Vol. 28 高等学校化学学报 No. 10
2007 年 10 月 CHEMICAL JOURNAL OF CHINESE UNIVERSITIES 1885 ~ 1888

基于随机森林与 Chemistry Development Kit 描述符的 P-gp 底物识别

马广立[1], 赵筱萍[2], 程翼宇[1]
(1. 浙江大学药物信息学研究所, 杭州 310027; 2. 浙江中医药大学, 杭州 310053)

摘要 应用随机森林方法、开放源代码软件-CDK(Chemistry Development Kit)描述符与 170 个化合物的训练数据集[其中 96 个为磷糖蛋白(P-gp)底物], 建立了 P-gp 底物的识别模型. 研究了 CDK 描述符与 P-gp 底物识别的关系, 结果表明, 原子极化性和电荷偏面积等分子属性对 P-gp 底物识别起到重要作用. 该模型对训练集的预测正确率为 99.42%; 对外部测试集(42 个化合物, 其中 24 个为 P-gp 底物)的预测结果为 P-gp 底物、非底物及总测试集的识别正确率分别为 87.50%, 83.33% 和 85.71%. 212 个化合物数据集上的 Leave-One-Out 交叉验证识别正确率为 77.4%.

关键词 磷糖蛋白; 随机森林; 模式识别
中图分类号 R914.2; O652; O641 **文献标识码** A **文章编号** 0251-0790(2007)10-1885-04

位于细胞膜上的磷糖蛋白(P-glycoprotein, P-gp), 由多药耐药(MDR)基因编码, 依赖于 ATP 供能的泵出转运子. P-gp 在许多正常组织及肿瘤细胞中皆有表达. 其对化疗药物的外排作用即耐药现象是导致癌症治疗失败的最重要因素之一. 同时, P-gp 的外排作用在药物吸收、分布、代谢和排泄(ADME)过程中也起到重要作用. 据此, 在药物发现阶段识别出 P-gp 底物对于药物筛选和开发具有重要意义.

将 P-gp 底物识别与其它的 QSAR/QSPR(定量结构活性/属性关系)或 CSAR/CSPR(分类结构活性/属性关系)建模[1~3]进行比较发现: (1) P-gp 具有多个结合位点; (2) 存在复杂的干扰因素, 如 P-gp 为细胞膜嵌入式蛋白. 早期研究结果表明, 化合物与 P-gp 作用的预测主要是对分子属性及作用过程的探讨[4~8]. 近几年来, 化合物与 P-gp 作用的预测研究主要集中在新出现的分子表征方法与计算方法方面[9~14].

本文利用随机森林方法和 CDK 描述符建立 P-gp 底物识别模型, 相对于以往研究具有更高的预测准确度. 同时, 随机森林的变量筛选能力揭示了分子描述符在 P-gp 底物识别中的重要性. 研究结果表明, 基于 CDK 描述符的随机森林模型对 P-gp 底物识别具有良好的预测能力和稳定性, 从而对研究药物的 ADME 性质预测和新药开发提供理论和应用上的支持. 由于 CDK 为开放源代码软件包, 所以本研究可以由任何单位实现并使用, 具有重要的应用价值.

1 研究方法

1.1 数据来源与分子描述符计算

本文所采用的药物及是否为 P-gp 底物的数据引自文献[13,15]. 合并数据集, 同时剔除相同药物及不能确定结构式的药物数据, 共 212 个化合物. 考虑到 P-gp 底物识别的复杂性, 本研究将较多化合物划入训练集, 即训练集与测试集的比例为 4∶1. 即从总数据集随机抽取 42 个化合物作为外部测试数据集, 其余 170 个化合物作为训练数据集. 训练集与外部测试数据集列于表 1 和表 2.

所用描述符是由 CDK 软件包计算完成. 由来自世界各地的 20 多位开发者共同开发的 CDK 是针对

收稿日期: 2007-01-24.
基金项目: 国家"九七三"计划项目(批准号: 2005CB523402)和浙江省自然科学基金(批准号: Y204418)资助.
联系人简介: 程翼宇, 男, 教授, 博士生导师, 主要从事药物信息学研究. E-mail: chengyy@zju.edu.cn

结构化学与生物信息学建立的免费开放源代码(Open-Source)Java 程序库. 目前版本的 CDK(Version 20050826)在 QSAR 建模方面提供了结构式、拓扑、几何、电荷及杂合等5类描述符[16]. 关于描述符及 CDK 的详细信息请参考 CDK 官方网站(http://cdk.sf.net).

