种生物活性，而且被证实是一种特异性CDK4抑制剂。虽然它能抑制多种癌细胞的体外增殖，但是它的强毒副作用限制了它的临床使用。

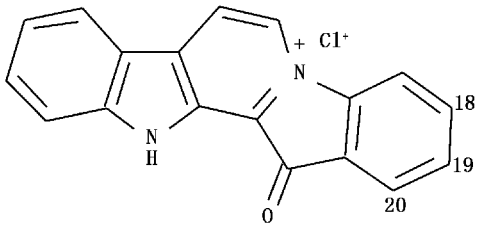

图1　Fascaplysin结构图

Fig.1　Fascaplysin structure

了解化合物的结构对其生物活性的影响是很重要的。QSAR是一种以结构描述子来表征化合物的结构与其生物活性之间关系的统计学方法。因此，有了QSAR模型，就可以通过化合物的结构来预测化合物的活性，为修饰先导化合物，获得新的更好生物活性提供帮助。目前流行的

收稿日期：2011-02-22　　**修回日期：**2011-06-06

基金项目：上海市教育委员会重点学科建设项目(J50702)

作者简介：谢天宝(1983—)，男，硕士研究生，研究方向为定量构效关系。E-mail：xietianbao1983@yahoo.com.cn

通讯作者：印春生，E-mail：csyin@shou.edu.cn

表征方法包括二维拓扑描述子、量化经验描述子、三维分子场描述子等。QSAR 已成为药物研发过程中一种重要工具。

本文对 44 种海洋 Fascaplysin[5] 类 CDK4 抑制剂使用 MEDV-13 和指示描述子联合表征其分子结构，首次建立了 QSAR 模型。将为 CDK4 抑制剂的研究提供另外一种方法，同时也能为设计新的抗肿瘤药物提供理论基础。

1 材料与方法

1.1 MEDV-13 算法

MEDV-13 是由刘树深[6]提出的基于 13 种原子类型和 43 种原子属性的分子距离矢量描述子的简称。MEDV-13 将有机化合物中出现的非氢原子如 C、N、O、S、P、F、Cl、Br、I 的不同原子环境划分为 13 种类型，并且根据原子的元素特性及成键特性将非氢原子细分为 43 种属性类型。MEDV-13[6-7]描述子的计算公式为：

$$x_r = m_{kl} = \sum_{i\in k, j\in l} \frac{q_i q_j}{d_{ij}^2}$$

$$(k,l=1,2,\cdots,13;\ l\geqslant k;\ r=1,2,\cdots,91) \quad (1)$$

式中：k、l 为各非氢原子类型；d_{ij} 为非氢原子 i、j 之间的最短拓扑距离即从原子 i 到原子 j 的各个路径中化学键数加和的最小值；q_i 与 q_j 为原子在实际分子环境中的 E-状态指数。

$$q_i = I_i + \sum_{j\neq i}^{all\ j} \frac{I_i - I_j}{d_{ij}^2} \quad (2)$$

式中：I 为原子固有属性。刘树深[6]在提出 MEDV-13 描述子时对其进行了修改。

$$I = \sqrt{\frac{v}{4}\frac{(2/n)^2\delta^v + 1}{\delta}} \quad (3)$$

式中：v 为原子价电子层的电子数；n 为原子最外层的主量子数；δ^v 和 δ 分别为原子在不同分子环境中构成不同共价键的电子描述符。

1.2 最佳子集回归

在大量的描述子变量中，有很大一部分对化合物生物活性不重要，如何从这些描述子中选择重要的变量并剔除不重要的变量就显得异常重要，最佳子集回归（BSR，又称为 VSMP 法）[8-10]是将描述子变量筛选和多元回归技术结合用于挑选最优描述子变量的方法，与传统的 ASR（All-subsets regression）方法相比，不再以模型的估计能力为目标，而是以 leave-one-out（LOO）交互检验相关系数以及均方根误差为目标函数进行变量筛选的方法。在 LOO 交互检验的基本过程中首先指定一个最优变量数（vn），然后每次抽出一个样本为检验集，以余下的样本作为训练集建立模型，并预测检验集，如此重复至所有样本都被抽出一次为止。

1.3 建模方法

本文以 MEDV-13 和指示描述子联合提取化合物结构信息，运用最佳子集回归法筛选最优描述子变量，最后通过多元线性回归法建立了 44 种海洋 Fascaplysin 类 CDK4 抑制剂的分子结构和活性数据 IC_{50} 之间的定量构效关系模型[5]，并验证了模型的预测能力，结果显示，该模型具有很高的稳定性和预测能力。

