ISSN: 2249 –4820

Chemistry & Biology Interface, 2012, 2, 3, 140-156

Chemistry & Biology Interface

An official Journal of ISCB, Journal homepage; www.cbijournal.com

Research Paper

Information rich Descriptors for Predicting P-Glycoprotein Substrates and Non-substrates: A Binary-QSAR approach

Ranajit Shinde[1], Shikhar Gupta[1] and C. Gopi Mohan[12]*

[1]*Department of Pharmacoinformatics, National Institute of Pharmaceutical Education and Research (NIPER), Sector 67, S.A.S. Nagar- 160 062, Punjab, INDIA.* [2]*Amrita Centre for Nanosciences and Molecular Medicine (ACNSMM), Amrita Institute of Medical Sciences, Ponekkara, Kochi- 682 041, Kerala State, India*

Received 14 March 2012; Accepted 23 April 2012

Keywords: P-glycoprotein; Permeability; Binary QSAR; *in silico* modeling, substrate

Abstract: A highly probabilistic binary quantitative structure-activity relationship (binary-QSAR) model has been developed to predict substrates and non-substrates of P-glycoprotein (Pgp). A total of 123 compounds, classified as Pgp substrates/non-substrates, on the basis of the efflux ratio from Pgp monolayer efflux assays were selected in this study. Binary-QSAR model is developed on training set of 99 diverse compounds (36 substrates and 63 non-substrates) using 12 information rich descriptors. Solubility, Lipinski violation score, partition coefficient, CYP2D6 enzyme substrate probability etc. were some of the important descriptors used in developing binary-QSAR model. This model showed excellent overall prediction accuracy of 100% on substrates and non-substrates for training set of 99 compounds. Further, the leave-one-out cross-validated prediction accuracy was 96.9% on substrates and non-substrates. When applied to the test set of 24 compounds (8 substrates and 16 non-substrates), model correctly predicted the behavior 6 out of 8 substrates (75%) and 15 out of 16 non-substrates (94.4%). These three mispredictions were found to lie in the limitation zones of Pgp monolayer efflux assay, where it is difficult to classify compounds as Pgp substrate or non-substrate. Present model can be seen as *in silico* simulation for predicting the result of *in vitro* Pgp monolayer efflux assay. The results suggest that it is a powerful tool to identify substrate or non-substrate nature of compounds, and can be used in high-throughput screening.

1. Introduction

P-glycoprotein (Pgp) is a well characterized transporter family of adenosine triposphate binding cassette. It is extensively distributed and expressed in normal cells of all species, such as, columnar epithelial cells of lower gastrointestinal tract, capillary endothelial cells of brain and testis, canalicular surface of hepatocyte and apical surface of proximal tubule in kidney [1]. Pgp limits oral absorption, restrict blood-brain barrier penetration and modulate hepatic, renal, or intestinal elimination [2-4]. Drugs from different therapeutics classes such as

Corresponding Author* : Phone: +91-172-2214682, Fax: +91-172-2214692, E-mail: cgopimohan@yahoo.com

ISSN: 2249 –4820

Chemistry & Biology Interface, 2012, 2, 3, 140-156

calcium channel blockers, neuroleptics, antiarrythmics, antimalarial, antifungal and anticancer agents are known as substrates of Pgp [5]. Over expression of Pgp have decreased efficacy of several drugs, and is a major cause of multi-drug resistance in different diseases [6]. Identification and classification of Pgp substrates from its non-substrates and inhibitors is essential in drug candidate selection and optimization.

Different experimental and theoretical methods have been developed to unravel the Pgp modulating activity of the compounds, by modeling of Pgp substrates, non-substrates and inhibitors. Experimental methods mainly include, (i) Pgp monolayer efflux assay, (ii) Pgp ATPase assay and (iii) Calcein-AM inhibition assay. Polli et al. with the use of Pgp monolayer efflux assay, analyzed 66 compounds, and showed that this assay method is more reliable to classify compounds as Pgp substrates at low/moderate permeability [5]. Theoretical methods include different physico-chemical properties such as logP, molecular weight, surface area, aromaticity, amphiphilicity, proton basicity, hydrogen bonding capacity; and molecular structural features such as two electron donor groups with a spatial separation of (2.5±0.3)Å/ or (4.6±0.6)Å or three electron donor groups with a spatial separation of outer two groups of (4.6±0.6)Å. These properties have been found to contribute significantly towards interactions of substrates with Pgp [7-12].

Different research groups have successfully developed QSAR models using combination of 2D and 3D descriptors. Both qualitative and quantitative molecular studies offer insights in this direction through different *in silico* models and thereby distinguish Pgp substrates from its non-substrates. Such models are very useful for the selection of lead candidate and thereby to reduce the attrition rate in the later stages of development [10-20].

The present paper describes development of binary-QSAR model using QuaSAR-binary module of Molecular Operating Environment (MOE) software (Ver-2003.02) [22] on a dataset of 123 compounds, assayed by monolayer efflux assay (MDR1-MDCK cell line) for Pgp activity [5,21], It predicts compound as Pgp substrate or non-substrate.

2. Methodology

2.1. Collection and Classification of Substrates and Non-substrates of Pgp

Compounds (substrates and non-substrates) assayed under uniform condition by monolayer efflux assay using MDR1-MDCK cell lines were collected from the literature [5,21]. In binary methodology, before building the QSAR model, one need to assign active (substrates) or inactive (non-substrate) category to the compounds. In this study we used Pgp efflux ratio [$P_{app\ B\rightarrow A}$ /$P_{app\ A\rightarrow B}$], obtained from the monolayer efflux assay, to assign substrate or non-substrate category to the compounds. Classification of substrates and non-substrates was performed in accordance with the following criteria:

Activity =1 [if a) efflux ratio > 2.0, or

(Pgp substrate) b) efflux ratio is between 1.5 to 2.0 and dropped to ~1.00 in presence of specific Pgp inhibitor GF120918]

and

Activity =0 [if a) efflux ratio < 1.5, or

(Pgp non-substrate) b) efflux ratio is between 1.5 to 2.0 and not dropped to ~1.00

ISSN: 2249 –4820

Chemistry & Biology Interface, 2012, 2, 3, 140-156

in presence of specific Pgp inhibitor GF120918]

Classification of compound as a substrate or a non-substrate was made on the basis of above said efflux ratio [5,21] and binary value '1' assigned to Pgp substrates (active) and '0' for Pgp non-substrates (inactive). Thus, total of 123 compounds assayed under uniform conditions and having high quality *in vitro* results, classified as substrates and non-substrates of Pgp, in which 99 (36 - substrates, 63 - non-substrates) were used as training set to develop binary-QSAR model, and 24 (8 - substrates, 16 - non-substrates) were used as test set to evaluate performance of the model.

