ELSEVIER

Computational models for predicting substrates or inhibitors of P-glycoprotein

Lei Chen, Youyong Li, Huidong Yu, Liling Zhang and Tingjun Hou

Institute of Functional Nano & Soft Materials (FUNSOM) and Jiangsu Key Laboratory for Carbon-Based Functional Materials & Devices, Soochow University, Suzhou, Jiangsu 215123, China

The impact of P-glycoprotein (P-gp) on the multidrug resistance and pharmacokinetics of clinically important drugs has been widely recognized. Here, we review *in silico* approaches and computational models for identifying substrates or inhibitors of P-gp. The advances in the datasets for model building and available computational models are summarized and the advantages and drawbacks of these models are outlined. We also discuss the impact of the recently reported crystal structures of P-gp on potential breakthroughs in the computational modeling of P-gp substrates. Finally, the challenges of developing reliable prediction models for P-gp inhibitors or substrates, as well as the strategies to surmount these challenges, are reviewed.

Introduction

P-glycoprotein (P-gp), a member of ATP-binding cassette (ABC) transporter family, significantly impacts the multidrug resistance (MDR) phenomenon and absorption, distribution, metabolism and elimination (ADME) properties of drugs [1–7]. It is normally expressed at many physiological barriers, including the intestinal epithelium, hepatocytes, renal proximal tubular cells, the adrenal gland and endothelial capillaries of the brain comprising the blood–brain barrier (BBB) [8], as well as being commonly over-expressed in tumor cell lines [9]. It is now clear that P-gp can transport many chemically and structurally unrelated drugs and agents [10], resulting in the MDR phenomenon that accounts for chemotherapeutic failure in the treatment of cancers. Moreover, P-gp functions as an energy-dependent hydrophobic efflux pump that exports a large number of hydrophobic compounds from cells [11]. It is observed that apical expression of P-gp in many human organs results in reduced drug intestinal absorption, and enhanced elimination into bile (liver) and urine (kidney) [3,12,13]. Therefore, P-gp has a great impact on the ADME properties of a variety of drugs [14].

P-gp can interact with large numbers of structurally diverse compounds, which suggests its multiple binding sites of different chemical environments. According to the interactions, these compounds can be classified into three categories: substrates, inhibitors and modulators [15]. Compounds actively transported by P-gp are classified as substrates, whereas those that compromise the transporting function of P-gp are classified as inhibitors. Modulators interact with the binding sites distinct from those of substrates, therefore reducing substrate binding through a negative allosteric interaction. Modulators and inhibitors exert the same final biological effect, restoring cell sensitivity to chemotherapeutic agents. Therefore, the term of inhibitor is often used synonymously with that of modulator [16].

Owing to the importance of P-gp on MDR and ADME, extensive studies have been carried out to identify P-gp substrates or develop more-potent, -selective and -specific P-gp inhibitors [6]. The polyspecificity (i.e. promiscuity) of P-gp in substrate and inhibitor recognition makes designing effective candidate compounds difficult [17]. Traditionally, experimental assays were used to assess the interactions and transport of new chemical entities with P-gp [6]. However, these experimental assays are expensive, laborious and time-consuming. Therefore, *in silico* models that provide rapid and cheap screening platforms for identifying P-gp inhibitors or substrates have been recognized to be valuable tools [18–20]. Numerous computational approaches or models based on quantitative structure–activity relationship (QSAR) analyses, pharmacophore modeling and molecular docking were developed to predict P-gp inhibitors or substrates [21–41]. The transporting mechanism, substrate properties and

Corresponding author:. Hou, T. (tingjunhou@hotmail.com), (tjhou@suda.edu.cn)

1359-6446/06/$ - see front matter doi:10.1016/j.drudis.2011.11.003

computational models for ABC transporters have been discussed in several reviews [42–46].

In this review, we survey the recent advances in computational approaches developed for the prediction of P-gp inhibitors or substrates. First, the structural characteristics of P-gp and the mechanism of the P-gp efflux pump are briefly introduced. Then, two fundamental aspects for the *in silico* predictions of P-gp inhibitors or substrates are systematically summarized, including the available datasets of P-gp inhibitors or substrates and the published computational models. In addition, we discuss the benefits of using the P-gp crystal structure and molecular docking approach for predicting P-gp substrates. Unfortunately, it was found that the docking scores could not distinguish substrates from non-substrates, suggesting that current molecular docking techniques could not deal effectively with the complex nature and the weak and unspecific ligand-binding properties of P-gp.

The structure of P-gp and the mechanism of P-gp transport

Throughout the past decade, the structure of P-gp has usually been characterized by homology modeling techniques [47–49]. However, earlier attempts to model the 3D structure of P-gp suffered from low sequence identity to the template protein, a prime example being the bacterial ABC transporter MsbA [48,49]. The lack of a reliable crystal structure becomes a major obstacle to design anti-MDR inhibitors. Encouragingly, in 2009, the X-ray structure of *apo* murine P-gp (PDB entry: 3G5U; resolution: 3.8 Å) and two additional P-gp X-ray structures in complex with two cyclopeptidic inhibitors (PDB entries: 3G60 and 3G61; resolution: 4.40 Å and 4.35 Å) were reported by Aller *et al.* [50]. The murine P-gp shares 87% sequence identity with the human homology. P-gp is a pseudo-symmetrical heterodimer with each monomer consisting of two bundles of six transmembrane (TM) helices (TMs 1–3, 6, 10, 11 and TMs 4, 5, 7–9, 12) and two nucleotide-binding domains (NBDs) separated by ~30 Å (Fig. 1). The function of NBDs is to bind ATP, thus providing energy for substrate binding. Two bundles of six TM helices form the inward-facing conformation that results in a large internal cavity open to the cytoplasm and the inner leaflet. TMs 4, 6 and TMs 10, 12 form two portals that provide access for the entry of hydrophobic molecules directly from the membrane. The crystal structures of P-gp in complex with QZ59 show that *RRR*- and *SSS*-QZ59 have different binding locations and orientations: *RRR*-QZ59 binds to one site per transporter whereas *SSS*-QZ59 binds to two sites, confirming the polyspecificity of P-gp.

To date, the mechanism for P-gp transporting substrates out of cells is still subject to considerable controversy [51–54]. One classic hypothesis was proposed to interpret the mechanism of the energy-driven drug transporter: (i) the inward-facing region of P-gp has high affinity to substrates and binds substrates using the energy provided by NBD dimerization; (ii) P-gp undergoes large structural changes from an inward-facing to an outward-facing conformation during the catalytic cycle; (iii) substrates are released as a consequence of decreased binding affinity caused by the changes in specific residue contacts or, alternatively, facilitated by ATP hydrolysis, which could disrupt NBD dimerization and reset the system back to inward-facing and reinitiate the transport cycle (Fig. 1) [55,56].