2 结果与讨论

通过 VSMP[9]程序，以 MEDV-13 指数为描述子变量（使用自编 matlab 程序计算），在变量 vn = 2,3,4,…,n/5 时，分别建立了 44 个 Fascaplysin 衍生物与它们的抑制活性之间的定量构效关系模型。模型的相关统计量见表 1。

表 1 44 个 Fascaplysin 衍生物最佳子集统计量随 *vn* 变化表

Tab. 1 Some statistics of optimal subset varied with *vn* of 44 Fascaplysin derivatives

vn	q^2	q	RMSEP	r^2	r	RMSEE	最佳子集
2	0.302 0	0.549 5	0.297 1	0.497 3	0.705 2	0.252 1	9 30
3	0.538 7	0.734 0	0.241 5	0.597 9	0.773 2	0.225 5	30 67 92
4	0.535 7	0.731 9	0.242 3	0.609 2	0.780 5	0.222 3	21 22 30 92
5	0.574 0	0.757 6	0.232 1	0.672 0	0.819 8	0.203 6	6 22 30 63 92
6	0.593 1	0.734 2	0.226 8	0.683 7	0.826 9	0.200 0	2 15 22 30 70 92
7	0.602 0	0.775 9	0.224 3	0.691 8	0.831 7	0.197 4	2 15 18 22 30 36 92
8	0.590 8	0.768 6	0.227 5	0.698 3	0.835 6	0.195 3	2 9 15 18 22 30 36 92

注：q 是模型 LOO 检验的预测相关系数；RMSEP 是模型预测均方根误差；r 是模型估计相关系数；RMSEE 是估计均方根误差；vn 为模型优化子集变量数。

从表 1 可以看出,尽管 6 变量时的 r^2 比 8 变量小,但 q^2 更大,因此选择 6 变量为最佳子集变量,6 变量模型为最优模型,模型的标准偏差 s 为 0.227。通过分析模型在 LOO 检验中预测得到的 pIC_{50}($pIC_{50} = -\log(IC_{50} \times 10^{-6})$)值,我们发现有 3 个样本的 pIC_{50} 计算值和原先的实验值之间误差均超过 2 倍偏差即 0.452(图 2),样本 2 误差为 0.467 3(估计时误差为 0.421 8),样本 13 误差为 -0.541 9(估计时误差为 -0.511 6),样本 30 误差为 -0.547 2(估计时误差为 -0.511 6)。这些误差值与其它的样本的误差值相比明显偏大,这 3 个样本应视为离群值,建模时删去,否则会影响模型的稳定性。

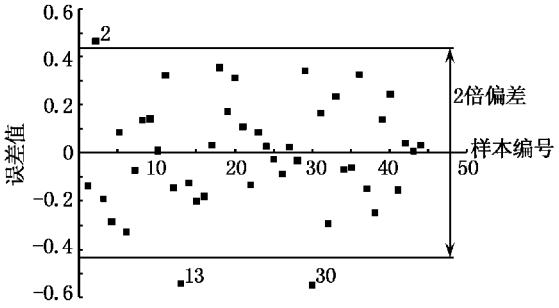

图 2　样本误差图

Fig. 2　Error of samples

模型的验证分为外部验证(外部测试集)和内部验证(LOO 检验),虽然 LOO 检验相关系数 q^2 常常作为判断模型好坏的一个重要指标(如判定是否有离群值及内部预测能力的好坏),但它并不能保证所建立的模型具有良好的对外部未知样本的预测能力,有时会带来较乐观的结果。一个好的 QSAR 模型不仅应该具有良好的校正能力,还必须同时具有对外部样本的良好预测能力,因此,为了获得一个具有良好外部预测能力的模型,我们将 41 个样本分为训练集和外部测试集。首先,将 41 个样本的生物活性值按大小排序,然后均匀提取出训练集(31 个样本)和测试集(10 个样本)。同样使用 VSMP 程序,得到的统计参数量见表 2。

显然,由表 2 知道,$vn = 6$ 时,有最佳模型,该模型具有良好的估计能力和 LOO 预测能力和较高的外部样本预测相关系数 R。模型等式如下

$$pIC_{50} = 0.1373x_2 - 0.273x_9 - 0.1293x_{15} + 0.0853x_{26} + 21.0691x_{67} - 0.2604x_{92} + 8.2587$$