2.2. Representation of Compounds

Compounds were drawn using ChemDraw Ultra version 6.0.1 and converted into 3D structures using Chem3D Pro version 6.0 [22]. 3D structures were then manually inspected to represent proper chirality, and then the geometry was optimized using molecular mechanics force field (MM2) method with rms gradient of 0.100. Molecular descriptors were calculated using MOE (Ver. 2003.02) and Cerius^2 (Ver. 4.10) software [23,24]. Molecular structure of the training and test set compounds is given in supplementary material Figure (A&B).

2.3. Molecular descriptors Selection, Definition with its Significance

Twelve molecular descriptors computed from two different software, MOE and Cerius^2, was used to build QSAR model. Descriptors were selected based on the knowledge about substrates and non-substrates and their interaction with Pgp. These molecular descriptors are shown in Table 1. The description and significance of these molecular descriptors can be elaborated by understanding its physiological behavior.

Descriptor "a_hyd" belongs to structural type and was used to quantify number of hydrophobic atoms present in a compound (Table 1). It can also be defined as pharmacophore atom type descriptor by assigning a type to each heavy atom in a compound using a rule based system. Ability of compound to cross the biological membrane was quantified in the form of partition coefficient descriptor (SlogP). "SlogP" i.e. log of the octanol/water partition coefficient (including implicit hydrogens) belongs to thermodynamic descriptor and is calculated using atomic contribution model [25]. Another descriptor, "Lip_violation", was used to represent absorptive power or permeation of a compound. It depends on molecular weight, AlogP98, Lipinski H-bond acceptor and Lipinski H-bond donor of the compound [26]. The electronic descriptor "PEOE_VSA_FPPOS" (or Q_VSA_FPPOS) defines total positive polar van der Waals surface area, which can influence hydrogen bonding character. This charge dependent descriptors prefixed with PEOE_ or Q_ use the partial charges stored with each structure in the database which is the sum of the v_i (van der Waals surface area of atom i calculated using a connection table approximation) such that q_i (partial charge of atom i) is non-negative. Descriptors chi0, PHI and IAC-Total indicates the degree of flexibility, size, shape and connectivity of atoms and the un-saturation in the compound. "Chi" descriptor is refinement of shape index that takes into consideration the contribution of covalent radii and hybridization states, making the shape of the compound [27]. "PHI" descriptor is based on structural properties which restrict a compound being "infinitely flexible", the model for which is an endless chain of

ISSN: 2249 –4820

Chemistry & Biology Interface, 2012, 2, 3, 140-156

$C(sp^3)$ atoms. The structural features for infinite flexibility restriction include (i) fewer atom, (ii) presence of rings, (iii) branching and (iv) the presence of atoms with covalent radii smaller than those of $C(sp^3)$ [28]. To represent mixture of atomic contribution to molecular refractivity and their connectivity, the thermodynamic descriptor "GCUT_SMR_0" was used. Solubility, one of the main factors controlling permeation was quantified using "ADME_solubility" descriptor (Table 1) [29]. Probability of compound to become substrate of enzyme CYP2D6, a member of the cytochrome P450 mixed-function oxidase system, was quantified using "ADMET_CYP2D6_PROB" descriptor. The cytochrome P450 2D6 model predicts CYP2D6 enzyme inhibition using 2D chemical structure as input. The model was developed from known CYP2D6 inhibition data on a structurally diverse set of 100 compounds where an ensemble of recursive partitioning trees was trained against 2D descriptors and 1D similarity data [30]. Number of atoms of type H_48 and N_66, used in calculation of AlogP98, were quantified using descriptor Atype_H_48 and Atype_N_66 respectively, and are presented in Table 1 [31].

2.4. Binary Methodology

Binary QSAR method builds the predictive binary models through the use of highly significant probabilistic and statistical inference. The predictive capacity of binary QSAR is not interpolative, because data fitting is not used, and is based on generalizations substantiated by the experimental data. In this work we have used binary methodology, based upon statistical probability estimation, first introduced by Labute [32], for the development of binary QSAR model. It correlates structural properties of compounds with a "binary" expression of biological activity and calculates probability value in the form of: 1 = active => substrate and 0 = inactive=> non-substrate. The derived binary-QSAR model predict the probability of new compound(s) to be substrate or non-substrate of Pgp. Binary-QSAR methodology have been successfully applied previously in several investigations such as estrogen receptor ligands, carbonic anhydrase II inhibitors and MAO inhibitors [33-35].

Binary method uses results $\{(y_i, x_i)\}$ of the experiment for a set of m compounds. y_i is either 0 or 1 (either "inactive" or "active") and x_i are the vectors that correspond to a set of n molecular descriptors $(x_{i1}, x_{i2,\ldots} x_{in})$, i.e., $x_{i=(} x_{i1}, x_{i2,\ldots} x_{in})$

The binary-QSAR analysis procedure is summarized in Figure 1. Briefly, a set of molecular descriptors are computed for each compound in a data set which is then transformed into a set of de-correlated and normalized set of variables, and the probability distribution is estimated based on Bayes' theorem. Quality of a binary-QSAR model is measured as follows: let m1 represent the number of substrates, m0 the number of non-substrates, c1 the number of substrates correctly labeled by the QSAR model and c0 the number of non-substrates correctly predicted by the QSAR model. Then three parameters of performance are calculated as: (i) accuracy on substrate, [A1=$c1/m1$]; (ii) accuracy on non-substrate, [A0=$c0/m0$]; (iii) overall accuracy on all of the compounds, [A=$(c0 + c1)/(m0 + m1)$]. The details of binary-QSAR methodology have been illustrated by Labute [32]. The computational procedure used for developing binary-QSAR model is depicted below as flow chart (Figure 1).

