支持向量机在构建白介素-1β转化酶抑制剂药效团模型中的应用*

□杨 晔 张燕玲 乔延江**

（北京中医药大学中药学院 北京 100102）

摘 要：以24个作用于人体外围血单核细胞药理模型的白介素-1β转化酶抑制剂作为研究对象，计算了其表征分子的拓扑、电子、几何结构等物理化学性质的1209个分子描述符，用CfsSubsetEval评价方法和BestFirst-D1-N5搜索方法筛选描述符，用Kennard-Stone方法选择训练集和测试集。分别采用支持向量机、决策树、贝叶斯网络、人工神经网络等机器学习方法建立分类预测模型并使用Catalyst/HipHop系统建立药效团模型。结果表明支持向量机优于其他分类模型，正、负样本的预测正确率均达到100%。最优药效团模型具有5个特征：2个疏水基团、2个脂性氢键受体、1个氢键给体；以此药效团进行中药数据库筛选得到384个候选白介素-1β转化酶抑制剂。利用支持向量机建立的分类预测模型对候选化合物的活性进行了预测，其中高活性化合物占96.6%，表明白介素-1β转化酶抑制剂药效团模型较准确地反映了高活性化合物的公共特征。该模型的建立有助于从中草药筛选新型白介素-1β转化酶抑制剂。

关键词：白介素-1β转化酶抑制剂 分子描述符 支持向量机 药效团 HipHop

白介素-1β转化酶（Interlenkin-1βcoverting Enzyme，ICE），是剪切IL-1β和IL-18前体的特异有效蛋白酶[1]。这两种细胞因子与多种炎症过度导致的疾病有关，如病毒性感染、肿瘤、自身免疫性疾病、风湿性疾病、心肌梗塞以及神经退行性疾病等。因此，白介素-1β转化酶已经成为治疗炎症性相关疾病的重要靶点[2]。

Catalyst/HipHop系统可用于构建定性药效团模型，其构建药效团模型时，只需满足分子结构多样性，不需要分子的活性值，适用范围广，但是难以评价模型的可靠性及其命中化合物活性的高低。支持向量机（Support Vector Machines，SVM）是20世纪90年代由Vapnik提出的基于统计学习理论，以结构风险最小化原则提出的机器学习方法[3]，具有所需样本量少、建模方便、计算简单、学习训练时间短、泛化能力强等优点。近年来，在药物分析、医学检验、工业生

收稿日期：2009-08-12

修回日期：2009-09-10

* 科技部国家“973”计划（2005CB523401）：组分配伍与饮片配伍的相关性研究，负责人：郑虎占；国家“973”计划（2006CB504703）：寒热药性的内在规律及共同属性研究，负责人：乔延江；中医药行业科研专项（200707010）：针对病毒性疾病的中药活性发现关键技术研究，负责人：朱晓新。

** 联系人：乔延江，本刊编委，教授，博士生导师，北京中医药大学副校长，主要研究方向：中药信息学研究，Tel：010-84738620，E-mail：yjqiao@263.net。

产、房地产投资等领域得到了广泛的应用[4~5]。本文尝试将 SVM 与 HipHop 方法结合，验证药效团模型的可靠性，评价命中化合物活性的高低，进一步阐释药物作用机理。

本文以 24 个 ICE 抑制剂作为研究对象，分别利用 SVM、决策树、贝叶斯网络、人工神经网络等机器学习方法建立了分类预测模型，并用 Catalyst/HipHop 系统建立药效团模型。以药效团模型搜索中草药数据库，用 SVM 建立的模型预测筛选所得中药成分活性的高低，验证药效团模型的可靠性，以期从中草药数据库中筛选出具有较高生物活性的 ICE 抑制剂。

