高等学校化学学报 2007 , 28 (10): 1885—1888 ISSN: 0251-0790 CN: 22-1131/O6

研究论文

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

收稿日期 2007-1-24 修回日期 网络版发布日期 2007-10-24 接受日期

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

分类号 R914.2 0652 0641

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%.

Key words P-glycoprotein(P-gp) Random forest Pattern recognition

DOI:

通讯作者 程翼宇 chengyy@zju.edu.cn