$n = 31$, $r^2 = 0.8535$, RMSEE = 0.1387, F = 23.3025 (估计)

$q^2 = 0.7724$, RMSEP = 0.1729 (LOO 预测)

$R^2 = 0.6635$ (外部测试集预测)

表 2　31 个 Fascaplysin 衍生物最佳子集统计量随 *vn* 变化表

Tab. 2　Some statistics of optimal subset varied with *vn* of 31 Fascaplysin derivatives

vn	q^2	q	RMSEP	r^2	r	RMSEE	最佳子集
2	0.624 4	0.790 2	0.222 1	0.662 5	0.831 9	0.210 5	30 67
3	0.704 8	0.839 5	0.196 9	0.780 6	0.883 5	0.169 8	2 22 30
4	0.688 4	0.829 7	0.192 3	0.815 3	0.902 9	0.155 8	21 22 30 92
5	0.758 4	0.870 9	0.178 1	0.839 3	0.916 1	0.145 3	21 22 30 67 92
6	0.772 4	0.878 9	0.172 9	0.853 5	0.923 9	0.138 7	2 9 15 26 67 92

该模型的预测值见表 3 (包括对建模前删除的样本 2、13、30 的预测)。其中,x_{92} 描述子是指示描述子,表示二联苯中苯环的相互位置(18、19、20 号 C 原子位置连接苯环时 x_{92} 值分别为 0.5、1.0、1.5)。依据分子电性距离矢量相关理论,与 x_2、x_9、x_{15}、x_{26}、x_{67} 这 5 个 MEDV-13 描述子变量相关的分子基团分别为-C、-C-、O = 、> C-、> N-。指示描述子 x_{92} 的系数是负值,表示该描述子与生物活性 pIC_{50} 负相关,二联苯的对位(18 号 C)位置连接苯环时对增加活性有利。x_2 系数为正,暗示二联苯上连接叔丁基有利于活性的提高,正如第 5(pIC_{50} = 4.950 8)和 11 号(pIC_{50} = 5.036 2)样本具有较高的活性。活性预测值和实验值关系见图 3。

为了进一步说明各个基团对活性影响的程度大小,我们使用碎片贡献率对描述子进行分析。计算碎片贡献率的方式如下[11-13]:

$$\Psi_r(i) = a_i \overline{TI}_i \quad (4)$$

$$\Psi_f(i) = r^2 |\Psi_r(i)| \div \sum |\Psi_r(i)| \times 100\% \quad (5)$$

式中:a_i 和 $\overline{TI}_i$ 分别为系数和模型中第 i 个拓扑描述子的平均值;r^2 为模型相关系数的平方;Ψ_f 为碎片贡献。

表3 模型活性观察值(pIC_{50})、实验值

Tab.3 Activities (pIC_{50}) observed, calculated by the optimal model

No.	化合物	观察值	预测值	No.	化合物	观察值	预测值
1	$C_{18}H_{11}N_2O$	6.259 6	6.229 6	23	$C_{24}H_{19}N_2OF$	4.721 2	4.559 0
2	$C_{24}H_{22}N_2O$	5.221 8	4.528 3	24	$C_{25}H_{22}N_2O$	4.585 0	4.517 1*
3	$C_{24}H_{21}N_2OF$	4.677 8	4.774*	25	$C_{24}H_{19}N_2OF$	4.602 1	4.639 1
4	$C_{25}H_{23}N_2O$	4.619 8	4.772 4*	26	$C_{24}H_{19}N_2OF$	4.568 6	4.664 2
5	$C_{28}H_{31}N_2O$	4.950 8	4.758 9	27	$C_{24}H_{19}N_2OF$	4.744 7	4.597 0
6	$C_{25}H_{23}N_2O_2$	4.309 8	4.402 1	28	$C_{25}H_{22}N_2O$	4.619 8	4.618 0
7	$C_{30}H_{26}N_2O$	4.455 9	4.424 9*	29	$C_{25}H_{22}N_2O$	4.958 6	4.588 5*
8	$C_{23}H_{20}N_2O$	4.795 9	4.573 8	30	$C_{25}H_{22}N_2O$	4.124 9	4.556 2
9	$C_{23}H_{19}N_2OF$	4.823 9	4.889 8*	31	$C_{25}H_{22}N_2O_2$	4.481 5	4.444 9
10	$C_{24}H_{22}N_2O$	4.744 7	4.884 9*	32	$C_{25}H_{22}N_2O_2$	4.102 4	4.597 5*
11	$C_{27}H_{28}N_2O$	5.036 2	4.922 5	33	$C_{24}H_{24}N_2OF$	4.920 8	4.769 0
12	$C_{24}H_{22}N_2O_2$	4.318 8	4.536 3	34	$C_{24}H_{24}N_2OF$	4.677 8	4.819 2
13	$C_{23}H_{20}N_2O$	4.065 5	4.493 1	35	$C_{24}H_{24}N_2OF$	4.769 6	4.739 5
14	$C_{23}H_{19}N_2OF$	4.420 2	4.718 5	36	$C_{25}H_{22}N_2O$	5.045 8	4.732 5
15	$C_{27}H_{29}N_2O$	4.420 2	4.769 5	37	$C_{25}H_{22}N_2O$	4.602 1	4.781 2
16	$C_{24}H_{23}N_{O2}$	4.102 4	4.247 3	38	$C_{25}H_{22}N_2O$	4.494 9	4.777 2
17	$C_{24}H_{23}N_2O_2$	4.275 7	4.199 4	39	$C_{25}H_{22}N_2O_2$	4.552 8	4.570 2
18	$C_{24}H_{23}N_2O_2$	4.585 0	4.264 8	40	$C_{25}H_{22}N_2O_2$	4.657 6	4.676 7
19	$C_{23}H_{20}N_2OF$	4.699 0	4.672 2	41	$C_{23}H_{21}N_3O$	4.585 0	4.654 4
20	$C_{24}H_{23}N_2O$	4.823 9	4.562 9	42	$C_{22}H_{19}N_3O$	4.522 9	4.686 6*
21	$C_{24}H_{19}N_2OF$	4.638 3	4.557 2	43	$C_{23}H_{20}N_3O$	4.408 9	4.441 4*
22	$C_{24}H_{19}N_2OF$	4.455 9	4.617 3	44	$C_{23}H_{20}N_3O$	4.443 7	4.542 4

注：*为预测值。

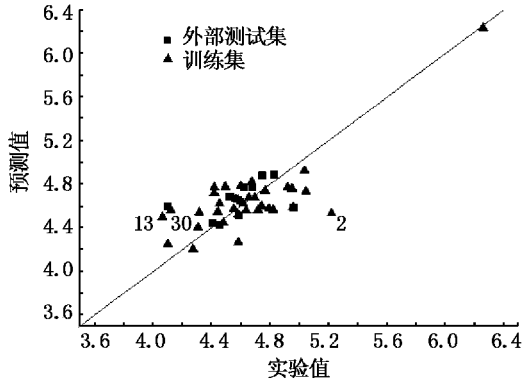

图3 44Fascaplysin类CDK4抑制剂的活性值(pIC_{50})与预测值相关关系

Fig.3 The relationship of the predicted values and observed values (pIC_{50})

从表4中看出，象征着C原子的x_{15}(-C-或=C-和 >C-)描述子对化合物活性有着主要贡献，达到66.02%，这是因为-C-、=C-、>C-基团是化合物的主要组成结构单元，分析x_{15}描述子数据知道，各个化合物的值很接近。x_{92}反映化合物结构中二联苯的苯环相互位置，它的贡献率为3.81%，意味着苯环在不同的位置，对活性有重要影响。描述子x_9和x_{67}中都含有O原子，表示亲水基团对化合物的活性也有一定影响。最佳描述子中没有包含卤代元素信息的描述子，暗示分子结构中卤代元素F对化合物的贡献微乎其微，x_2中包含有甲基的信息，说明甲基取代基有助于提高化合物的活性。

表4 描述子的碎片贡献率(Ψ_r、Ψ_f)

Tab.4 The relative and fraction contribution (Ψ_r and Ψ_f) of individual descriptor

描述子	基团	Ψ_r 和 Ψ_f
x_2	-C and -C-、=C-	0.325 2(6.06%)
x_9	-C-、=C- and O=	-0.076 3(1.43%)
x_{15}	-C-、=C-and >C-	-3.545 0(66.02%)
x_{26}	>C- and >C-	0.410 6(7.65%)
x_{67}	>N- and -O-	0.021 7(0.40%)
x_{92}		-0.204 4(3.81%)

3 结论

以分子电性距离矢量(MEDV-13)及指示描述子对44种有机化合物的分子结构与抑制活性之间的定量关系进行研究，建立了一个六变量QSAR模型，通过内部样本的LOO交互检验和外