一、材料与方法

1. 化合物的活性数据

试验样本来源于 MDDR（MDL Drug Data Report:Version2007.2）数据库中作用于人外围血单核细胞（Human Peripheral Blood Mononcytes，PBMC）药理模型的 24 个有抑制 ICE 活性的化合物。虽然这些分子有活性值，但由于样本分子活性值不能平均分布于 4 个数量级，故本文未用 Catalyst/HupoGen 系统做定量模型的计算。根据本文分子的活性值，将 $lg1/IC_{50} \geqslant 0$ 的化合物归入活性类，共 17 个化合物，标记为+；而将 $lg1/IC_{50} < 0$ 的化合物归入无活性类，共 7 个化合物，标记为-。本文采用 Kennard-Stone（KS）方法[6]选择训练集和测试集，KS 法可以保证训练集中样本按空间距离分布均匀，使训练集具有较好的代表性。选取了 2/3 作为训练集，1/3 作为测试集，训练集包含 11 个活性分子，5 个非活性分子；测试集包含 6 个活性分子，2 个非活性分子。图 1、图 2 分别列出了选取的训练集和测试集化合物结构、活性以及分类的试验结果。本文中所有分子的稳定几何结构均由 HyperChem7（http://www.hyper.com）中 MM+力场优化得到。