Table 1 Training set*

Compd.	Sub.	Compd.	Sub.	Compd.	Sub.	Compd.	Sub.	Compd.	Sub.	Compd.	Sub.	Compd.	Sub.
CGP-41251	Y	Ritonavir	Y	Digitoxin	Y	NSC268251	N	NSC676617	N	Hydroxyrubicin	Y	Methotrexate	Y
CP100356	Y	S 9788	Y	Digoxigenin	Y	NSC314622	N	NSC676618	N	Mthylprednisolone	Y	Mitoxantrone	Y
Doxorubicin	Y	Saquinavir	Y	Diltiazem	Y	NSC364080	N	NSC678047	N	Pristinamycin IA	Y	Ondansetron	Y
Estradiol	Y	Spiperone	Y	Docetaxel	Y	NSC49899	N	NSC686028	N	Trifluoperazine	Y	Perphenazine	Y
Etoposide	Y	Tacrolimus	Y	Emetine	Y	NSC606532	N	Nigericin	N	Triflupromazine	Y	Fluphenazine	Y
Haloperidol	Y	Teniposide	Y	Endosulfan	Y	NSC615985	N	PSC-833	N	Chlorpromazine	Y	Neostigmine	Y
Idarubicin	Y	Terfenadine	Y	Epothilone	Y	NSC623083	N	Paraquat	N	Corticosterone	Y	Valinomycin	Y
Indinavir	Y	Topotecan	Y	Isosafrole	Y	NSC630148	N	Phosmet	N	Cyclosporin-A	Y	Cinchonidine	Y
LY335979	Y	Toremifene	Y	Pafenolol	Y	NSC630357	N	Prednisolone	N	Dexamethasone	Y	Clotrimazole	Y
Loperamide	Y	Verapamil	Y	Rapamycin	Y	NSC630721	N	Progesterone	N	Dexniguldipine	Y	Daunomycin	Y
Losartan	Y	Vinblastine	Y	Safingol	Y	NSC633528	N	Propranolol	N	Staurosporine	Y	Cefoperazone	Y
Methadone	Y	Vincristine	Y	Vindoline	Y	NSC639677	N	Ranitidine	N	Trimethoprim	Y	Thioridazine	Y
Morphine	Y	Yohimbine	Y	Aldicarb	N	NSC648403	N	Tamoxifen	N	Mitomycin_c	Y	Fexofenadine	Y
Nelfinavir	Y	Acebutolol	Y	Aldoxycarb	N	NSC653278	N	Testosterone	N	Phenobarbital	Y	Domperidone	Y
Nicardipine	Y	Adriamycin	Y	Amantadine	N	NSC664565	N	Triamterene	N	Catharanthine	Y	Pararosaniline	Y
Nifedipine	Y	Amiodarone	Y	Atrazine	N	NSC667532	N	Aminocarb	N	Dipyridamole	Y	Chlorambucil	N
Paclitaxel	Y	Azidopine	Y	CHAPS	N	NSC667533	N	Mannitol	N	Daunorubicin	Y	Methotrexate	N
Phenytoin	Y	Bisantrene	Y	Carbaryl	N	NSC667551	N	Reserpic	N	Mithramycin	Y	Tetraphenyl-	Y
Prazosin	Y	Cefotetan	Y	Carboplatin	N	NSC667558	N	Triforine	N	Chlorpheniramine	N	phosphonium	
Puromycin	Y	Celiprolol	Y	Carmustine	N	NSC668354	N	Vinclozolin	N	Cyclophosphamide	N	S-Farnesyl-	N
Quinidine	Y	Cimetidine	Y	Farnesol	N	NSC671400	N	Lidocaine	N	Podophyllotoxin	N	cysteine	
Quinine	Y	Colchicine	Y	Fluorouracil	N	NSC674508	N	Lindane	N	Fluazifop-butyl	N	Deoxypodo-	N
Rhodamine	Y	Deprenil	Y	Itraconazole	N	NSC676602	N	Melphalan	N	Methoxychlor	N	phyllotoxin	
Rifampicin	Y	Dibucaine	Y	Leptophos	N	NSC676615	N	Mevinphos	N	Pyridostigmine	N		
Phenoxazine	Y	Flupenthixol	Y	Mirex	N	NSC676616	N	Midazolam	N	Propiconazole	N		

* Sub.: the compound is P-gp substrate. Y means yes; N means no.