The fact that large numbers of structurally diverse compounds interact with P-gp suggests that multiple binding sites could be involved in substrate and/or inhibitor binding [16]. The type and number of binding sites is still not clear [56,57]. Two 'functional' drug-binding sites have been identified within P-gp, based on their mutual interactions in the transport process; the H-site, which binds Hoechst 33342, and the R-site, which binds rhodamine 123 (R123) [57]. The X-ray crystallographic studies of P-gp showed that two drugs can bind to different 'small' binding sites in a single large flexible binding pocket [50]. Another regulatory site was also found in P-gp, and the binding to this site led to a dramatic change in the properties of the transported substrate-binding site [58].

In silico predictions of P-gp inhibitors or substrates

Rather than developing computational models based on complicated statistical techniques, earlier attempts have been made to find a set of simple rules based on structural and functional

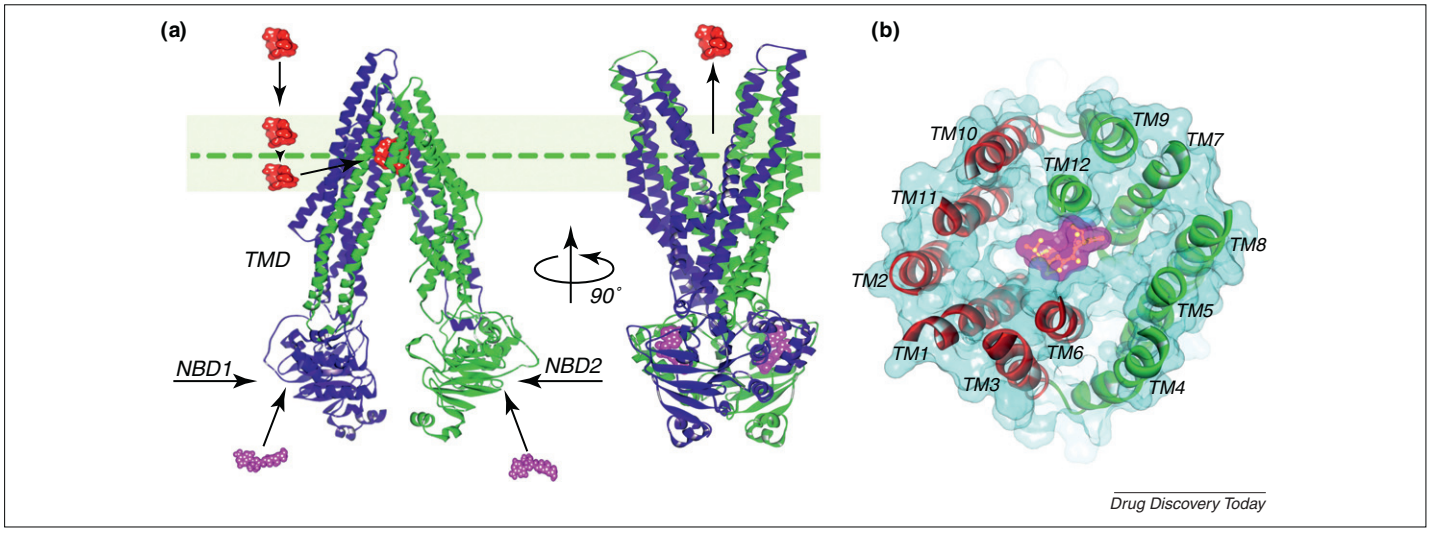

FIGURE 1

(a) Transport cycle for substrate efflux pumped by P-glycoprotein. Substrates are colored red and ATP is magenta. **(b)** Ligand-binding site on the inward facing conformation of human P-glycoprotein. The cyclopeptidic inhibitor *RRR*-QZ59 is co-crystallized and shown in magenta. The human P-glycoprotein (P-gp) was modeled based on the crystal structure of murine P-gp (PDB entry: 3G60) using the Modeller program in Discovery Studio (version 2.5).

TABLE 1

The theoretical models for predicting P-glycoprotein substrates and inhibitors

Refs	Method	Model	Descriptors	Dataset		Performance
				Training	Test	
Penzotti [29]	CONAN	Classification	Pharmacophore-based descriptors	144	45	Training[a]: accuracy = 80%; test[b]: accuracy = 63%
Gombar [24]	LDA	Classification	Electrotopological state values, shape indices and molecular properties	95	58	Training: SE[c] = 100%, *SP*[d] = 90.6% test: accuracy = 86.2%
Xue [35]	SVM	Classification	159 descriptors	74	25	SE = 84.2%, SP = 66.7%, accuracy = 80%
Crivori [22]	PLSD	Classification	Volsurf descriptors	53	272	Training: accuracy = 88.7%; test: accuracy=72.4%
Sun [31]	Bayes	Classification	Atom typing descriptors and fingerprints	424	185	Test: accuracy=82.2%
Cabrera [36]	TOPS-MODE	Classification	TOPS-MODE descriptors	163	40	Training: SE = 82.4%, SP = 79.17%, accuracy = 80.9%; test: accuracy = 77.5%
Wang [32]	BRNN	Correlation	249 descriptors	43	14	Training: r^2 = 0.756, test: r^2 = 0.728
Lima [27]	SVM, *k*NN, DT, binary QSAR	Classification	MolconnZ, AP, VolSurf and MOE Descriptors	144	51	Training: accuracy = 94%; test: accuracy = 81%
Huang [39]	SVM, PS	Classification	79 descriptors	163	40	Training: 95.5%; test: 90%
Muller [64]	PLS	Correlation	CoMFA and CoMSIA descriptors	28	30	Training: r^2 = 0.82; test: r^2 = 0.6
Wu [34]	MLR, SVM	Correlation	423 CODESSA descriptors	56	14	Training: r^2 = 0.85; test: r^2 = 0.81
Chen [21]	RP, NBC	Classification	Fingerprints and molecular properties	973	300	Training: SE = 84.7%, SP = 82.1%, accuracy = 88.9%; test: SE = 79.2%, SP = 83.8%, accuracy = 81%
Cianchetta [37]	PLS	Correlation	Almond and Volsurf descriptors	109	20	Training: r^2 = 0.83, LOO q^2 = 0.75; test: r^2=0.72
Ekin [23]		Pharmacophore		27	19	Training: r^2 = 0.77; test: Spearman r = 0.68
				21	19	Training: r^2 = 0.88; test: Spearman r = 0.7
				17	19	Training: r^2 = 0.86; test: Spearman r = 0.46
Li [26]	DT	Classification	Pharmacophore models	163	97	Training: accuracy = 87.7%; test: accuracy = 87.6%
Wang [41]	SVM	Classification	ADRIANA.Code, MOE and ECFP4 fingerprints	212	120	Training: LOO accuracy = 75%; test: accuracy = 88%

[a] Training represents training set.
[b] Test represents test set.
[c] SE represents sensitivity.
[d] SP represents specificity.

features that can characterize the interactions between a substrate or inhibitor and P-gp [12,59–61]. For example, Seelig suggested a set of well-defined structural elements required for an interaction with P-gp [61]. These recognition elements were formed by two (type I unit) or three (type II unit) electron donor groups with a fixed spatial separation. Type I units consisted of two electron donor groups with a spatial separation of 2.5 ± 0.3 Å, and type II units contained either two electron donor groups with a spatial separation of 4.6 ± 0.6 Å or three electron donor groups with a spatial separation of the outer two groups of 4.6 ± 0.6 Å. All molecules that contained at least one unit (i.e. type I or type II) were predicted to be P-gp substrates. These simple rules can be understood easily and used by laboratory scientists as well as computational chemists; however, they are too simple to characterize P-gp substrates or inhibitors effectively [26].