2. 分类预测模型的建立

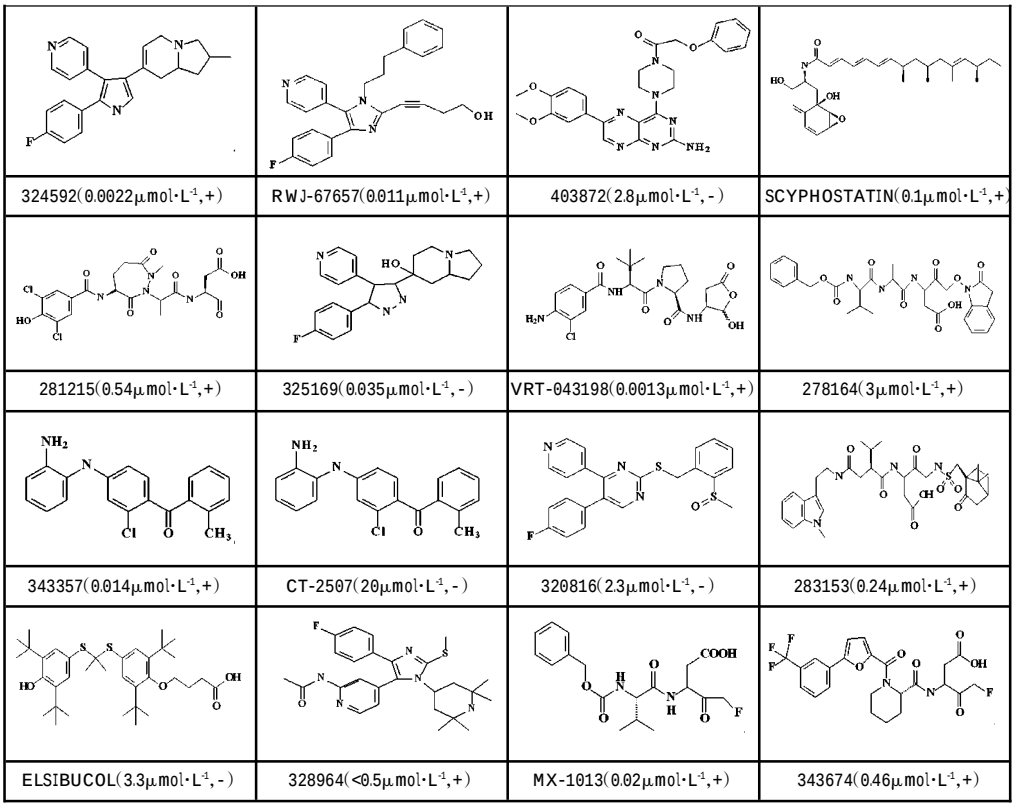

图 1　训练集化合物结构、活性及其分类

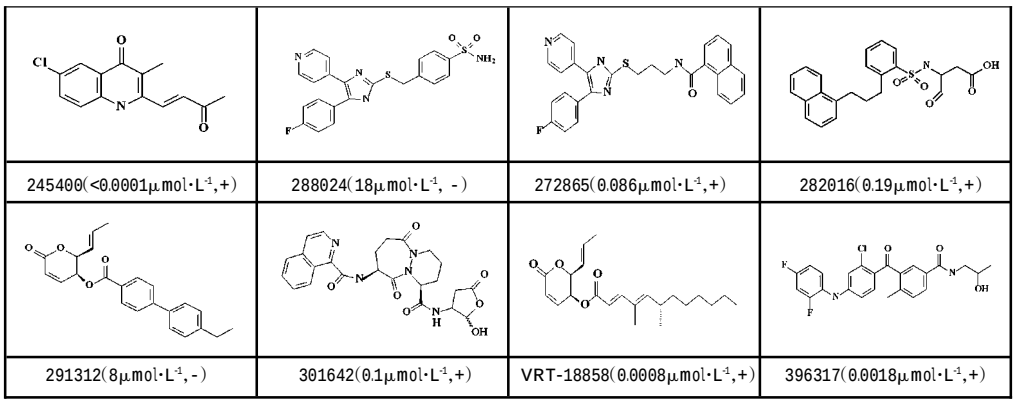

图 2　测试集化合物结构、活性及其分类

(1)筛选描述符。

分子描述符是化合物的结构和物理化学性质的表征，化合物的生物活性往往与表征分子关键性质的描述符相关。本文利用 Dragon 软件（Software Vision 2.1）计算了包括分子拓扑、电子、几何等结构信息在内的 1209 个描述符。将这些描述符进行相关性分析，去除相关性≥0.95 的描述符，描述符缩减为 624 个，主要包括拓扑、BCUT、2D 自相关、RDF、3D-MoRSE、GETAWAY、WHIM 等[7]描述符。

为了消除不同量纲对变量的影响，对所得的分子描述符数据进行标准化处理，即对每个变量先减去样本平均值，再除以样本方差。为了使所建立的模型不产生过拟合，使所建立的模型具有较好的预测能力，本文采用 Weka（Version 3.4.7）机器学习平台中的 CfsSubsetEval 评价方法和 BestFirst-D1-N5 搜索方法[8]，通过十折交叉验证筛选得到与类相关性最大、相互之间关联性较低的描述符。

(2)支持向量机。

SVM 是一种模式识别和分类工具，属于有监督的学习方法。其基本思想是针对二类分类问题：如果训练集是线性可分的，则在高维空间中寻找一个最优分类超平面来实现对样本空间的划分，所得到的超平面应当满足分类间最大化原则；如果训练集线性不可分，则利用核函数映射，将输入向量投影到一更高维空间，划分正负样本[9]。目前，常用的核函数主要包括线性、多项式、RBF 以及 Sigmoid 等多种形式。(C，γ)是本文使用的 RBF 核函数最重要的参数对。其中 C 为误差惩罚参数，C 越小，惩罚越小，从而使训练误差变大，系统的泛化能力越差，C 太大，与置信范围相关的权重相应变小，系统的泛化能力也会下降；γ 控制着径向基函数的振幅，针对具体的分类情况，这两个参数应有不同的取值。关于 SVM 的原理与算法的详细描述已有文献报道[10]，本文不作赘述。

本文 SVM 算法采用台湾大学林智仁（Lin Chin-Jen）提供的网络共享算法 Libsvm（http://www.csie.ntu.edu.tw/~cjlin/libsvm/）。

(3)其他机器学习方法。

本文使用新西兰怀卡托大学开发的 Weka（version 3.4.7）机器平台[11]建立的决策树（Decision Tree）J48 算法、贝叶斯网络（Bayesian Network）K2 算法以及反向传播的神经网络（Artificial Neural Network）算法，分别建立 ICE 抑制剂活性分类预测模型，参数均采用默认值。

3. 药效团模型的建立

(1)训练集的选择及构象分析。

训练集的选择同用于建立分类预测模型的分子。采用 Accelrys 公司的 Catalyst4.11 软件包进行药效团模型的计算[12]，分子的三维结构在 Catalyst 中的 View compound 模块中完成，采用 Catalyst 中的 2D Beautify 对化合物的二维结构进行优化，然后利用 Generate Standard 3D 来调整化合物的键长和键角，最后用 3D Minimize 寻找其能量最低的三维空间构象。为了最大限度覆盖每个分子与受体结合时各种构象，采用 Best Quality 模式，能量间隔为83.74kJ·mol^{-1}，最大构象数目设为 250 个，对所构建的分子进行处理，得到一系列的低能构象。