Table 2 Test set*

Compd.	Sub.	Compd.	Sub.	Compd.	Sub.	Compd.	Sub.	Compd.	Sub.	Compd.	Sub.	Compd.	Sub.
Flunitrazepam	Y	Nimodipine	Y	Calphostin_c	Y	Leupeptin	Y	NSC309132	N	Morphine-6-	Y	Practolol	N
GF120918	Y	Promazine	Y	Cefazolin	Y	Vinorelbine	Y	NSC617286	N	glucuronide		Sumatriptan	N
Gallopamil	Y	Reserpine	Y	Chloroquine	Y	BIBW 22	N	NSC666331	N	Dehydrocorti-	N	Trypan blue	N
Hydrocortisone	Y	Actinomycin_d	Y	Digitoxigenin	Y	Cortodoxone	N	NSC667560	N	costerone		Cytarabine	N
Ivermectin	Y	Aldosterone	Y	Digoxin	Y	Dialifos	N	NSC674570	N	Epipodophyl-	N		
Lovastatin	Y	Amprenavir	Y	Epirubicin	Y	Dieldrin	N	NSC676593	N	lotoxin			
Monensin	Y	Bepridil	Y	Erythromycin	Y	Epinephrine	N	NSC676610	N				

* Sub.: the compound is P-gp substrate. Y means yes; N means no.

1.2 随机森林方法

随机森林方法是分类树方法与统计重采样技术的有机整合. 由于随机森林的诸多特性, 而被广泛应用于生物信息学及药物筛选等领域[17]. 随机森林构建分类模型的计算过程参考文献[18], 其中袋外错误率(Out of Bag Error, OOB Error)为使用一定数量的样本构建随机森林, 预测其余样本(袋外样本)的错误率平均下降准确率(Mean Decrease Accuracy)为替换袋内变量, 并记录袋外的预测值, 然后比较替换与未替换的 OOB 预测变化值; Gini 用于度量每一个节点的分裂质量, 平均下降 Gini(Mean Decrease Gini)为一个变量对随机森林中每棵树每一个节点的 Gini 均值.

本研究的算法流程主要分为3个步骤, 即: 利用随机森林筛选描述符, 优化随机森林树数量, 生成最终的随机森林模型. 利用随机森林自身变量的筛选能力挑选出对 P-gp 底物识别最为重要的描述符. 在确保模型预测能力不降低的情况下, 尽量减少最终预测模型输入变量的数量. 同时, 使用筛选

出来的描述符阐述底物分子属性与 P-gp 之间的关系. 研究中随机森林与变量筛选软件由免费开源统计软件 R 实现(http://www.r-project.org).

2　结果与讨论

2.1　随机森林的结构确定

用 CDK 计算出每个化合物的结构式、拓扑、几何、电荷及杂合等 5 类近百个描述符. 利用随机森林判断变量对分类结果贡献大小的能力进行变量筛选. 根据图 1, 综合考虑变量数量与 OOB 错误率的关系, 挑选出 10 个描述符. 然后, 剔除其中 3 个与其它描述符相关系数高于 0.95 的描述符. 最后, 模型保留的 7 个描述符为: apol(原子极化性之和), CPSA.8, CPSA.11, CPSA.21(电荷偏面积), BCUT.5(奇异值描述符), chi1(碳连接指数)和 weightedPath.0(WP.0, Wiener 路径值). 其中 apol 属于结构式描述符; CPSA.8, CPSA.11, CPSA.21, BCUT.5属于杂合描述符; chi1 和 weightedPath.0 属于拓扑描述符. 7 个描述符之间最高相关系数为 0.76, 表明不存在高相关性.