3. Results and Discussion

ISSN: 2249 –4820

Chemistry & Biology Interface, 2012, 2, 3, 140-156

3.1. Development of binary-QSAR model

Binary-QSAR model was developed using 123 compounds which had been assayed under uniform conditions in which 44 compounds are substrates and 79 compounds are non-substrates of Pgp. Compounds were grouped into training set (99 Compounds) and test set (24 Compounds) based on their permeability value ($P_{app\ A\rightarrow B}$) [5,21]. Permeability was shown as a very significant factor affecting Pgp efflux [36]. While grouping the compounds it was assured that, both training and test set will contain nearly same range of permeability values so that robust model could be built for different sets of permeability values.

The computational procedure for developing binary-QSAR model, using MOE software, is depicted in Figure 1. Compounds were first assigned binary value (1 or 0) based on its class (substrate or non-substrate) as explained above. Then twelve important descriptors as mentioned above, representing different physicochemical properties of compounds in training set, were used to build binary-QSAR model. Threshold value was set default and smoothing parameter (σ^2) was optimized by testing the cross-validated accuracy of the model over several trial values. When σ^2=0.25 value was used, the cross-validated statistics gave accuracy of 50% on substrates, 92% on non-substrates, with total accuracy of 77% on the model. There is considerable improvement in the cross-validated statistics of the model by decreasing the σ^2 value further and 0.01 gave the best statistics with accuracy of 92% on substrates, 100% on non-substrates with total accuracy of 97% on the model.

3.2. Quality measurement of individual Molecular descriptor

Quality of the 12 individual molecular descriptors in the training set of 99 compounds was measured using MOE software. Percentage accuracy of each molecular descriptor to predict substrates and non-substrates of Pgp for this training set is presented in Table 2. Three parameters of performance (A, A0 and A1) were calculated for each descriptor (details given in “binary method” section). Total accuracy (A) - fraction of observations correctly predicted by the descriptor, accuracy on non-substrates (A0) - fractions of correctly predicted sample non-substrates, accuracy on substrates (A1) - fractions of correctly predicted sample substrates. The results indicate that the percentage predictive accuracy on non-substrates (A0) is better than that of substrates (A1) (Table 2). The total accuracy (A) of all the 12 information rich individual molecular descriptor showed more than 60% contribution and gave good statistical significance for building binary-QSAR model (Table 2). Accuracy on A0 might be affected because of the dominant number of 63 non-substrates as compared to 36 substrates in the training set of 99 compounds. Contribution of all 12 individual molecular descriptors has minimized the biased effect, and showed good statistical significance on total accuracy (A).

Pgp activity relates mainly to structural, electronic, thermodynamic and ADME properties of the compound and the descriptor space we have chosen cover all this aspects significantly. Biological significance of these descriptors is well correlated to the binding, transport and solubility of compounds. Molecular hydrophobicity is considered as brain uptake of drugs and the descriptors mainly contributing to this effect include a_hyd, Lip_violation and ADME_Solubility. This is because hydrophobicity is a major

ISSN: 2249 –4820

Chemistry & Biology Interface, 2012, 2, 3, 140-156

determining factor in a compound's absorption, distribution in the body, penetration across vital membranes and biological barriers, metabolism and excretion. Besides predicting the likely transport of a compound around the body, it also impacts formulation, dosing, drug clearance, and toxicity. Hence it plays a critical role in helping late stage attrition in the drug discovery process.

3.3. Robustness of binary-QSAR model

Binary value and predicted probability values obtained using binary-QSAR methodology for training set of 99 compounds is presented in Table 4. Quality of the binary-QSAR model was measured using three parameters of performance (A, A0 and A1) as given in "binary method" section. Model robustness also requires high compound to-variables ratio to be significant and is greater than 8.2 (99/12). The stability of the model is validated by its prediction accuracy on the training set using cross validation statistics. The sensitivity (i.e. ability to correctly identify substrates) and specificity (i.e. ability to correctly identify non-substrates) of developed binary-QSAR model is explained below. Cross-validated statistical accuracy of the model is presented in Table 3. This model has 11 principle components as shown in supplementary material-Table A. Correlation coefficient (in Table A)- indicates whether or not the substrate and non-substrate distributions are correlated (0 = perfectly correlated, 1 = perfectly uncorrelated).

Developed binary-QSAR model is highly predictive and shows excellent cross-validated statistics as shown in Table 3. Leave-one out (LOO) cross validation procedure was done by leaving one compound out, for building the model, and then testing the left out compound. Total percentage predictive accuracy on all the compounds in the training set was 96.9% (accuracy on its substrates and non-substrates). Cross-validated predictive accuracy in training set was 100% on non-substrates and 91.6% on substrates as shown in Table 3. The χ^2 test of significance (*p*-value) was also used to judge the quality of the model. The binary-QSAR model described above shown a highly significant *p*-value of 2.30e-017 for all the compounds and 1.16e-018 for the substrates/non-substrates (Table 3). These low *p*-values indicate that the molecular descriptors and binary-QSAR model contributes in a significant way for the prediction of Pgp substrates and non-substrates. The p-value quoted under accuracy on actives (substrates)/inactives (non-substrates) is the probability that both these accuracies would differ from the "chance accuracy" as much as they do, if the substrates and non-substrates in the QSAR model are uncorrelated with those in the sample. In other words, our binary-QSAR model is highly sensitive in identifying Pgp substrates and highly specific in identifying Pgp non-substrates. Present dataset due to its structural diversity seems to be quite valid for LOO cross-validation analysis.

Among 12 descriptors, Lipinski violation (Lip_violation) count descriptor predicted substrates with 30.6% accuracy while non-substrates with accuracy of 92.1%, (Table 2). The above Lipinski analysis clearly shows that absorption is an important determinant of Pgp efflux. This result is in concordance with experimental findings of efflux studies on drugs, which showed that unfavorable chemical features of P-gp substrates limit passive permeability and thus are more susceptible to P-gp-mediated efflux [5, 36]. Thus the compound with poor passive absorption has greater prevalence to become Pgp substrate and vice versa.