(2)模型计算。

Catalyst/HipHop 是计算药效团模型的常用系统，能根据训练集中活性化合物的三维空间排布产生共同的化学特征[13]。在构建药效团模型时，将活性最高的两个化合物 324592 和 VRT-043198 的 Principal 值设为 2，MaxOmitFeat 值设为 0，活性最低的化合物 CT-2507 的 Principal 值设为 0，MaxOmitFeat 值设为 2，其他化合物的 Principal 和 MaxOmitFeat 值均设为 1。Misses，FeatureMisses，CompleteMisses 分别设为 4、4、3，Spacing 值设为 200。除以上参数外，其他参数均采用默认值。

(3)数据库搜索。

利用 Catalyst 系统中的 Fast Flexible Database Search 模式对存储多个构象的三维数据库进行柔性搜索，对搜索结果命中的化合物采用 Best Fit/Compare 模式进行化合物多构象与药效团匹配，根据匹配程度预测化合物活性[14]。

二、结果与讨论

1. 分类预测模型的建立和比较

为了考察筛选描述符的合理性，采用 CfsSubsetEval 评价方法和 BestFirst-D1-N5 搜索方法筛选用于各类机器学习方法模型的 624 个描述符。其中，CfsSubsetEval 逐一评估每个描述符的预测能力和它们之间的重复程度，之后挑选那些与类有高度关联但相互之间关联程度却较低的描述符；BestFirst-D1-N5 通过返回进行贪心式爬山搜索，它可以从一个空的

描述符集正向搜索，或从一个满集反向搜索，或从中间的一个点开始并向前后两个方向，通过考虑所有可能的单个描述符加入及删除进行搜索。筛选结果使描述符减少到 6 个，主要包括拓扑描述符，二维自相关描述符，WHIM 描述符以及 GETAWAY 描述符[15~18]，所选的描述符见表 1。

为了评价各机器学习方法对训练集的拟合能力以及对测试集的预测能力，用敏感性（SE）、专一性（SP）以及准确性（Q）来评价模型对正样本，负样本及总体的预测正确率，其定义分别为：SE=TP/(TP+FN)，SP=TN/(TN+FP)，Q=(TP+TN)/(TP+TN+FP+FN)，其中 TP 和 TN 分别代表对测试集中预测正确的正样本数和负样本数。FP 和 FN 分别代表预测错误的负样本数和正样本数。各类机器学习方法分类建模的结果见表 2。

从表 2 可以看出，贝叶斯网络及神经网络算法均对负样本有很高的预测正确率，均达到 100%，但决策树对负样本的预测正确性只有 50%。决策树、贝叶斯网络、神经网络算法对正样本的正确率均为 83.33%，预测能力不太理想。与之相比，SVM 对正、负样本的预测正确率均最高，均达到 100%。其原因为 SVM 采用逐一交互检验的方法确定 C 和 γ，它的原理是把数据训练集分为 n 份，用 n-1 份来预测剩下的一份数据，C 和 γ 作为变量，均方差 RMS 作为目标函数，RMS 最小时的 C 和 γ 分别为 128 和 0.0078125，这样做既可以寻找到最适于建立模型的参数，又防止了模型对于训练集的过拟合。因此，与其他分类预测方法比较，SVM 方法能较好地解决了小样本的学习分类问题，适用于 ICE 抑制剂的活性分类预测模型的建立，具有较优的预测结果。

2. 药效团模型的建立和优化

利用 16 个训练集分子计算得到 10 个得分较高的药效团模型，得分值 132.399~153.488。主要的药效特征包括氢键受体（HB Acceptor，HBA）、脂性氢键受体（HB Acceptor lipid，HBALi）、疏水基团（Hydrophobic，H）、芳香疏水基团（Hydrophobic aromatic，Har）和氢键给体（HB Donor，D）等。对这 10 个药效团模型进行聚类分析，得到 4 类药效团模型特征，其中第 1 类为 1、3、4、6 号药效团，其药效特征为 HHDAliAli；第 2 类为 2、5、7 号药效团，其药效特征为 HHDAliA；第 3 类为 8、10 号药效团，其药效特征为 HHAliAli；第 4 类为 9 号药效团，其药效特征为 HarHAliAli。

MDDR 收录了 177981 个已报道活性化合物组成数据库 D；搜索 MDDR 中具有抑制 ICE 活性的化合物 223 个组成数据库 A；以药效团为提问结构搜索数据库 D，命中化合物为 H_t；以药效团作为提问结构搜索数据库 A，命中的活性化合物为 H_a；A% 为药效团命中的活性化合物所占比例，即 $A\%=(H_a/A)\times100$；E 为辨识有效性指数，反映了药效团模型区分活性和非活性化合物的能力，E 值越大，说明模型区分活性化合物和非活性化合物的能力越强，即 $E=\frac{H_a/H_t}{A/D}=\frac{H_a\times D}{H_t\times A}$，详细介绍请参考文献[19]。