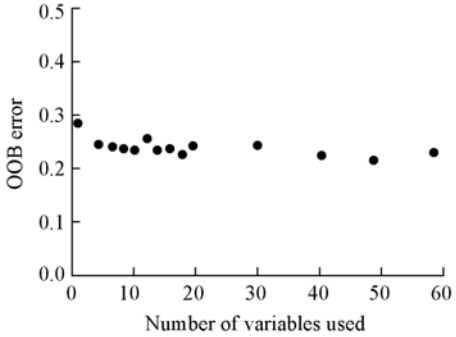

Fig. 1　Relationship between number of variables used and OOB error

根据随机森林的特性, 在实际建模中无需对 CDK 描述符进行筛选. 在随机森林建模之前使用了变量筛选步骤, 主要是为了找出对 P-gp 底物识别贡献最大的描述符集. 同时, 通过比较 CDK 描述符全集与筛选出描述符的子集(P-gp 底物识别能力), 并未发现显著差别, 即使采用筛选出来的 7 个描述符可以替代 CDK 描述符用于 P-gp 底物识别建模.

随机森林的分类结果最终由森林中每棵分类树的结果投票所得, 所以分类树的数量对 P-gp 底物识别具有重要影响. 在分类树数量为 120 时, OOB 错误率及底物与非底物识别错误率为最小, 所以本研究选择 120 作为分类树数量.

2.2　随机森林模型

模型中 7 个描述符对 P-gp 底物识别的重要性由平均下降准确率和平均下降 Gini 决定. 图 2 显示了每个描述符的平均下降准确率和平均下降 Gini 及其排序. 分子的电荷表面积相关属性(CPSA.11, CPSA.23, CPSA.8)在模型中起到了重要作用, 表明分子表面的电荷属性对 P-gp 底物的识别具有重要意义. 分子表面电荷属性对分子进入细胞膜具有重要意义, 这与 Seelig 的假说相符[7]. 同时, CDK 描述符中包含与 LogP 高度相关的 XLogP. 随机森林模型并没有将 XLogP 选入模型, 表示在本研究中 XLogP 对于 P-gp 底物识别不具有重要作用. 这与 Litman 等[5]的观点一致. 与 XLogP 相同, 氢键受体数量对于 P-gp 底物识别也不具有重要影响. 虽然分子极化性对于 P-gp 底物识别具有一定影响, 但其机理还有待于进一步探讨.

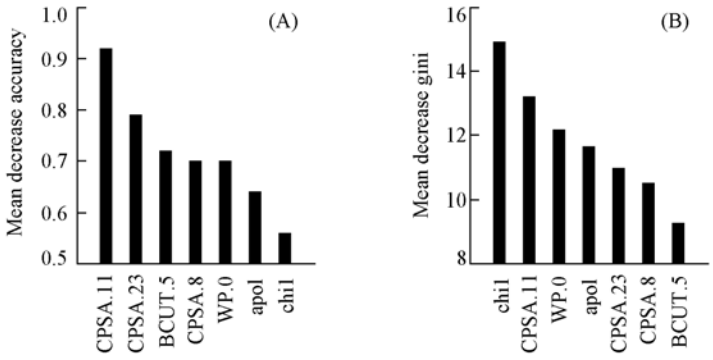

Fig. 2　Importance of each descriptors

最终 P-gp 底物识别随机森林模型由 7 个描述符和 120 棵分类树构成. 该模型对训练集的分类正确率为 99.42%. 表 3 为该模型对外部测试集的分类正确率. 对外部测试集总的分类正确率为 85.71%; 对底物识别的正确率为 87.50%; 对非底

Table 3　Misclassification rates on the external test set

Actual	Predicted		Misclassification rate(%, Total)
	Sub. Y	Sub. N	
Subtrate	21	3	87.50
Nonsubstrate	3	15	83.33(85.71)

物识别的正确率为 83.33%. 该模型对底物的识别准确率要高于对非底物的识别准确率. 基于 7 个描述符和 120 棵分类树的随机森林方法在 212 个化合物的全集之上 Leave-One-Out 交叉验证正确率为 77.4%. 相对于以往研究[10~14], 本模型采用了更少的描述符, 得到了更高的分类正确率. 表明随机森林方法与 CDK 描述符的组合是药物 QSAR 建模中一个有效的研究工具.