ISSN: 2249 –4820

Chemistry & Biology Interface, 2012, 2, 3, 140-156

Another descriptor, number of hydrophobic atoms (a_hyd) have highest accuracy of 61.1%, followed by descriptor IAC-Total of 55.5% and descriptor PEOE_VSA_FPPOS of 52.8%. Significant contribution of descriptor (a_hyd) suggests hydrophobic interactions with Pgp, an important factor for becoming its substrate. The present analysis is in agreement with other *in silico* studies where it was shown that the multiple hydrophobic points are required for binding to Pgp [8-12].

Screening studies for Pgp drug interactions identified a number of clinically important drugs as Pgp substrates which are very diverse e.g. anthracyclines (doxorubicin, daunorubicin), alkaloids (reserpine, vincristine, vinblastine), specific peptides (valinomycin, cyclosporine), steroid hormones (aldosterone, hydrocortisone) and local anaesthetics (dibucaine) [37]. Even dye molecules (Rhodamine 123) and pharmaceutical excipients exhibited Pgp substrate activity. Examples of some non-substrates are alprenolol, amantadine, ametryptyline, atenolol, biperidine, bromocryptine, etc. [38] These compounds show distinct differences in permeability, molecular weight and polar surface area. It was observed that almost all non-substrates are within the limits of the Lipinski's rule of five. Most of the Pgp substrates belong to the upper limits of molecular weight (>500) and total polar surface area (>75). Our model also supports these observations viz Lipinski violation count (lip_violation) and number of hydrophobic atoms (a_hyd) is shown to be important descriptors to identify compounds as Pgp substrates and non-substrates. It was also shown that unfavorable chemical features of Pgp substrates limits the passive permeability and hence become more susceptible to Pgp efflux. Thus poor passive absorption and higher hydrophobic interactions are important factor in making compounds substrate to Pgp.

3.4. Validation of binary QSAR Model

Robustness and the true test of any QSAR model's performance is its accuracy on a set of compounds not included in the training set. Therefore, the developed binary-QSAR model was applied to an additional test set of 24 compounds, not included in the training set. Table 5 shows the test set compounds, its experimental efflux ratio (B→A/A→B), Pgp experimental status for substrate/or non-substrate, observed binary activity of compound for substrate-1/non-substrate-0); predicted binary activity of compounds to be substrate/or non-substrate and predicted Pgp binary status of compounds. QSAR model correctly predicted 21/24 compounds (87.5%) as either substrates or non-substrates. Among eight substrates, six compounds correctly predicted by the *in silico* screen with overall prediction sensitivity of 75%. Better prediction specificity of 94.4% was obtained with 17 of the 18 non-substrates as shown in Table 5. A lower sensitivity, i.e., a higher rate of false negative predictions was expected to some extent due to non-substrate biased training data set.

The three mispredicted (or outlier) compounds in the developed binary QSAR model are nortriptyline, erythromycin and diltiazem compounds. Nortriptyline compound was non-substrate, which was predicted as substrate while erythromycin and diltiazem substrates was predicted as non-substrates and is shown in Table 6. For nortriptyline, mean permeability P_{app} (A→B) is high, 337 nm/s (efflux ratio 1.39). *In vitro* monolayer efflux assay cannot accurately determine the substrate nature of highly permeable compounds. At the same time, calcein-AM efflux assay has

ISSN: 2249 –4820

Chemistry & Biology Interface, 2012, 2, 3, 140-156

confirmed the binding of nortriptyline to the Pgp. Thus, QSAR model is likely correct in identifying this compound as substrate. The outlier erythromycin has a very limited permeability of 0.94 nm/s (efflux ratio 14.4) at which it is very difficult to accurately determine the property to become substrate of Pgp in the monolayer efflux assay. Diltiazem have efflux ratio of 1.53 or 1.64 which suggest that it falls in non-confident zone (B→A/A→B ratio 1.5 to 2.0), where classification of compound as substrates/non-substrates is uncertain. In addition, diltiazem is absorbed almost completely through the intestine. This suggests that diltiazem is borderline substrate and has weak substrate activity towards Pgp [5,21]. The present results clearly suggest that false positive prediction on nortriptyline compound and false negative prediction on erythromycin and diltiazem compound (Table 6), derived from binary-QSAR model are due to experimental limitations of monolayer Pgp efflux assay.

3.5 Practical implications of the Model

Review of literature survey has shown different *in silico* qualitative and quantitative molecular models to offer insights into the molecular determinants of Pgp substrates, non-substrates and/ or inhibitors. Attempts to develop QSARs for Pgp substrates/non-substrates/inhibitors/modulators, to link their physico-chemical properties with the biological activity, have been dealt with the difficulties. Different groups have successfully linked combination of 2D and 3D physicochemical properties of compound, to represent their substrate and non-substrate nature towards Pgp, using different statistical approaches [10, 13-18,39]. It has been observed that more complex descriptors and powerful statistical methods of molecular modeling are necessary for identification of Pgp substrates and non-substrates [17]**.** Therefore we decided to develop an *in silico* QSAR model, to surrogate experimental monolayer Pgp efflux assay, using different topological and physicochemical descriptors.

Further, carrying out *in vitro* monolayer Pgp efflux assay is a cumbersome procedure and requires lot of laboratory work (culturing MDCK type II cells, maintaining Pgp expression, simultaneous determination of permeability, end point determination using LC/MS/MS). Also this assay experiments enable determination of substrate activity of only dozen compounds per week. But *in silico* binary-QSAR model can rapidly and reliably identify the compounds to be substrate from their structure. This will circumvent the requirement of time, manpower, hard work and money in performing *in vitro* monolayer Pgp efflux assay. The present model can also prioritize small subset of potential hit compounds in chemical libraries or potential lead compounds to be obtained from high throughput screening, which will further guide in performing *in vitro* monolayer Pgp efflux assay experiments.