Extensive computational models, based on 2D-QSAR, 3D-QSAR, pharmacophore modeling and molecular docking techniques, have been developed to predict P-gp inhibitors or substrates. The theoretical models reported for predicting P-gp inhibitors or substrates are summarized in Table 1. Moreover, a variety of statistical techniques as well as machine learning approaches, including multiple linear regression (MLR) [27], partial least square discriminant analysis (PLSD) [22], linear discriminant analysis (LDA) [24,36], decision tree (DT) [21], support vector machine (SVM) [34,35,39,41], Kohonen self-organizing map (SOM) [33] and Bayesian classifier [21,31], have been employed to develop the theoretical models.

Experimental datasets for model developments

The preparation of relevant datasets with high quality and quantity is the first step toward constructing models with high confidence. Traditionally, the public datasets only have a limited number of compounds (less than or close to 200) [29,32,35,36,39]. Gombar *et al.* compiled a dataset of 98 molecules, which consists of 32 non-substrates and 66 substrates

identified by *in vitro* monolayer efflux assays [24]; Penzotti *et al.* reported a dataset of 195 compounds, among which are 108 P-gp substrates and 87 non-substrates [29]; Xue *et al.* assembled a dataset of 201 compounds, which includes 116 substrates and 85 non-substrates of P-gp [35]. In 2005, Sun reported an extensive validated dataset of 609 compounds provided by Dr Klopman [31]. The reversal factor (RF) was used to measure the ability to reverse MDR. From the 609 compounds in the dataset, 378 compounds were active, with an RF value greater than 2.0, and the remaining compounds were inactive. In 2011, Wang *et al.* reported a large dataset of 332 compounds, which included 206 P-gp substrates and 126 non-substrates [41].

Recently, we reported the largest dataset for P-gp inhibitors and non-inhibitors available to date [21]. This dataset has 1273 structurally diverse molecules, consisting of 797 P-gp inhibitors and 476 non-inhibitors. The two most important sources are the experimental data of 609 compounds with the multiple drug resistance reversal (MDRR) activity reported by Bakken and Jurs [62], and the experimental data of 347 compounds reported by Ramu and Ramu [63,64]. The experimental MDRR ratio was used as a criterion to determine whether a compound is an inhibitor or not: if the MDRR ratio was less than 4 the compound was categorized into the non-inhibitor class; if the MDRR ratio was greater than 5 the compound was categorized into the inhibitor class; if the MDRR ratio was $\geq$4 and $\leq$5 the compound was considered to be moderately active and not included in the dataset (the dataset is available at: http://cadd.suda.edu.cn/admet, accessed November 2011). It is worth noting that the assays used for assessing P-gp inhibition do not truly reflect a direct measure of P-gp inhibition. However, based on the primary assumption that only the EC_{50} shift is caused by the inhibition of P-gp, the assay is one of the classical approaches used within the field of oncology.

The available datasets do not appear to be robust and reliable, because the data from different experimental protocols are usually mixed together. The class (i.e. inhibitor or non-inhibitor, and substrate or non-substrate) of a compound needs to be checked carefully if it can be found in multiple publications [21].

Theoretical models based on QSAR

In 2004, Gombar *et al.* employed LDA to construct a classification model for P-gp substrates [24]. The training set consisting of 95 compounds was classified as 63 substrates and 32 non-substrates based on the results from *in vitro* monolayer efflux assays. The LDA classifier with 27 descriptors gave a sensitivity of 100% and a specificity of 90.6% in the cross-validation test; moreover, a prediction accuracy of 86.2% was obtained on an external test set of 58 compounds. The analysis of these 27 descriptors in the final classifier suggested that the ability to partition into membranes, molecular bulkiness and the counts and electrotopological values of certain isolated and bonded hydrides were important structural features of substrates. However, the training set used by Gombar *et al.* was not extensive enough; therefore, the chemical space covered by the current model might be limited.

In 2005, Cabrera *et al.* developed a linear discriminant model for a dataset of 163 compounds, which includes 91 substrates and 72 non-substrates [36]. The final model based on nine TOPS-MODE descriptors achieved a sensitivity of 82.42% and a specificity of 79.17% for the training set. For the external validation set with 40 compounds (22 substrates and 18 non-substrates), the model gave a prediction accuracy of 77.5%. Analysis of the descriptors in the model evidenced that the standard bond distance, the polarizability and the Gasteiger–Marsilli atomic charge affected the interaction between P-gp and substrates.

In 2006, Crivori *et al.* applied PLSD analysis to classify 22 P-gp substrates and 31 P-gp non-substrates based on the VolSurf descriptors [22]. The model had an accuracy of 88.7% for the training set, but it only achieved an accuracy of 72.4% for the external set of 272 compounds. Then the authors applied PLSD analysis to construct the classifier to distinguish P-gp substrates from P-gp inhibitors based on the GRIND descriptors. The classifier discriminated between 69 substrates and 56 inhibitors with an average accuracy of 82%.

In 2005, Cianchetta *et al.* developed a 3D-QSAR model using the Almond and VolSurf descriptors for a diverse set of 129 compounds, which were evaluated for P-gp inhibition using the calcein-AM method assay [37]. These compounds were divided into a training set of 109 compounds and a test set of 20 molecules. Statistical analysis showed that after Fractional factorial design (FFD) fractional selection has been implemented the PLSD model with three latent variables gave the best prediction for the training set: $r^2 = 0.8252$; leave-one-out (LOO) $q^2 = 0.7459$; leave-two-out (LTO) $q^2 = 0.7456$; and random grouping (RG) $q^2 = 0.7400$. It is encouraging that this model achieved a square correlation coefficient of $r^2 = 0.7160$ for the tested molecules. By analyzing the Almond descriptors in the PLSD model, the authors proposed the following pharmacophore hypothesis: two hydrophobic groups 16.5 Å apart and two hydrogen-bond-acceptor groups 11.5 Å apart.

In 2005, Sun built a naive Bayesian classifier to categorize MDRR agents into active and inactive classes based on atom-type-based molecular descriptors and fingerprints [31]. The whole dataset was split into a training set of 424 molecules and a test set of 185 molecules. The classifier built from the training set predicted the MDRR activities of the tested compounds with a success rate of 82.2%. The author believed that the model based on atom-typing descriptors and naive Bayesian classification offered extra information for the rational design of MDRR agents.