从 1、2 类药效团中分别选择两个药效团，即 1、3、2、5 号药效团，分别修改代表药效特征空间允许误差值（Tolerance），分别将这 4 个药效团的 Tolerance 值改为默认值的 0.9、0.8、0.7 倍。搜索结果见表 3，其中，1—$1t_{0.7}$ 分别表示 1 号药效团及各药效特征的 Tolerance 改为原来的 0.9 倍、0.8 倍、0.7 倍，得到的药效团模型，同理 2，3，5 号药效团。

综合比较各类优化结果，选取辨识有效性指数 E 及活性化合物有效命中率分别为 3.80 和 22.87%，Tolerance 为默认值的 0.8 倍的 3 号药效团作为初选

表 1　筛选描述符的结果

No.	Abbreviation	Full name
1	X4Av	Average valence connectivity index chi-4
2	MTAS8m	Moran autocorrelation-lag 8/weighted by atomic masses
3	MTAS1p	Moran autocorrelation-lag 1/weighted by atomic polarizabilities
4	GATS2v	Geary autocorrelation-lag2/weighted by atomic van der Waals volumes
5	E3u	3rd component accessibility directional WHIM index/unweighted
6	R1v	R autocorrelation of lag 1/weighted by atomic van der Waals volumes

表 2　不同机器学习方法分类结果比较

Method	No. of descriptors	SE(%)	SP(%)	Q(%)
SVM	6	100	100	100
Decision Tree	6	83.33	50	71.43
Bayesian Network	6	83.33	100	85.71
Artificial Neural Network	6	83.33	100	85.71

药效团的模型。各特征之间的距离允许误差为±1Å，将各特征之间的距离允许误差修改为±0.8Å。在修改各特征之间的距离允许误差（Distance Tolerance）得到的搜索结果中，活性化合物命中率A%相差较小，经比较，最终确立的最优药效团模型具有5个特征，包括2个疏水基团、2个脂性氢键受体、1个氢键给体，其辨识有效性指数E及活性化合物命中率A%分别为5.71和18.83%，H1-Ali1之间距离为e±0.8Å，药效团模型见图3。

3. 药效团模型可靠性验证

利用所得最优药效团模型，搜索中药化学数据库（Traditional Chinese Medicine Database, TCMD 2005），命中384个活性化合物。利用SVM建立的分类预测模型预测筛选的中药成分活性高低，结果表明，有371个活性化合物属于高活性抑制剂，命中率高达96.6%，说明药效团模型合理有效，反映了高活性化合物的公共特征。筛选有3项以上不符合利平斯基（Lipinsk）规则即类药5规则[20]的中药有效成分，最后确定了148个中药有效成分。经文献检索，这148个中药有效成分中，抗肿瘤、抗病毒类中药占69%，代表药有紫杉、绞股蓝、苦楝皮等。行气活血类中药占14%，代表药有合欢皮、蜘蛛香、远志等。剩余的17%成分分别具有抗炎、祛风湿、补益气血等功效。临床研究表明，ICE抑制剂主治病毒性感染、肿瘤、神经退行性疾病、风湿性疾病、心肌梗塞、炎症反应、自身免疫性疾病等，与本试验命中的中药有效成分相符合，进一步证实了本试验结果的可靠性。