4. Conclusions

We have developed binary-QSAR model using dataset of 123 compounds and 12 information rich descriptors, for predicting substrates and non-substrates of Pgp. Lipinski violation count (lip_violation) and number of hydrophobic atoms (a_hyd) are important descriptors for identifying compounds as substrates and non-substrates of Pgp. The present result indicated poor passive absorption and higher hydrophobic interactions is significant feature for making compounds substrate of Pgp. However, the three wrong predictions in test set were found to be because of limitation of the *in*

ISSN: 2249 –4820

Chemistry & Biology Interface, 2012, 2, 3, 140-156

vitro monolayer Pgp efflux assay. This model can be seen as *in silico* simulation of *in vitro* monolayer Pgp efflux assay, and can be used in the early stage of drug discovery to prevent the compounds from its late stage attrition.

Acknowledgements

One of the authors C. Gopi Mohan acknowledges DST, New Delhi for financial support for this work. Shikhar Gupta acknowledges DBT, New Delhi for the award of Junior research fellowship.

Table 1. Molecular descriptors and its corresponding software

Descriptor-tag	Descriptor type | Definition	Software used
a_hyd	Structural descriptor | Number of hydrophobic atoms	MOE
SlogP	Thermodynamic descriptor | Log of the octanol/ water partition coefficient	
Lip_violation	ADME descriptor | Absorptive power of compound	
PEOE_VSA_FPPOS	Electronic descriptors | Fractional positive polar van der Waals surface area.	
Chi0	Topological descriptor | Atomic connectivity index (order 0)	
GCUT_SMR_0	Thermodynamic descriptor | Mixture of atomic contribution to molecular refractivity and their connectivity	
ADME_Solubility	ADME descriptor | Solubility factor controlled by absorption	Cerius[2]
ADMET_CYP2D6_PROB	ADME descriptor | Probability of compound to become substrate of enzyme CYP2D6	
Atype_N_66	Thermodynamic descriptor | Number of atoms of type N_66 used in calculation of AlogP98	
Atype_H_48	Thermodynamic descriptor | Number of atoms of type H_48, used in calculation of AlogP98	
PHI	Topological descriptor | molecular flexibility index	
IAC-Total	Information content descriptor | The atoms in the compound are partitioned into equivalence classes corresponding to their atomic numbers	

Table 2. Percentage predicted accuracy of individual molecular descriptor.

S. No.	Molecular Descriptor	A (%)	A0 (%)	A1 (%)
1	GCUT_SMR_0	65.70	90.50	22.20
2	SlogP	67.70	85.70	36.10
3	Lip_violation	69.70	92.10	30.60
4	PEOE_VSA_FPPOS	70.70	81.00	52.80

ISSN: 2249 –4820

Chemistry & Biology Interface, 2012, 2, 3, 140-156

5	chi0	71.70	85.70	47.20
6	a_hyd	76.80	85.70	61.10
7	ADME_Solubility	68.70	87.30	36.10
8	ADMET_CYP2D6_PROB	66.70	90.50	25.60
9	Atype_N_66	64.60	90.50	19.40
10	Atype_H_48	67.70	95.20	19.40
11	PHI	70.70	85.70	44.40
12	IAC-Total	74.70	85.70	55.60

Table 3. Cross-validated statistics of Pgp substrates and non-substrates in training set of 99 compounds.

	Cross-validated statistics	Chance accuracy
Total percentage accuracy (A)	96.9%	55.0%
χ^2 test of significance (*p*-value)	2.30e-017	
Accuracy on substrates (A1)	91.6%	33.0%
Accuracy on non-substrates (A0)	100.0%	67.0%
χ^2 test of significance (*p*-value)	1.16e-018	

Table 4. Training set compounds; Pgp efflux ratio (B→A/A→B); observed and predicted binary activity of compounds (Substrate-1/ Non-substrate-0).

Compound Name (Therapeutic Indication)	(B→A/A→B) Ratio[a,b]	Pgp status[c] (exptl.)	Binary activity (observed)	Binary activity (predicted)	Pgp status[c] (predicted)
Acrivastine (antihistamine)	3.71	S	1	0.99	S
Amantadine (antiviral, antiparkinsonian)	0.84[a],0.95[b]	N	0	0	N
Amitriptylline (antidepressant)	1.34	N	0	0	N
Amprenavir (antiviral)	32.4[a],29.0[b]	S	1	0.94	S
Astemizole (antihistamine)	2.22	S	1	0.99	S
Atenolol (antihypertensive)	1.24	N	0	0	N
Biperiden (antiparkinsonian)	0.95	N	0	0	N
Bromocriptine (antiparkinsonian)	1.26	N	0	0	N
Bufuralol (antihypertensive)	0.78	N	0	0	N
Buspirone (anxiolytic)	0.95	N	0	0	N
Carbamazepine (anticonvulsant)	0.98	N	0	0	N
Cetirizine (antihistamine)	8.65	S	1	0.99	S
Chloroquine (antimalerial)	3.83	S	1	1	S
Chlorpheneramine (antihistamine)	1.14[a],0.93[b]	N	0	0	N
Chlorpromazine (antipsychotic)	1.09[a],1.27[b]	N	0	0	N

ISSN: 2249 –4820

Chemistry & Biology Interface, 2012, 2, 3, 140-156

Chlorprothixene (antipsychotic)	1.26	N	0	0	N
Cimetidine (anti-ulcerative)	2.19[a],4.77 [b]	S	1	1	S
Clarythromycin (antibiotic)	31.3	S	1	1	S
Clemastine (antihistamine)	1.29	N	0	0	N
Clomipramine (antidepressant)	1.42	N	0	0	N
Colchicine (antiarthritis)	11.34	S	1	0.94	S
Cyclobenzaprine (muscle relaxer)	0.97	N	0	0	N
Daunorubicin (antineoplastic)	14.2	S	1	1	S
Desipramine (antidepressant)	1.03	N	0	0	N
Dexamethasone (corticosteroid)	12.4	S	1	1	S
Diphenhydramine (antihistamine)	0.91	N	0	0	N
Dipyridamole (vasodilating agent)	22.7	S	1	1	S
Domperidone (antiemetic)	31.2	S	1	1	S
Doxapram (respiratory stimulant)	1.41	N	0	0	N
Doxepin (antidepressant)	1.15	N	0	0	N
Doxorubicin (antineoplastic)	0.67	N	0	0	N
Doxylamine (antihistamine)	0.84	N	0	0	N
Eletriptan (antimigraine)	31.5[a],44.7[b]	S	1	0.94	S
Emetine (antiprotozoal)	29.2	S	1	1	S
Etoposide (antineoplastic)	2.8	S	1	1	S
Famciclovir (antiviral)	3.17	S	1	1	S
Flumazenil (benzodiazepine antagonist)	0.92	N	0	0	N
Fluoxetine (antidepressant)	1.18	N	0	0	N
Flurazepam (anxiolytic)	0.88	N	0	0	N
Fluvoxamine (antidepressant)	1.2	N	0	0	N
Guanabenz (antihypertensive)	0.91	N	0	0	N
Haloperidol (antipsychotic)	1.04	N	0	0	N
Hoechst 33342 (mutagent)	7.75	S	1	1	S
Imipramine (antidepressant)	1.05	N	0	0	N
Indomethacin (antiinflammatory)	0.97	N	0	0	N
Itraconazole (antifungal)	0.97	N	0	0	N
Ketoconazole (antifungal)	1.02	N	0	0	N
Labetolol (antihypertensive)	8.85	S	1	1	S