In 2006, Lima *et al.* [27] developed a set of classification models for a dataset of 195 diverse substrates and non-substrates by employing various combinations of optimization methods and descriptor types [27,29]. In the modeling process, four descriptor sets were used, including 381 molecular connectivity indices, 173 atom pair (AP) descriptors, 72 VolSurf descriptors and 189 descriptors calculated by Molecular Operating Environment (MOE), and four modeling techniques were used, including, *k*-nearest neighbors (*k*NN) classification QSAR, binary QSAR, DT and SVM. Every QSAR modeling technique was combined with each descriptor type to create 16 (4 methods $\times$ 4 descriptors) combinatorial QSAR models. The best models based on SVM and either AP or VolSurf descriptors achieved high correct classification rates with 94% and 81% for the training and test sets, respectively.

In 2008, Muller *et al.* developed 3D-QSAR models using Comparative Molecular Field Analysis (CoMFA) and Comparative Molecular Similarity Index Analysis (CoMSIA) approaches for 28 P-gp inhibitors, including 24 structurally related derivatives of tariquidar and four XR compounds [65]. The best 3D-QSAR models achieved an internal predictive squared correlation coefficient higher than

0.8. The models were then validated by an external test set of 30 XR compounds, and the best CoMSIA model gave a predictive squared correlation coefficient of 0.6. It should be noted that the CoMFA and CoMSIA models were developed based on a series of homologs; therefore, they do not have general predictive capability.

In 2009, Wu *et al.* applied MLR and SVM techniques to construct the hybrid QSAR models to predict MDR modulating activities for 70 compounds [34]. First, the heuristic method was applied to select the descriptors using the CODESSA^{TM} program, and then the prediction models were built by MLR, SVM and hybrid QSAR modeling. The best hybrid model gave root-mean-square (RMS) errors of 0.33 units for the training set, 0.47 for the test set and 0.36 for the whole set, and the corresponding correlation coefficients (r^2) were 0.85, 0.81 and 0.84, respectively. One big concern regarding Wu's models is whether the application domain is large enough for the compounds outside the chemical space covered by the training set because the dataset used in modeling is small.

Recently, Wang *et al.* built several classification models to predict whether a compound is a P-gp substrate or not, based on a large dataset with 332 distinct structures [41]. Each molecule was represented by three sets of molecular descriptors, including ADRIANA.Code, MOE and ECFP_4 fingerprints. The classification models were constructed by SVM based on a training set, which includes 131 P-gp substrates and 81 P-gp non-substrates. The best model gave 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 that have significance in separating substrates and non-substrates.

More recently, we developed a set of classification models for a large dataset of 1273 molecules [21]. The whole dataset was randomly split into a training set of 973 molecules and a test set of 300 molecules. First, the DTs were built from the training set using the recursive partitioning (RP) technique and validated by an external test set of 300 compounds. The best DT correctly predicted 83.5% of the P-gp inhibitors and 67% of the P-gp non-inhibitors in the test set. Second, the naive Bayesian categorization modeling was applied to establish classifiers for the P-gp inhibitors and non-inhibitors. The Bayesian classifier displayed an average correct prediction for 81.7% of 973 compounds in the training set with LOO cross-validation procedure and 81.2% of 300 compounds in the test set. By establishing multiple Bayesian classifiers with and without molecular fingerprints, the impact of molecular fingerprints on classification was evaluated by the prediction accuracy for the test set. We found that the inclusion of molecular fingerprints could improve the prediction significantly. Moreover, as an unsupervised learner without tuning parameters, the Bayesian classifier employing fingerprints highlights the important structural fragments favorable or unfavorable for P-gp inhibition. These important fragments predicted by the Bayesian classifier could be useful for experimental scientists when designing molecules with better P-gp inhibition. The 15 good and 15 bad fragments ranked by the Bayesian scores are summarized in Fig. 2. The relationships between important fragments and P-gp inhibition have been discussed by Chen *et al.* [21].

Theoretical models based on pharmacophore modeling

In 2002, Ekins *et al.* proposed a set of pharmacophore models for P-gp inhibitors [38]. The pharmacophore models generated from the inhibition of digoxin transport in Caco-2 cells, vinblastine and calcein accumulation in P-gp-expressing LLC-PK1 cells, as well as vinblastine binding in vesicles derived from CEM/VLB100 cells were used to rank the experimental data for the inhibition of verapamil binding in Caco-2 cells. The pharmacophore model based on 27 inhibitors of digoxin transport in Caco-2 cells consisted of four hydrophobes and one hydrogen-bond acceptor. This model possessed an observed versus predicted correlation of $r^2 = 0.77$ for the training set and high prediction accuracy for the tested molecules. Moreover, the authors found that the digoxin pharmacophore model could give a good rank for the data from the inhibition of verapamil binding. All five P-gp inhibitor pharmacophores were merged to uncover the features that occupy similar regions in space. This analysis suggested the presence of at least four distinct groups of features, consisting of two hydrophobic domains along with a hydrogen-bond acceptor region and an aromatic ring region, both of which were near one of the hydrophobic domains [23].

In 2002, Pajeva and Wiese developed a general pharmacophore model using the GASP program developed by Tripos for 18 structurally diverse MDR substrates and modulators that bind to the verapamil binding site of P-gp [28]. The pharmacophore model was composed of two hydrophobic points, three hydrogen-bond acceptor points and one hydrogen-bond donor point. They proposed a hypothesis to explain the broad structural variety of the P-gp substrates and inhibitors: (i) the verapamil binding site of P-gp has several points that are involved in hydrophobic and hydrogen-bond interactions; (ii) different drugs can interact with different receptor points in different binding modes.

In 2002, Penzotti *et al.* developed a multiple-pharmacophore model, composed of a set of two-to-four-point pharmacophores to discriminate between P-gp substrates and non-substrates [29]. The whole dataset of 195 compounds for pharmacophore modeling was split randomly into a training set of 144 compounds and a test set of 51 compounds. The final multiple-pharmacophore model was composed of 100 two-, three- and four-point pharmacophores. These compounds matching at least 20 of the 100 pharmacophores in the ensemble were likely to be P-gp substrates. The model offered an overall classification accuracy of 80% for the training set, but only 63% for the test set.

In 2004, Langer *et al.* constructed a general pharmacophore model for inhibitors of P-gp based on a training set of 15 propafenone-type modulators [25]. The pharmacophore model consisted of one hydrogen-bond acceptor, one hydrophobic core, two aromatic hydrophobic areas and one positive ionizable group. The model was validated by 105 compounds from an in-house library. The 105 propafenone-type inhibitors in the test set were ranked according to their EC_{50} values. Within the top 30% compounds ($n = 35$) only three were incorrectly predicted; and within the bottom 30% compounds ($n = 35$) 28 substances (80%) were predicted as being completely inactive.