表3 1、2、3、5号药效团模型优化后搜索结果

No.	Feature	D[a]	A[b]	Ht[c]	Ha[d]	A%[e]	E[f]
1	H H D Ali Ali	177981	223	42712	97	43.50	1.81
$1t_{0.9}$	H H D Ali Ali	177981	223	26057	75	33.63	2.30
$1t_{0.8}$	H H D Ali Ali	177981	223	12414	50	22.42	3.21
$1t_{0.7}$	H H D Ali Ali	177981	223	4160	22	9.87	4.22
2	H H D H A	177981	223	41478	95	42.60	1.83
$2t_{0.9}$	H H D H A	177981	223	25298	73	32.74	2.30
$2t_{0.8}$	H H D H A	177981	223	12036	49	21.97	3.25
$2t_{0.7}$	H H D H A	177981	223	3908	20	8.97	4.08
3	H H D Ali Ali	177981	223	42138	93	41.70	1.76
$3t_{0.9}$	H H D Ali Ali	177981	223	27214	73	32.74	2.14
$3t_{0.8}$	H H D Ali Ali	177981	223	10713	51	22.87	3.80
$3t_{0.7}$	H H D Ali Ali	177981	223	4046	20	8.97	3.56
5	H H D H A	177981	223	47423	103	46.19	1.73
$5t_{0.9}$	H H D H A	177981	223	29577	82	36.77	2.21
$5t_{0.8}$	H H D H A	177981	223	14250	60	26.91	3.36
$5t_{0.7}$	H H D H A	177981	223	4566	23	10.31	4.02

a)D: No. of compounds in database; b)A: No. of active compounds; c)Ht: No. of hits; d)Ha: No. of hits in active compounds; e)A%: Percent ratio of the actives in the hit list; f) E: Enrichment。

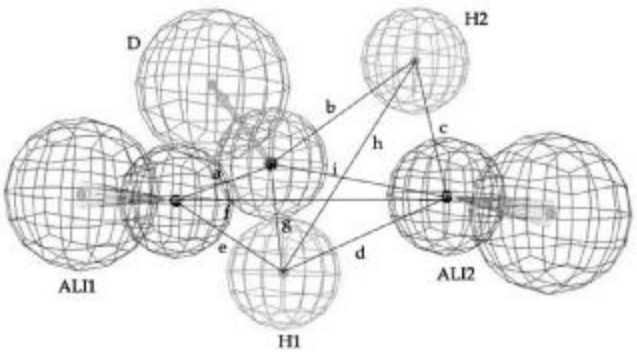

图3 最优药效团模型

药效团和机器学习方法均可以建立预测模型，搜寻小分子三维结构数据库，确定有活性的先导化合物。然而，定性的药效团模型无法预测命中化合物活性的高低，分类预测模型无法阐释与配体结合的靶点信息。两者结合，不仅可以根据重要活性的原子和基团空间关系，反推出与之结合的靶点的立体形状、结构、性质等信息，得到虚拟的靶点模型，阐释药物作用机理，还可以进一步验证模型的可靠性，确定命中化合物活性的高低。

三、结 论

本文利用相同的训练集分子分别构建了分类预测模型和药效团模型。其中，各类机器学习方法构建的分类预测模型中，SVM构建的活性预测模型对正、负样本的预测正确率均达到100%。采用Catalyst/HipHop系统构建ICE抑制剂药效团，具有2个疏水基团，2个脂性氢键受体和1个氢键给体。利用药效团模型进行中药化学数据库搜索结果，经SVM建立的分类预测模型验证，说明药效团模型合理有效，反映了高活性化合物的公共特征，并筛选得一系列有该活性的化合物，为新型ICE抑制剂的开发及药物作用机理阐释提供科学依据。

参考文献

1 David, G. P., Patricia, M., Ronald, L., et al. Identification and Characterization of a Novel Class of Interleukin-1 Post-Translational Processing Inhibitors. Pharmacology and Experimental Therapeutics, 2001, 299(1):187.

2 马学琴,张亚辉,周忠良,等. 炎症 Caspase 与相关疾病. 生命科学, 2006, 18(5):454~455.

3 Vapinkvn. The Nature of Statistical Learning theory. New York: Springer Verlag, 2000.

4 Xue, Y., Yap, C. W., Sun, L. Z., et al. In Silico Prediction of Pregnane X Receptor Activators by Machine Learning Approache. Chem. Inf. Comput. Sci, 2004, 44:1497.

5 Burbidge, R., Trotter, M., Buxton, B., et al. Drug design by machine learning: support vector machines for pharmaceutical data analysis. Comput. Chem, 2001, 26:5.

6 Galvão, R., Araujo, M., José, G., et al. A method for calibration and validation subset partitioning. Talanta, 2005, 67(4):736~740.

7 Adam, F., Lingyi, Z., Harshinder, S. QSAR Study of Skin Sensitization Using Local Lymph Node Assay Data. Int. J. Mol. Sci., 2004, 5:56~66.