ISSN: 2249 –4820

Chemistry & Biology Interface, 2012, 2, 3, 140-156

Lidocaine (anesthetic)	0.99[a],0.83[b]	N	0	0	N
Loratidine (antihistamine)	1.9	S	1	0.99	S
Lorcainide (antiarrythmic)	1.45	N	0	0.01	N
Mannitol (diuretic)	0.82[a],0.88[b]	N	0	0	N
Maprotiline (antidepressant)	1.05	N	0	0	N
Mebendazole (anthelmintics)	0.91	N	0	0	N
Meprobamate (anxiolytic)	0.97	N	0	0	N
Mequitazine (antihistamine)	2.81	S	1	0.96	S
Metergoline(analgesic, antipyretic)	1.21	N	0	0	N
Methotrexate (antineoplastic)	0.67	N	0	0	N
Methysergide (antimigraine)	4.33	S	1	0.99	S
Metoprolol (antihypertensive)	1.21	N	0	0	N
Midazolam (anesthetic)	0.81[a],1.0[b]	N	0	0	N
Mitoxantrone (antineoplastic)	3.38	S	1	1	S
Nalbuphine (analgesic)	2.17	S	1	1	S
Naloxone (narcotic antagonist)	1.29	N	0	0	N
Naltrexone (narcotic antagonist)	1.04	N	0	0	N
Nelfinivir (antiviral)	8.86[a],22.3[b]	S	1	1	S
Neostigmine (cholinergic agent)	1.81[a],2.23[b]	S	1	1	S
Nicardipine (antihypertensive, antianginal)	1.08	N	0	0	N
Nifedipine (antihypertensive, antianginal)	1.26	N	0	0	N
Nitrazepam (anticonvulsant)	1.17	N	0	0	N
Nitrendipine (antihypertensive)	0.8	N	0	0.01	N
Nordazepam (anxiolytic)	0.93	N	0	0	N
Noscapine (antitussive)	1.03	N	0	0	N
Oxprenolol (antihypertensive)	1.37	N	0	0	N
Perphenazine (antipsychotic)	1.47	N	0	0.04	N
Pheniramine (antihistamine)	1.41	N	0	0.04	N
Pirenzapine (anti-ulcerative)	3.63	S	1	1	S
Prazosin (antidepressant)	4.63	S	1	1	S
Procyclidine (antiparkinsonian)	0.95	N	0	0	N
Progabide (anticonvulsant)	0.88	N	0	0.11	N
Promethazine (antihistamine)	1.27	N	0	0	N
Propranolol (antihypertensive)	1.04[a],1.04[b]	N	0	0.02	N
Protriptylene (antidepressant)	2.37	S	1	0.55	S
Puromycin (antibiotic)	3.1	S	1	1	S

ISSN: 2249 –4820

Chemistry & Biology Interface, 2012, 2, 3, 140-156

Quinidine (antiarrythmic)	27.2	S	1	1	S
Reserpine (antihypertensive)	3.71	S	1	1	S
Scopolamine (antiemetic)	1.14	N	0	0	N
Selegiline (antiparkinsonian)	0.76	N	0	0	N
Sulfasalazine (anti-inflammatory, bowel)	1.65	N	0	0	N
Terfenadine (antihistamine)	4.66[a],2.88[b]	S	1	1	S
Testosterone (androgen)	0.73	N	0	0.08	N
Trazodone (antidepressant)	0.94	N	0	0	N
Trimethoprim (antibacterial)	3.61[a],1.94[b]	S	1	1	S
Trimipramine (antidepressant)	0.92	N	0	0	N
Vincristine (antineoplastic)	6.31	S	1	1	S
Vinorelbine (antineoplastic)	69.8	S	1	1	S
Yohimbine (antiadrenergic)	1.17	N	0	0	N
Zolmitriptan (antimigraine)	2.48	S	1	0.97	S
Zolpidem (sedative)	1.14	N	0	0	N

[a]Values taken from (5), [b]Values taken from (21).

[c] Abbreviations: S, substrate; N, non-substrate.

Table 5. Test set compounds; Pgp efflux ratio (B→A/A→B); observed and predicted binary activity of compounds (Substrate-1/ Non-substrate-0).

Compound Name (Therapeutic Indication)	B→A/A→B Ratio[a,b]	Pgp status[c] (exptl.)	Binary activity (observed)	Binary activity (predicted)	Pgp status[c] (predicted)
Alprenolol (antihypertensive)	1.01	N	0	0	N
Antipyrine (local analgesic)	0.94	N	0	0	N
Clonidine (antihypertensive)	0.99	N	0	0	N
Cyclosporin A(immunosuppresives)	9.61	S	1	1	S
Diltiazem (anti-anginal)	1.64[a],1.53[b]	S	1	0	N
Erythromycin (antibiotic)	14.4	S	1	0	N
Guanfacine (antihypertensive)	1.23	N	0	0	N
Indinavir (antiviral)	20.3[a],24.6[b]	S	1	0.9	S
Ketamine (anesthetic)	0.93	N	0	0	N
Levomeprazine (antipsychotic)	1.54	S	1	0.79	S
Loperamide (antidiarrheal)	7.77 [a],9.9 [b]	S	1	1	S
Mephentermine (vasopressor)	0.87	N	0	0	N
Mexilitene (antiarrythmic)	0.87	N	0	0	N
Monensin (antibiotic)	2.88	S	1	1	S
Nortriptylene (antidepressant)	1.39	N	0	0.91	S

