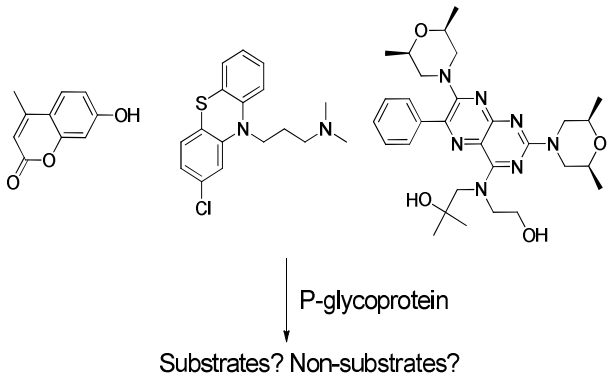
P-glycoprotein Substrate Models Using Support Vector Machines Based on a Comprehensive Dataset

Zhi Wang[1,2], Andreas Bender[2], Robert C. Glen[2], Aixia Yan[1]

[1]State Key Laboratory of Chemical Resource Engineering, Department of Pharmaceutical Engineering, P.O. Box 53, Beijing University of Chemical Technology, 15 BeiSanHuan East Road, Beijing 100029, P. R. China.

[2]Unilever Centre for Molecular Science Informatics, Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge, CB2 1EW, United Kingdom



P-glycoprotein (P-gp) is one of the major ABC transporters and involved in many essential processes such as lipid and steroid transport across cell membranes, but also in the uptake of drugs such as HIV protease and reverse transcriptase inhibitors. Despite its importance, reliable models predicting substrates of P-gp are scarce. In this study, we have built several computational models to predict whether or not a compound is a P-gp substrate, based on the largest dataset yet published, employing 332 distinct structures. Each molecule is represented by ADRIANA.Code, MOE and ECFP_4 fingerprint descriptors. The models are computed using a support vector machine based on a training set which includes 131 substrates and 81 non-substrates that were evaluated by 5-, 10-fold and leave-one-out (LOO) cross-validation. The best model gives a Matthews Correlation Coefficient of 0.73 and a prediction accuracy of 0.88 on the test set. Examination of the model based on ECFP_4 fingerprints revealed several substructures which could have significance in separating substrates and non-substrates of P-gp, such as the nitrile and sulfoxide functional groups which have a higher frequency in non-substrates than in substrates. In addition structural isomerism in sugars was found to result in remarkable differences regarding the likelihood of a compound to be a substrate for P-gp.