Similar to Penzotti's work [29], Li *et al.* developed multiple pharmacophore models for differentiating P-gp substrates and non-substrates [26]. A comprehensive set of four-point pharmacophores was generated based on 163 compounds (91 substrates and 72 non-substrates). Nine significant pharmacophores were applied to generate a simple classification tree. The analysis of multiple pharmacophores revealed that hydrogen-bond acceptor,

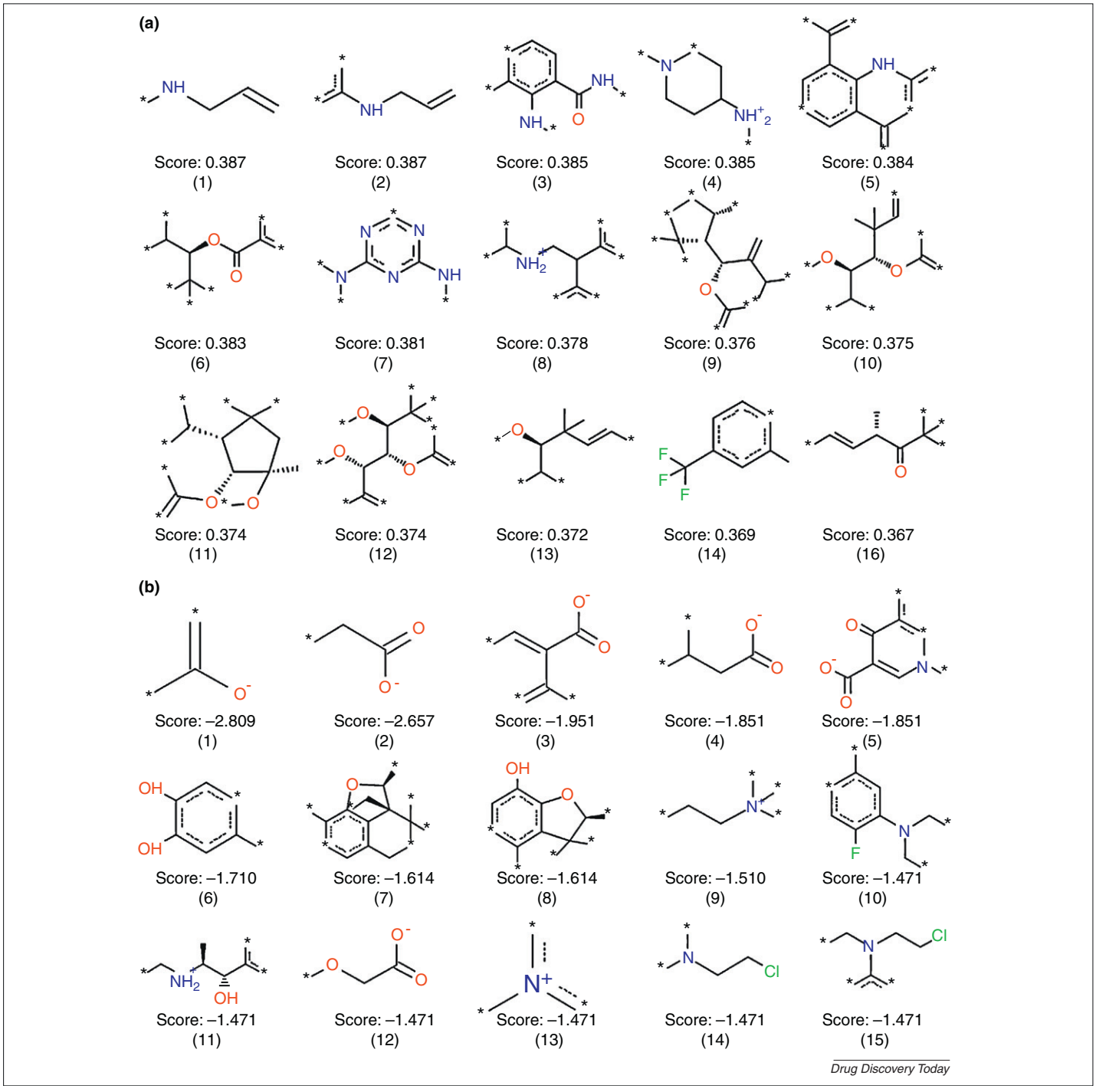

FIGURE 2

(a) The 15 good and **(b)** 15 bad fragments for P-glycoprotein inhibition identified by the Bayesian classifier based on molecular properties and the FCFP_4 fingerprint set.

positive ionizable, aromatic ring and hydrophobic groups were essential features for substrate activity. The classification tree achieved an overall accuracy of 87.7% for the training set and 87.6% for the external test set of 97 molecules.

Theoretical models based on molecular docking

In the earlier stages of P-gp study, the QSAR and pharmacophore modeling techniques were the usual methods used to predict the P-gp inhibitors or substrates owing to the lack of available crystal structures for P-gp. In 2009, the X-ray structures of murine P-gp were reported by Aller *et al.* [50]. The crystal P-gp structures provide good starting points for molecular docking studies.

In many studies, the homology models were developed to characterize the putative ligand-binding sites or investigate the possible conformations of P-gp in different states [49,66–68]. However, only a few publications showed the P-gp models were used to dock compounds into the putative ligand-binding sites [40,69].

In 2009, Becker *et al.* presented four 3D models of P-gp describing two different states along the catalytic cycle using the X-ray structures of Sav1866 and MsbA as the templates in homology modeling [69]. The inter-residue distances of the theoretical models correlated well with distances derived from cross-linking data. One of the nucleotide-free 3D models was used to dock four different ligands, including verapamil, rhodamine B, colchicines and vinblastine, into the central binding cavity harbored by the TM domains. The docked poses for each ligand were found to interact with the residues that have been experimentally identified as binding to a specific ligand. Docking studies indicate that no access route is large enough to allow the entry of one ATP molecule into the catalytic site of the nucleotide-bound models suggesting that these structures should undergo changes to accommodate their ligands. However, the binding poses of the studied ligands given by theoretical predictions could not be validated by solid experimental evidence.

Recently, Pajeva *et al.* tried to dock a series of compounds into the P-gp-binding cavity based on the recently resolved P-gp structure [40]. The docked structures confirmed the P-gp pharmacophoric features identified, and revealed the interactions of some functional groups and atoms in the structures with particular protein residues. However, the accuracy of docking could not be evaluated because the authors did not compare the docking scores with the experimental binding affinities.

Is having the crystal structure of P-gp enough to predict substrate binding?

The polyspecificity of P-gp could be the main obstacle to carrying out molecular docking studies for P-gp. The cavity formed by P-gp encloses a volume of ~6000 $Å^3$ [50], which provides ample space for P-gp to bind two or even more small molecules simultaneously.