ISSN: 2249 –4820

Chemistry & Biology Interface, 2012, 2, 3, 140-156

Practolol (antiarrythmic)	1.32	N	0	0	N
Promazine (antipsychotic)	1.15	N	0	0	N
Pyridostigmine (cholinergic agent)	1.24	N	0	0	N
Ranitidine (anti-ulcerative)	1.35	N	0	0	N
Ritonavir (antiviral)	54.4	S	1	1	S
Sumatriptan (antimigraine)	1.37[a],1.48[b]	N	0	0	N
Tacrine (cognitive stimulant)	0.93	N	0	0	N
Warfarin (anticoagulant)	0.83	N	0	0	N
Zimeldine (antidepressant)	1.01	N	0	0.13	N

[a]Values taken from [5], [b]Values taken from [21].
[c] Abbreviations: S, substrate; N, non-substrate.

Table 6. Experimental mean apical to basolateral permeability P_{app} A to B with BA/AB ratio along with the observed and predicted binary activity for the unpredicted (or outlier) compounds.

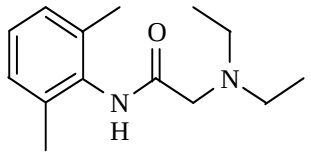
Compound Name	Mean P_{app} A to B[a] (nm/s)	BA/AB Ratio[a,b] (*In vitro* analysis)	Binary activity (observed)	Pgp status[c] (observed)	Binary activity (predicted)	Pgp status[c] (predicted)
Nortriptyline	337[b]	1.39	0	N	0.91	S[d]
Erythromycin	0.94[a]	14.4	1	S	0.00	N[e]
Diltiazem	413[a], 430[b]	1.64, 1.53	1	S	0.00	N[e]

[a]Values taken from [5], [b]Values taken from [21].
[c]Abbreviations S, substrate; N, non-substrate.
[d]False positive; [e]False negative

ISSN: 2249 –4820

Chemistry & Biology Interface, 2012, 2, 3, 140-156

Compound



Binary Classification
[Substrate (1) /Non-substrate (0)]

Descriptors Calculation

QSAR Table

Compounds	Activity	MW	SlogP	chi0	-	-	*n*
1	0	314	0.93	11.33	-	-	-
2	1	410	0.27	15.66	-	-	-
3	0	250	0.67	20.15	-	-	-
4	0	275	0.68	22.36	-	-	-
.	1	-	-	-	-	-	-
.	1	-	-	-	-	-	-
.	0	-	-	-	-	-	-
m	0	-	-	-	-	-	-

Principle Component Analysis
(PCA)

Conditional Probability Analysis
(Bayes' Theorem)

Binary-QSAR Model

Figure 1. Flow chart of binary-QSAR methodology implemented in MOE.

ISSN: 2249 –4820

Chemistry & Biology Interface, 2012, 2, 3, 140-156

References

[1] Chan, L.M.; Lowes, S.; Hirst, B.H. The ABCs of drug transport in intestine and liver: efflux proteins limiting drug absorption and bioavailability. *Eur. J. Pharm. Sci.*, **2004**, *21*, 25-51.

[2] Padowski, J.M.; Pollack, G.M. Pharmacokinetic and pharmacodynamic implications of P-glycoprotein modulation. *Methods Mol. Biol.*, **2010**, *596*, 359-384.

[3] Varma, M.V.; Sateesh, K.; Panchagnula, R. Functional role of P-glycoprotein in limiting intestinal absorption of drugs: contribution of passive permeability to P-glycoprotein mediated efflux transport. *Mol. Pharm.* **2005,** *2*, 12-21.

[4] H. Potschka. Targeting regulation of ABC efflux transporters in brain diseases: a novel therapeutic approach. *Pharmacol. Ther.*, **2010**, *125*, 118–127.

[5] Polli, J.W.; Wring, S.A.; Humphreys, J.E.; Huang, L.; Morgan, J.B.; Webster, L.O.; Serabjit-Singh, C.S. Rational use of in vitro P-glycoprotein assays in drug discovery. *J. Pharmacol. Exp. Ther.*, **2001**, *299*, 620-628.

[6] Varma, M.V.; Ambler, C.M.; Ullah M.; Rotter, C.J.; Sun, H.; Litchfield, J.; Fenner, K.S.; El-Kattan, A.F.; Targeting intestinal transporters for optimizing oral drug absorption. *Curr. Drug Metab.*, **2010**, 11,730-742.

[7] Seelig, A. A general pattern for substrate recognition by P-glycoprotein. *Eur. J. Biochem.*, **1998**, *251*, 252-261.

[8] Chen, L.; Li, Y.; Yu, H.; Zhang, L.; Hou, T. Computational models for predicting substrates or inhibitors of P-glycoprotein. *Drug Discov.Today*, **2011,** in press.

[9] Li-Blatter, X.; Beck, A.; Seelig A. P-Glycoprotein-ATPase Modulation: The Molecular Mechanisms. *Biophy. J.*, **2012**, 102, 1383-1393.

[10] Schwaha, R.; Ecker, G.F. Use of shape similarities for the classification of P-glycoprotein substrates and nonsubstrates. *Future Med. Chem.*, **2011,** 3, 1117-1128.

[11] Estrada, E.; Molina, E.; Nodarse, D.; Uriarte, E. Structural contributions of substrates to their binding to P-Glycoprotein. A TOPS-MODE approach. *Curr. Pharm. Des.*, **2010,** 16, 2676-2709.

[12] Wang, Z.; Chen, Y.; Liang, H.; Bender, A.; Glen, R.C.; Yan, A. P-glycoprotein substrate models using support vector machines based on a comprehensive data set. *J. Chem. Inf. Model.*, **2011,** 51, 1447-1456.

[13] Hammann, F.; Gutmann, H.; Jecklin, U.; Maunz, A.; Helma, C.; Drewe, J. Development of decision tree models for substrates, inhibitors, and inducers of p-glycoprotein. *Curr. Drug Metab.*, **2009,** 10, 339-346.