部样本的预测验证,结果表明,所构建模型具有良好的估计能力和预测能力,MEDV-13 和指示描述子很好地描述了海洋 Fascaplysin 类 CDK4 抑制剂的分子结构。对最佳子集变量组合进行分析,得出影响化合物活性的主要分子结构单元为 =C-(-C-)、>C-、>N-、-O-以及二联苯中苯环的位置,苯环在对位时对活性有利。进一步分析各描述子的碎片贡献率表明:虽然 x_{15}描述子所表示的-C-、=C-和 >C-基团之间的相互作用对 CDK4 抑制活性的影响起主导作用,甲基取代基有助于提高化合物的活性,亲水基团对化合物的活性也有一定影响,卤代元素 F 对化合物的贡献微乎其微。

参考文献:

[1] 杨书梅. CDK4 酶抑制剂的计算机辅助分子设计研究[D]. 浙江:浙江大学,2007:3.

[2] 卢晓玲,郑燕玲,陈海敏,等. Fascaplysin 通过诱导细胞凋亡抑制 HeLa 细胞体外增殖[J]. 药学学报,2009,44(9):980 - 986.

[3] SCHWARTSMANN G, RATAIN M J, CRAGG G M, et al. Anticancer drug discovery and development throughout the world[J]. Journal of Clinical Oncology, 2002, 20(18): 47 - 59.

[4] ROLL D M, IRELAND C M, LU H S M, et al. Fascaplysin, an unusual antimicrobial pigment from the marine sponge Fascaplysinopsis sp[J]. The Journal of Organic Chemistry, 1988, 53(14):3276 - 3278.

[5] JENKINS P R, WILSON J, EMMERSON D, et al. Design, synthesis and biological evaluation of new tryptamine and tetrahydro-β-carboline-based selective inhibitors of CDK4 [J]. Bioorganic and Medicinal Chemistry, 2008, 16(16): 7728 - 7739.

[6] 刘树深. 有机物分子电性距离矢量表征及其应用[M]. 北京:高等教育出版社,2005:35 - 46.

[7] 胡儒柱. 几类与农药有关的有机物定量构效关系研究[D]. 广西:桂林理工大学,2009:4 - 5.

[8] CUI S H, YANG J, LIU S S, et al. Predicting bioconcentration factor values of organic pollutants based on medv descriptors derived QSARs[J]. Science in China Series B: Chemistry, 2007, 50(5):587 - 592.

[9] LIU S S, LIU H L, YIN C S, et al. VSMP: A Novel Variable Selection and Modeling Method Based on the Prediction [J]. Journal of Chemical Information and Computer Science, 2003, 43(3):964 - 969.

[10] QIN L T, LIU S H, LIU H L, et al. A new predictive model for the bioconcentration factors of polychlorinated biphenyls (PCBs) based on the molecular electronegativity distance vector (MEDV)[J]. Chemosphere, 2008, 70(9):1577 - 1587.

[11] LU C H, GUO W M, YIN C S. Quantitative structure-retention relationship study of the gas chromatographic retention indices of saturated esters on different stationary phases using novel topological indices[J]. Analytica Chimica Acta, 2006, 561(1/2):96 - 102.

[12] HU R Z, YIN C S, WANG Y, et al. QSPR study on GC relative retention time of organic pesticides on different chromatographic columns[J]. Journal of Separation Science, 2008, 31(13):2434 - 2443.

QSAR study on CDK4 inhibitors of marine Fascaplysin analogues based upon MEDV-13

XIE Tian-bao, YIN Chun-sheng, YANG Hong
(*College of Marine Sciences, Shanghai Ocean University, Shanghai* 201306, *China*)

Abstract: The unique marine ecosystem is more and more recognized as a source of extremely potent natural anticancer agents. The Molecular Electronegativity Distance Vector (MEDV-13) was used to describe the chemical structure of 44 Fascaplysin analogues, and resulting in a quantitative structure-property relationship (QSAR) model of six parameters on IC_{50} for CDk4 using variable selection and modeling based on prediction (VSMP), which was validated ty LOO method and external test set were respectively. The obtained model shows good estimation ability and strong predictive power for the external samples with a calibrated correlation coefficient of $r = 0.923\ 9$ and LOO validated correlation coefficient of $q = 0.878\ 9$. Last, the influences of each index on activity were calculated by its relative or fraction contribution. The results revealed that the main structural factors influencing the bioactivities are =C-(or-C-)、>C-、>N-、-O- and benzene site in biphenyl, and para-biphenyl is advantageous to activity.

Key words: Fascaplysin; CDK4; MEDV-13; QSAR