To evaluate the prediction capability of molecular docking, we docked 245 diverse molecules, comprising 157 P-gp substrates and 88 P-gp non-substrates collected from the literature [24,33,35,61], into the binding cavity of P-gp. Two crystal structures of P-gp in complex with *RRR*-QZ59 and *SSS*-QZ59 (PDB entries: 3G60 and 3G61 [50]) were used as the receptor models in molecular docking studies. The molecular docking was accomplished by the Glide package (Schrödinger, version 2010). All the structures were docked and scored by two Glide precision modes: SP (standard precision) and XP (extra precision). For each structure, the binding pose with the best score was saved.

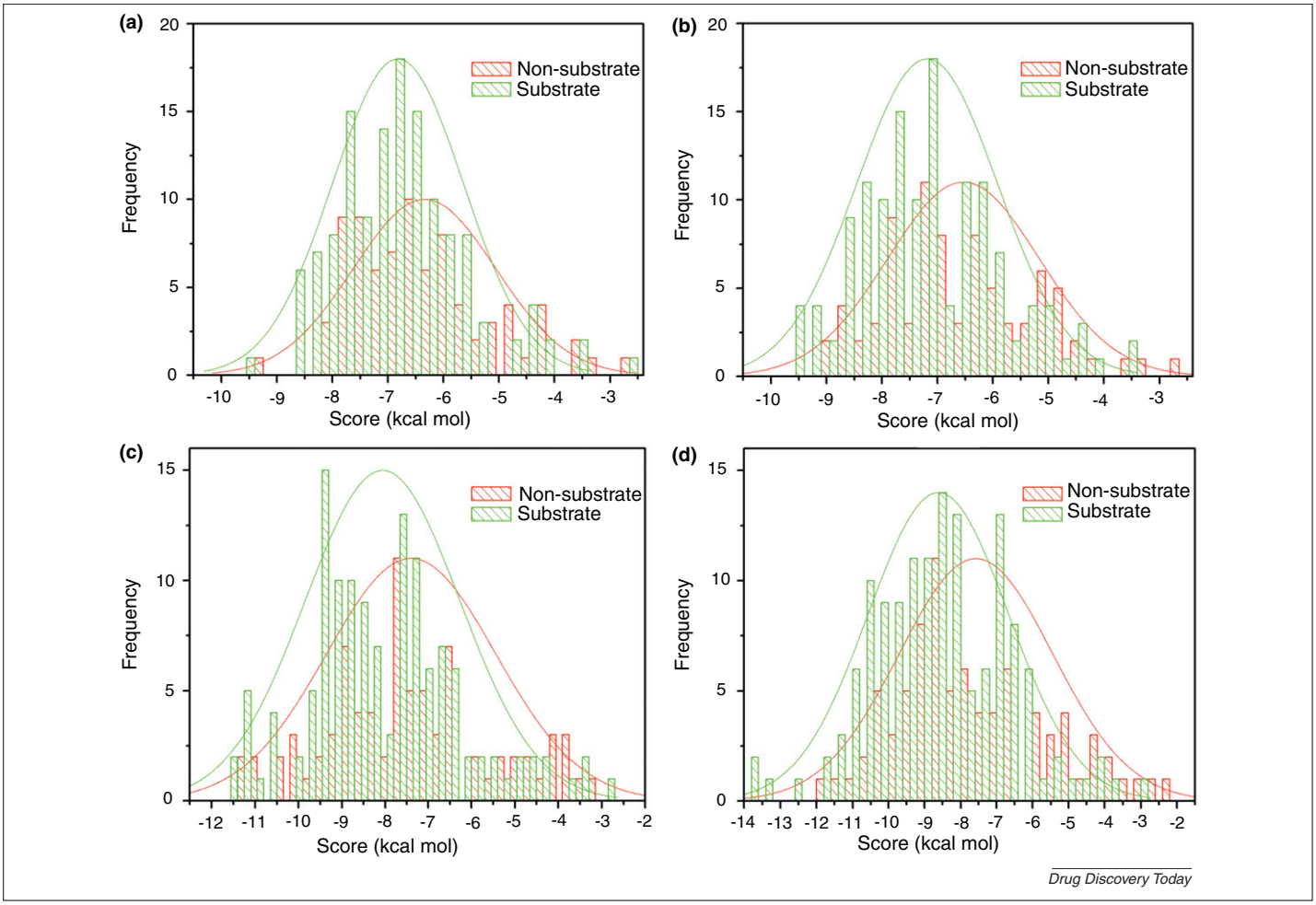

FIGURE 3

The distributions of the docking scores for P-glycoprotein substrates and non-substrates. **(a)** Docking was based on 3G60 and the SP precision mode; **(b)** docking was based on 3G60 and the XP precision mode; **(c)** docking was based on 3G61 and the SP precision mode; **(d)** docking was based on 3G61 and the XP precision mode.

The distributions of the molecular docking scores for the substrates and non-substrates are shown in Fig. 3. Although the mean values of the docking scores for the P-gp substrates are lower than those for the non-substrates (Fig. 3), the distributions of the substrates and the non-substrates still overlap greatly, which obviously shows that, based on the docking scores, the P-gp substrates and the non-substrates cannot be distinguished clearly.

The failure of the molecular docking study to distinguish the substrates from the non-substrates might be explained by the polyspecific nature of substrate binding and only one active binding pocket used in the same docking environment. We believe that it is still difficult to use P-gp homology models or X-ray structures successfully for prospective molecular docking studies. We might be able to apply the molecular docking to generate more-reliable predictions with more co-crystallized ligands solved in the future.

Current challenges and future directions

A lot of effort has been dedicated to predict P-gp inhibitors or substrates and understand the mechanism of action for the P-gp inhibitors or substrates. Currently, only limited *in silico* models can give satisfactory predictions. How to improve the prediction accuracy of the models still remains a significant challenge.

The lack of reliable and extensive experimental data is undoubtedly a major obstacle to developing accurate computational models. We have reported the largest dataset for P-gp inhibitors [21]. The dataset includes 1273 structurally diverse molecules, among which 797 molecules are P-gp inhibitors and 476 molecules are P-gp non-inhibitors. However, for P-gp substrates, large datasets are still needed. Even the largest dataset reported by Wang *et al.* only contains 332 P-gp substrates and non-substrates [41]. Therefore, further development on the availability of P-gp data for the public domain is still necessary.

As mentioned above, the large binding site of P-gp accommodating multiple binding modes and diverse chemical structures means that it is a demanding proposition for modeling. It is really difficult to develop a global model for P-gp inhibitors or substrates that could have different binding mechanisms. It is possible that the combination of two or more models, based on different principles, can give higher confidence for predicting P-gp inhibitors or substrates. Li's work gives us some clues to aid development of integrated models [26]. Li and co-workers developed multiple pharmacophore models for P-gp substrates and integrated them by a classification tree, which achieved high prediction accuracy. In the near future, based on the large datasets, we will be able to develop multiple prediction models with high prediction accuracy and integrate them into a single prediction platform.