8 董琳,邱泉,于晓峰,等译. 数据挖掘: 实用机器学习技术. 北京: 机械工业出版社, 2006:282~283.

9 冯雪松, 刘雅茹, 王大成, 等. 支持向量回归-紫外分光光度法用于测量小儿氨酚匹林咖啡因片含量的方法研究. 广东药学院学报, 2006,22(1):44.

10 陈晓梅,饶含兵,黄文丽,等. 机器学习方法用于二氢叶酸还原酶抑制剂的活性预测.高等学校化学学报,2007,28(11):2174~2175.

11 董琳,邱泉,于晓峰,等译. 数据挖掘: 实用机器学习技术. 北京: 机械工业出版社, 2006:269~277.

12 鲍红娟, 张燕玲, 乔延江. HMG-CoA 还原酶抑制剂三维药效团的构建. 物理化学学报, 2008, 24(2):302.

13 Yasuhisa Kurogi, Osman F. Güner. Pharmacophore Modeling and Three-dimensional Database Searching for Drug Design Using Catalyst. Current Medicinal Chemistry, 2001, 8(9):1039~1040.

14 鲍红娟, 张燕玲, 乔延江. 5-HT3 受体拮抗剂药效团模型的构建. 高等学校化学学报, 2008, 29(6):1128.

15 Prasanna, S., Doerksen, R. Topological Polar Surface Area: A Useful Descriptor in 2D-QSAR. Curr Med Chem, 2009, 16(1):21~41.

16 Saíz-Urra, L., González, M., Teijeira, M. 2D-autocorrelation descriptors for predicting cytotoxicity of naphthoquinone ester derivatives against oral human epidermoid carcinoma. Bioorg Med Chem, 2007, 15 (10): 3565~3571.

17 González, M., Suárez, P., Fall, Y., et al. Quantitative structure-activity relationship studies of vitamin D receptor affinity for analogues of 1alpha,25-dihydroxyvitamin D3. 1: WHIM descriptors. Bioorg Med Chem Lett, 2005, 15(23):5165~5169.

18 Saíz-Urra, L., González, M., Fall, Y., et al. Quantitative structure-activity relationship studies of HIV-1 integrase inhibition. 1. GETAWAY descriptors. Eur J Med Chem, 2007, 42(1):64~70.

19 Osman, F. G. Pharmacophore Perception, Development, and Use in Drug Design. La Jolla,California: International University Line, 2000:197~198.

20 姜凤超主编. 药物设计学. 北京: 化学工业出版社,2007:210.

Application of the Support Vector Machine in Constructing Interlenlein-1βooverting Enzyme Inhibitors Pharmacophore Model

Yang Ye, Zhang Yanling, Qiao Yanjiang

(School of Chinese Pharmacy, Beijing University of Chinese Medicine, Beijing 100102,China)

Abstract: 1209 molecular descriptors, including topological descriptors, electronic descriptors, and geometric descriptors, were calculated to characterize the physicochemical properties for 24 ICE inhibitors from human peripheral blood monocytes. CfsSubsetEval evaluation method and BestFirst-D1-N5 search method were applied to the variable selection. The Kennard-Stone method was adopted to select the training set and the testing set. Machine learning methods, including the Support Vector Machine (SVM), Decision Tree, Bayesian Network, and Artificial Neural Network, were used to develop the classification models. Meanwhile, three-dimensional pharmacophore models were generated by program Catalyst/HipHop. It was shown that the SVM outperformed other classification models and the prediction accuracy of positive and negative samples reached 100%. The best pharmacophore consisted of five features: two hydrophobic (H), two hydrogen-bond acceptor lipid (Ali), and one hydrogen-bond donor (D). 384 candidate ICE inhibitors were selected from the Traditional Chinese Medicine Database by the pharmacophore. On the other hand, using the SVM classification model to predict the activity of candidate compounds, highly active compounds accounted for 96.6%. The result suggested that the ICE inhibitors pharmacophore model accurately reflected the characteristics of highly active compounds. The pharmacophore model may contribute to screening of new ICE inhibitors from the Traditional Chinese Medicine.

Keywords: ICE inhibitors; Molecular descriptor; Support Vector Machine; Pharmacophore; HipHop

(责任编辑:李沙沙,责任译审:张立崴)