[14] Broccatelli, F.; Carosati, E.; Neri, A.; Frosini, M.; Goracci, L.; Oprea, T.I.; Cruciani, G. A novel approach for predicting P-glycoprotein (ABCB1) inhibition using molecular interaction fields. *J. Med. Chem.*, **2011,** 54, 1740-1751.

[15] Wang, Y.H.; Li Y.; Yang, S.L.; Yang, L. Classification of substrates and inhibitors of P-glycoprotein using unsupervised machine learning approach, *J. Chem. Inf. Model*, **2005**, *45*, 750-757.

[16] Didziapetris, R.; Japertas, P.; Avdeef, A.; Petrauskas, A. Classification analysis of P-glycoprotein substrate specificity. *J. Drug Target*, **2003**, *11*, 391-406.

[17] Gombar, V.K.; Polli, J.W.; Humphreys, J.E.; Wring, S.A.; Serabjit-Singh, C.S. Predicting P-glycoprotein substrates by a quantitative structure-activity relationship model. *J. Pharm. Sci*., **2004**, *93*, 957-968.

[18] Cianchetta, G.; Singleton, R.W.; Zhang, M.; Wildgoose, M.; Giesing, D.; Fravolini, A.; Cruciani, G.; Vaz R.J. A pharmacophore hypothesis for P-glycoprotein substrate recognition using GRIND-based 3D-QSAR. *J. Med. Chem.*, **2005**, *48*, 2927-2935.

[19] Cabrera, M.A.; Gonzalez, I.; Fernandez, C.; Navarro, C.; Bermejo, M. A topological substructural approach for the prediction of P-glycoprotein substrates. *J. Pharm. Sci.*, **2006**, *95*, 589-606.

[20] Crivori, P.; Reinach, B.; Pezzetta, D.; Poggesi, I. Computational models for identifying potential P-glycoprotein substrates and inhibitors. *Mol. Pharm*., **2006**, *3*, 33-44.

[21] Mahar Doan, K.M.; Humphreys, J.E.; Webster, L.O.; Wring, S.A.; Shampine, L.J.; Serabjit-Singh, C.J.; Adkison, K.K.; Polli, J.W. Passive permeability and P-glycoprotein-mediated efflux differentiate central nervous system (CNS) and non-CNS marketed drugs. *J. Pharmacol. Exp. Ther*., **2002**, *303*, 1029-1037.

[22] CambridgeSoft Co., 100 CambridgePark Drive, Cambridge, MA 02140 USA. http://www.cambridgesoft.com

[23] MOE, Chemical Computing Group Inc., Montreal, H3A 2R7 Canada, http://www.chemcomp.com.

[24] http://www.accelrys.com

[25] Wildman, S.A.; Crippen, G.M.; Prediction of Physicochemical Parameters by Atomic Contributions. *J. Chem. Inf. Comput. Sci*., **1999**, *39*, 868-873.

[26] Lipinski, C.A.; Lombardo, F.; Dominy, B.W.; Feeney, P.J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev*., **2001**, 46, 3-26.

[27] Hall, L.H.; Kier, L.B. The nature of structure-activity relationships and their relation to molecular connectivity. *Eur. J. Med. Chem*., **1997**, 12, 307-312.

[28] Hall, L.H.; Kier, L.B. The Molecular Connectivity Chi indexes and Kappa Shape Indexes in Structure-Property Modeling. *Reviews of Computational Chemistry*., **1991**, *2*, 367-422.

[29] Cheng, A.; Merz, Jr.K. Prediction of aqueous solubility of a diverse set of compounds using quantitative structure-property relationships. *J. Med. Chem*., **2003**, *46*, 3572-3580.

[30] Susnow, R.G.; Dixon, S.L. Use of robust classification techniques for the prediction of human cytochrome P450 2D6 inhibition. *J. Chem. Inf. Comput. Sci*., **2003**, *43*, 1308-1315.

[31] Ghose, A.K.; Viswanadhan, V.N.; Wendoloski, J.J. Prediction of Hydrophobic (Lipophilic) Properties of Small Organic Molecules Using Fragmental Methods:

ISSN: 2249 –4820

Chemistry & Biology Interface, 2012, 2, 3, 140-156

An Analysis of ALOGP and CLOGP Methods. *J. Phys. Chem*., **1998**, *102*, 3762-3772.

[32] Labute, P. Binary QSAR: A new method for the determination of quantitative structure activity relationships. *Pac. Symp. Biocomput*., **1999**, 444-455.

[33] Gao, H.; Lajiness, M.S.; Van Drie, J. Enhancement of binary QSAR analysis by a GA-based variable selection method. *J. Mol. Graph. Model*, **2002**, *20*, 259-268.

[34] Labute, P.; Nilar, S.; Williams, C. A probabilistic approach to high throughput drug discovery. *Comb. Chem. High Throughput Screen*, **2002**, *5*, 135-145.

[35] Gao, H.; Williams, C.; Labute, P.; Bajorath, J. Binary quantitative structure-activity relationship (QSAR) analysis of estrogen receptor ligands. J. Chem. Inf. Comput. Sci., **1999**, *39*, 164-168.

[36] Varma, M.V.; Sateesh, K.; Panchagnula, R. Functional role of P-glycoprotein in limiting intestinal absorption of drugs: contribution of passive permeability to P-glycoprotein mediated efflux transport. Mol. Pharm., **2005**, *2*, 12-21.

[37] Stouch, T.R.; Gudmundsson, O. Progress in understanding the structure-activity relationships of P-glycoprotein. *Adv. Drug Deliv. Rev*., **2002**, *54*, 315-328.

[38] Varma, M.V.; Ashokraj, Y.; Dey, C.S.; Panchagnula, R. P-glycoprotein inhibitors and their screening: a perspective from bioavailability enhancement. Pharmacol. Res., **2003**, *48*, 347-359.

[39] Wanchana, S.; Yamashita, F.; Hara, H.; Fujiwara, S.; Akamatsu, M.; Hashida, M. Two and Three Dimensional QSAR of carrier-mediated transport of beta-lactam antibiotics in Caco-2 cells. *J. Pharm. Sci*., **2004**, *93*, 3057-3065.