The high-resolution structures of P-gp are available now, but there are limited results for rationally translating this information into developing prediction models with satisfactory reliability. Unlike most pharmacological targets, P-gp can recognize a broad variety of compounds with relatively weak binding affinities. The weak and unspecific binding properties of P-gp amplify the inherent defects in molecular docking approaches and limit the use of those protein structures in a broader sense. How to incorporate the structural information and develop the structure-based prediction models for P-gp inhibitors or substrates remains a serious problem.

Acknowledgments

The project is supported by the National Science Foundation of China (Grant No. 20973121 and Grant No. 21173156) and the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD).

References

1 Fromm, M.F. (2000) P-glycoprotein: a defense mechanism limiting oral bioavailability and CNS accumulation of drugs. *Int. J. Clin. Pharmacol. Ther.* 38, 69–74
2 Gottesman, M.M. and Ling, V. (2006) The molecular basis of multidrug resistance in cancer: the early years of P-glycoprotein research. *Febs Lett.* 580, 998–1009
3 Kim, R.B. *et al.* (1998) The drug transporter P-glycoprotein limits oral absorption and brain entry of HIV-1 protease inhibitors. *J. Clin. Investig.* 101, 289–294
4 Leslie, E.M. *et al.* (2005) Multidrug resistance proteins: role of P-glycoprotein, MRP1, MRP2, and BCRP (ABCG2) in tissue defense. *Toxicol. Appl. Pharmacol.* 204, 216–237
5 Marzolini, C. *et al.* (2004) Polymorphisms in human MDR1 (P-glycoprotein): recent advances and clinical relevance. *Clin. Pharmacol. Ther.* 75, 13–33
6 Polli, J.W. *et al.* (2001) Rational use of in vitro P-glycoprotein assays in drug discovery. *Int. J. Clin. Pharmacol. Ther.* 299, 620–628
7 Szakacs, G. *et al.* (2004) The molecular mysteries underlying P-glycoprotein-mediated multidrug resistance. *Cancer Biol. Ther.* 3, 382–384
8 Sharom, F.J. (2008) ABC multidrug transporters: structure, function and role in chemoresistance. *Pharmacogenomics* 9, 105–127
9 Kartner, N. *et al.* (1983) Cell surface P-glycoprotein associated with multidrug resistance in mammalian cell lines. *Science* 221, 1285–1288
10 Szakacs, G. *et al.* (2006) Targeting multidrug resistance in cancer. *Nat. Rev. Drug Discov.* 5, 219–234
11 Ambudkar, S.V. *et al.* (2003) P-glycoprotein: from genomics to mechanism. *Oncogene* 22, 7468–7485
12 Gottesman, M.M. and Pastan, I. (1993) Biochemistry of multidrug resistance mediated by the multidrug transporter. *Annu. Rev. Biochem.* 62, 385–427
13 Fromm, M.F. *et al.* (1999) Inhibition of P-glycoprotein-mediated drug transport – a unifying mechanism to explain the interaction between digoxin and quinidine. *Circulation* 99, 552–557
14 Szakacs, G. *et al.* (2008) The role of ABC transporters in drug absorption, distribution, metabolism, excretion and toxicity (ADME-Tox). *Drug Discov. Today* 13, 379–393
15 Colabufo, N.A. *et al.* (2010) Perspectives of P-glycoprotein modulating agents in oncology and neurodegenerative diseases: pharmaceutical, biological, and diagnostic potentials. *J. Med. Chem.* 53, 1883–1897
16 Colabufo, N.A. *et al.* (2010) Substrates, inhibitors and activators of P-glycoprotein: candidates for radiolabeling and imaging perspectives. *Curr. Top. Med. Chem.* 10, 1703–1714
17 Ecker, G.E. *et al.* (2009) Predicting ligand interactions with ABC transporters in ADME. *Chem. Biodivers.* 6, 1960–1969
18 van de Waterbeemd, H. and Gifford, E. (2003) ADMET *in silico* modelling: towards prediction paradise? *Nat. Rev. Drug Discov.* 2, 192–204
19 Ekins, S. *et al.* (2007) Future directions for drug transporter modelling. *Xenobiotica* 37, 1152–1170
20 Hou, T.J. and Xu, X.J. (2004) Recent development and application of virtual screening in drug discovery: an overview. *Curr. Pharm. Des.* 10, 1011–1033
21 Chen, L. *et al.* (2011) ADME evaluation in drug discovery. 10. Predictions of P-glycoprotein inhibitors using recursive partitioning and naive Bayesian classification techniques. *Mol. Pharm.* 8, 889–900
22 Crivori, P. *et al.* (2006) Computational models for identifying potential P-glycoprotein substrates and inhibitors. *Mol. Pharm.* 3, 33–44
23 Ekins, S. *et al.* (2002) Application of three-dimensional quantitative structure–activity relationships of P-glycoprotein inhibitors and substrates. *Mol. Pharmacol.* 61, 974–981
24 Gombar, V.K. *et al.* (2004) Predicting P-glycoprotein substrates by a quantitative structure–activity relationship model. *J. Pharm. Sci.* 93, 957–968

