Research Article

JOURNAL OF CHEMOMETRICS

Received: 8 October 2010, Revised: 25 November 2010, Accepted: 1 December 2010, Published online in Wiley Online Library: 17 February 2011

(wileyonlinelibrary.com) DOI: 10.1002/cem.1376

Development of a highly specific ensemble of topological models for early identification of P-glycoprotein substrates

Mauricio Di Ianni[a], Alan Talevi[a]*, Eduardo A. Castro[b] and Luis E. Bruno-Blanch[a]

P-glycoprotein (Pgp) is an ATP-dependent efflux transporter protein associated with multidrug resistance in several diseases such as cancer, epilepsy and AIDS. It is preferentially expressed in organs and tissues that function as a barrier (e.g. the gut walls or the blood–brain barrier) or promote the elimination of xenobiotics from the organism (e.g. liver and kidney). Pgp limits drug bioavailability; thus, the recognition of Pgp substrates at the early stages of the drug development cycle is essential for the development of new chemotherapeutic agents to deal with multidrug resistance issues. Here we present the development of several classifier models based on topological descriptors to identify potential Pgp substrates, aimed to be applied as secondary filter in virtual screening campaigns. Receiver Operating characteristic (ROC) curves show that combination of individual models, through data fusion, in a three-model ensemble, allows attaining higher areas under the curve and an overall better behavior in terms of sensitivity or specificity. The individual discriminant functions (dfs) presented have a performance similar to that of the previously reported models and, remarkably, our models only include low-dimensional (up to 2D) molecular descriptors, which makes them adequate for the virtual screening of increasingly large virtual chemical repositories.

Supporting information may be found in the online version of this article.

Keywords: Pgp; topological models; data fusion

1. INTRODUCTION

The old drug discovery and development paradigm focused on introducing structural modifications to a lead compound in order to improve the potency of the drug (i.e. its ability to interact with the molecular target), with little attention being paid to other important aspects of a drug such as bioavailability or toxicity. As a consequence of this scheme, around 40% of the drug development projects failed because of inappropriate drug disposition characteristics (inability of the drug to reach its biological molecular target in adequate amounts) and about 20% ended up in failure due to toxicity issues [1]. Recent data suggest, however, that at present the attrition rate related to poor pharmacokinetics has been considerably reduced, and this can be ascribed to integration of *in vitro* and *in silico* Absorption, Distribution, Metabolism and Elimination (ADME) filters at early stages of the drug development process [2,3].

Among the ADME-related drug properties that raise more interest is drug affinity to P-glycoprotein (Pgp, also known as MDR1 protein or ABCB1 protein). Pgp is an ATP-dependent transporter that functions as a transmembrane efflux pump that translocates its substrates from its intracellular domain to its extracellular domain [4]. It is located in many organs and tissues that protect the organism from potentially toxic xenobiotics (a chemical that is found in a given organism but which is not normally produced or expected to be present in it), such as the epithelial cells of the gastrointestinal tract, the canalicular membrane of hepatocytes, the luminal membrane of the proximal tubule cells in the kidneys, the blood–brain barrier and others. When expressed in a normal tissue, Pgp acts in three main ways: it limits drug entry into the body after oral administration; it promotes drug elimination into urine and bile and; once a drug has reached general circulation, it limits the amount of drug that reaches certain sensitive tissues and organs (e.g. the brain and the testis). Pgp is also expressed in tumor cells, and similar transporters have been identified in infectious agents. A central characteristic of Pgp is its broad substrate specificity (i.e. substrate promiscuity, the ability to bind and transport a wide range of structurally unrelated chemicals): over-expression of Pgp thus determines multidrug resistance issues (decreased sensitivity to a wide range of structurally and/or functionally unrelated chemotherapeutic agents) [5]. Pgp over-expression

* *Correspondence to: A. Talevi, Medicinal Chemistry, Department of Biological Sciences, Faculty of Exact Sciences, University of La Plata (UNLP)—CCT La Plata CONICET, 47 & 115, La Plata (B1900AJI), Buenos Aires, Argentina. E-mail: atalevi@biol.unlp.edu.ar*

a *M. Di Ianni, A. Talevi, L. E. Bruno-Blanch*
Medicinal Chemistry, Department of Biological Sciences, Faculty of Exact Sciences, University of La Plata (UNLP)—CCT La Plata CONICET, 47 & 115, La Plata (B1900AJI), Buenos Aires, Argentina

b *E. A. Castro*
Instituto de Investigaciones Fisicoquímicas Teóricas y Aplicadas (INIFTA), Departament of Chemistry, Faculty of Exact Sciences, University of La Plata (UNLP)—CCT La Plata CONICET, La Plata, Buenos Aires, Argentina

may be intrinsic to a given patient (e.g. due to polymorphisms in the MDR1 gene) [6] or it may be triggered by disease or by drugs themselves [7,8]. Pgp over-expression is clinically relevant: it is associated with multidrug resistance throughout a very broad array of health conditions such as cancer [9], AIDS [10] and many central nervous system diseases [11], among others. Two strategies are thus being explored by medicinal chemists and pharmacologists to cope with Pgp-mediated multidrug resistance: (a) developing Pgp inhibitors and their co-administration together with already known drugs that are recognized as Pgp substrates and (b) identifying potential Pgp recognition at the beginning of the drug development cycle to select or design drug candidates that are less likely to be transported by Pgp [12].