参 考 文 献

[1] REN Tian-Rui(任天瑞), SHEN Bin(沈斌), PEI Jian-Feng(裴剑锋), *et al.*. Chem. J. Chinese Universities(高等学校化学学报)[J], 2005, **26**(3): 546—549

[2] MA Yi(马翼), JIANG Lin(姜林), LI Zheng-Ming(李正名), *et al.*. Chem. J. Chinese Universities(高等学校化学学报)[J], 2004, **25**(11): 2031—2033

[3] LI Ji-Lai(李吉来), HANG Ye-Chao(杭烨超), GENG Cai-Yun(耿彩云), *et al.*. Chem. J. Chinese Universities(高等学校化学学报)[J], 2007, **28**(1): 117—120

[4] Bain L. J., Mclachlan J. B., Leblanc G. A.. Environmental Health Perspectives[J], 1997, **105**(8): 812—818

[5] Litman T., Skovsgaard T., Zeuthen T., *et al.*. Biochimica et Biophysica Acta —— Molecular Basis of Disease[J], 1997, **1361**(2): 159—168

[6] Klopman G., Shi L. M., Ramu A.. Molecular Pharmacology[J], 1997, **52**(2): 323—334

[7] Seelig A.. European Journal of Biochemistry[J], 1998, **251**(1/2): 252—261

[8] Seelig A., Landwojtowicz E.. Eur. J. Pharm. Sci.[J], 2000, **12**(1): 31—40

[9] Ekins S., Kim R. B., Leake B. F., *et al.*. Mol. Pharmacol.[J], 2002, **61**(5): 974—981

[10] Pajeva I. K., Wiese M.. J. Med. Chem.[J], 2002, **45**(26): 5671—5686

[11] Penzotti J. E., Lamb M. L., Evensen E., *et al.*. J. Med. Chem.[J], 2002, **45**(9): 1737—1740

[12] Gombar V. K., Polli J. W., Humphreys J. E., *et al.*. J. Pharm. Sci.[J], 2004, **93**(4): 957—968

[13] Xue Y., Yap C. W., Sun L. Z., *et al.*. J. Chem. Inf. Comput. Sci.[J], 2004, **44**(4): 1497—1505

[14] Wang Y. H., Li Y., Yang S. L., *et al.*. J. Chem. Inf. Model[J], 2005, **45**(3): 750—757

[15] Cabrera M. A., Gonzalez I., Fernandez C., *et al.*. J. Pharm. Sci.[J], 2006, **95**(3): 589—606

[16] Steinbeck C., Han Y., Kuhn S., *et al.*. J. Chem. Inf. Comput. Sci.[J], 2003, **43**(2): 493—500

[17] Svetnik V., Liaw A., Tong C., *et al.*. J. Chem. Inf. Comput. Sci.[J], 2003, **43**(6): 1947—1958

[18] Breiman L.. Machine Learning[J], 2001, **45**(1): 5—32

Identification of P-gp Substrates Using a Random Forest Method Based on Chemistry Development Kit Descriptors

MA Guang-Li[1], ZHAO Xiao-Ping[2], CHENG Yi-Yu[1*]

(1. *Pharmaceutical Informatics Institute*, *Zhejiang University*, *Hangzhou* 310027, *China*;
2. *Zhejiang Chinese Medical University*, *Hangzhou* 310053, *China*)

Abstract A model to identify P-glycoprotein(P-gp) substrate was constructed with a random forest method based on open source software CDK(Chemistry Development Kit) descriptors and a training data set which contained 170 compounds(96 P-gp substrates). The study on the relationship between CDK descriptors and P-gp substrates indicates that sum of the atomic polarizabilities and charged partial surface area play important roles in identifying P-gp substrates. An external test data set containing 42 compounds(24 P-gp substrates) was employed. The correct classification rate on the training set is 99.42% and the correct classification rates for P-gp substrates, non-substrates and the total compounds on the test set are 87.50%, 83.33% and 85.71%, respectively. Leave-One-Out cross-validation correct classification rate (212 compounds) was 77.4%.

Keywords P-glycoprotein(P-gp); Random forest; Pattern recognition

(Ed.: H, J, Z)