25 Langer, T. *et al.* (2004) Lead identification for modulators of multidrug resistance based on *in silico* screening with a pharmacophoric feature model. *Arch. Pharm.* 337, 317–327
26 Li, W.X. *et al.* (2007) Significance analysis and multiple pharmacophore models for differentiating P-glycoprotein substrates. *J. Chem. Inf. Model.* 47, 2429–2438
27 Lima, P.D.C. *et al.* (2006) Combinatorial QSAR modeling of P-glycoprotein substrates. *J. Chem. Inf. Model.* 46, 1245–1254
28 Pajeva, I.K. and Wiese, M. (2002) Pharmacophore model of drugs involved in P-glycoprotein multidrug resistance: explanation of structural variety (Hypothesis). *J. Med. Chem.* 45, 5671–5686
29 Penzotti, J.E. *et al.* (2002) A computational ensemble pharmacophore model for identifying substrates of P-glycoprotein. *J. Med. Chem.* 45, 1737–1740
30 Schmid, D. *et al.* (1999) Structure–activity relationship studies of propafenone analogs based on P-glycoprotein ATPase activity measurements. *Biochem. Pharmacol.* 58, 1447–1456
31 Sun, H.M. (2005) A naive Bayes classifier for prediction of multidrug resistance reversal activity on the basis of atom typing. *J. Med. Chem.* 48, 4031–4039
32 Wang, Y.H. *et al.* (2005) An *in silico* approach for screening flavonoids as P-glycoprotein inhibitors based on a Bayesian-regularized neural network. *J. Comput. Aided Mol. Des.* 19, 137–147
33 Wang, Y.H. *et al.* (2005) Classification of substrates and inhibitors of P-glycoprotein using unsupervised machine learning approach. *J. Chem. Inf. Model.* 45, 750–757
34 Wu, J.H. *et al.* (2009) Quantitative structure activity relationship (QSAR) approach to multiple drug resistance (MDR) modulators based on combined hybrid system. *Qsar Combinatorial Sci.* 28, 969–978
35 Xue, Y. *et al.* (2004) Prediction of P-glycoprotein substrates by a support vector machine approach. *J. Chem. Inf. Comput. Sci.* 44, 1497–1505
36 Cabrera, M.A. *et al.* (2006) A topological substructural approach for the prediction of P-glycoprotein substrates. *J. Pharm. Sci.* 95, 589–606
37 Cianchetta, G. *et al.* (2005) A pharmaeophore hypothesis for P-glycoprotein substrate recognition using GRIND-based 3D-QSAR. *J. Med. Chem.* 48, 2927–2935
38 Ekins, S. *et al.* (2002) Three-dimensional quantitative structure–activity relationships of inhibitors of P-glycoprotein. *Mol. Pharm.* 61, 964–973
39 Huang, J.P. *et al.* (2007) Identifying P-glycoprotein substrates using a support vector machine optimized by a particle swarm. *J. Chem. Inf. Model.* 47, 1638–1647
40 Pajeva, I.K. *et al.* (2009) Combined pharmacophore modeling, docking, and 3D QSAR studies of ABCB1 and ABCC1 transporter inhibitors. *Chemmedchem* 4, 1883–1896
41 Wang, Z. *et al.* (2011) P-glycoprotein substrate models using Support Vector Machines based on a comprehensive data set. *J. Chem. Inf. Model.* 51, 1447–1456
42 Ha, S.N. *et al.* (2007) Mini review on molecular modeling of P-glycoprotein (Pgp). *Curr. Top. Med. Chem.* 7, 1525–1529
43 Demel, M.A. *et al.* (2008) *In silico* prediction of substrate properties for ABC-multidrug transporters. *Expert Opin. Drug Metab. Toxicol.* 4, 1167–1180
44 Ecker, G.F. *et al.* (2008) Computational models for prediction of interactions with ABC-transporters. *Drug Discov. Today* 13, 311–317
45 Demel, M.A. *et al.* (2009) Predicting ligand interactions with ABC transporters in ADME. *Chem. Biodivers.* 6, 1960–1969
46 Seeger, M.A. and van Veen, H.W. (2009) Molecular basis of multidrug transport by ABC transporters. *Biochim. Biophys. Acta* 1794, 725–737
47 Stenham, D.R. *et al.* (2003) An atomic detail model for the human ATP binding cassette transporter P-glycoprotein derived from disulfide cross-linking and homology modeling. *FASEB J.* 17, 2287–2289
48 Seigneuret, M. and Garnier-Suillerot, A. (2003) A structural model for the open conformation of the mdr1 P-glycoprotein based on the MsbA crystal structure. *J. Biol. Chem.* 278, 30115–30124
49 Pajeva, I.K. *et al.* (2004) Structure-function relationships of multidrug resistance P-glycoprotein. *J. Med. Chem.* 47, 2523–2533
50 Aller, S.G. *et al.* (2009) Structure of P-glycoprotein reveals a molecular basis for poly-specific drug binding. *Science* 323, 1718–1722
51 Hollenstein, K. *et al.* (2007) Structure and mechanism of ABC transporter proteins. *Curr. Opin. Struct. Biol.* 17, 412–418
52 Locher, K.P. (2009) Structure and mechanism of ATP-binding cassette transporters. *Philos. Trans. R. Soc. Lond. B: Biol. Sci.* 364, 239–245
53 Mourez, M. *et al.* (2000) Role, functional mechanism and structure of ABC (ATP-binding cassette) transporters. *M S-Med. Sci.* 16, 386–394
54 Oldham, M.L. *et al.* (2008) Structural insights into ABC transporter mechanism. *Curr. Opin. Struct. Biol.* 18, 726–733
55 Tombline, G. *et al.* (2005) Involvement of the ''occluded nucleotide conformation'' of P-glycoprotein in the catalytic pathway. *Biochemistry* 44, 12879–12886
56 Sauna, Z.E. and Ambudkar, S.V. (2007) About a switch: how P-glycoprotein (ABCB1) harnesses the energy of ATP binding and hydrolysis to do mechanical work. *Mol. Cancer Ther.* 6, 13–23
57 Sharom, F.J. *et al.* (2005) New insights into the drug binding, transport and lipid flippase activities of the P-glycoprotein multidrug transporter. *J. Bioenerg. Biomembr.* 37, 481–487
58 Martin, C. *et al.* (2000) Drug binding sites on P-glycoprotein are altered by ATP binding prior to nucleotide hydrolysis. *Biochemistry* 39, 11901–11906
59 Zamora, J.M. *et al.* (1988) Physical-chemical properties shared by compounds that modulate multidrug resistance in human leukemic cell. *Mol. Pharmacol.* 33, 454–462
60 Pearce, H.L. *et al.* (1989) Essential features of the P-glycoprotein pharmacophore as defined by a series of reserpine analogs that modulate multidrug resistance. *Proc. Natl. Acad. Sci.* 86, 5128–5132
61 Seelig, A. (1998) A general pattern for substrate recognition by P-glycoprotein. *Eur. J. Biochem.* 251, 252–261
62 Bakken, G.A. and Jurs, P.C. (2000) Classification of multidrug-resistance reversal agents using structure-based descriptors and linear discriminant analysis. *J. Med. Chem.* 43, 4534–4541
63 Ramu, A. and Ramu, N. (1992) Reversal of multidrug resistance by phenothiazines and structurally related compounds. *Cancer Chemother. Pharmacol.* 30, 165–173
64 Ramu, A. and Ramu, N. (1994) Reversal of multidrug resistance by bis(phenylalkyl)amines and structurally related compounds. *Cancer Chemother. Pharmacol.* 34, 423–430
65 Muller, H. *et al.* (2008) Functional assay and structure–activity relationships of new third-generation P-glycoprotein inhibitors. *Bioorg. Med. Chem.* 16, 2448–2462
66 O'Mara, M.L. and Tieleman, D.P. (2007) P-glycoprotein models of the apo and ATP-bound states based on homology with Sav1866 and MalK. *Febs Lett.* 581, 4217–4222
67 Globisch, C. *et al.* (2008) Identification of putative binding sites of P-glycoprotein based on its homology model. *Chemmedchem* 3, 280–295
68 Stockner, T. *et al.* (2009) Data-driven homology modelling of P-glycoprotein in the ATP-bound state indicates flexibility of the transmembrane domains. *Febs J.* 276, 964–972
69 Becker, J.P. *et al.* (2009) Molecular models of human P-glycoprotein in two different catalytic states. *Bmc Struct. Biol.* 9, 3