Most of the recently reported computational models to recognize Pgp substrates or inhibitors are either phamacophore hypothesis or QSAR models that rely on 3D molecular descriptors [12–23], with some exceptions such as the work by Huang *et al*. [24] and Cabrera *et al*. [25], who reported models that include either constitutional or topological (0D–2D) descriptors. Since currently public chemical repositories, such as Pubchem or ZINC databases, hold several millions of drug-like compounds and the small organic compound chemical space grows exponentially toward the hypothetical estimate of 10^{60} to 10^{100} feasible chemical entities [26,27], 3D models might not be the most efficient choice for exploration of this vast universe, since they require previous conformational analysis to obtain a probable conformer or an ensemble of probable conformers of the repository compounds. Alternatively, low-dimensional, conformation-independent descriptors (0D–2D) do not require any pre-processing of the molecular structures and their computation is extremely inexpensive in terms of computational time (a more detailed discussion on the advantages and disadvantages of 2D and 3D approaches can be found elsewhere [28]). Structure-based approaches to develop new drugs to circumvent Pgp are yet to be explored, due to the unavailability of a high-resolution structure of this efflux transporter. Because of the intrinsic difficulties in crystallizing transmembrane proteins [29], it was not until lately that a 3.8 Å crystal structure was described [30].

Here we report the development of an ensemble of QSAR topological models to be applied in virtual screening campaigns in order to discard potential Pgp substrate among drug candidates. This ensemble of models is aimed to efficiently explore large virtual repositories of chemical compounds to select chemical entities capable of dealing with Pgp-related multidrug resistance issues, and it has been conceived as an ADME filter to be used in virtual screening campaigns. The use of an ensemble of models instead of a single model is proposed here as a strategy to increase the specificity of the model and to consider the fact that multiple binding sites have been described in Pgp, so it would be difficult to identify Pgp substrates that are recognized by different sites of the protein with a single model. A detailed discussion on these subjects is presented in Sections 2 and 4.

2. METHODS

2.1. Dataset

A 250-compound diverse dataset containing 104 substrates and 146 non-substrates was extracted from the literature [12,16,18,31]. The dataset was split into a 125-compound training set and a 125-compound test set by systematic random sampling (the compounds of each category were sorted alphabetically and one in two compounds was kept for the test set; no periodic pattern exists between the name of the compounds used in this study and their structure or pharmacologic activity). We decided for an even partition of the dataset into training and test sets on the basis of a recent report that suggests that external validation results may be more reliable when even, random partitions of medium-size datasets are considered [32]. The name and structures of the drugs that compose the training and the test sets are presented as supplementary information so that the reader may examine its chemical diversity.

2.2. Molecular descriptor calculation and modeling technique

Dragon software for molecular descriptors calculation, version 4.0 (Milano Chemometrics, 2003) was used for the calculation of 867 low-dimensional (0D–2D) descriptors, distributed along 12 blocks of descriptors, e.g. constitutional descriptors, topological descriptors, connectivity indices, Galvez topological charge indices and others. Since such a high number of descriptors may result in chance correlations between the modeled property and a subset of descriptors, 30 subsets of descriptors obtained from random combinations of the blocks of Dragon low-dimensional descriptors were considered as independent pools of descriptors, each combination containing around 200 molecular descriptors. For example, the first pool of descriptors (207 descriptors) emerged from the combination of the following Dragon blocks of descriptors: edge adjacency indices, topological charge indices, eigenvalue-based indices and functional group counts; the second pool of descriptors (160 descriptors) combined walk and path counts, 2D autocorrelations and topological charge indices, and so on. We will refer to these random combinations of descriptors blocks as 'random pools' from now on.

A 31st pool of descriptors that we will call 'rational pool' was also considered by reviewing past reports of models related to Pgp affinity and designing a subset with those Dragon descriptors possibly related to key features identified in previous modeling efforts [12–24,33–35]. For example, several of the previous studies on Pgp affinity models reported that the numbers of H-bond donors and acceptors are important features for the recognition event; therefore, the number of H-bond donors, the number of H-bond acceptors, the number of primary, secondary and tertiary N atoms and the number of OHs were included in the rational pool. Some previous models suggested that the size of a molecule may be important for recognition: therefore, molecular weight and molar refractivity were included in the rational pool, since they are clearly correlated with molecular size and molecular volume. Several 2D autocorrelations possibly related to pharmacophore features of reported pharmacophores were also considered. In this way, we arrived at a 90-descriptor rational pool. Descriptors with constant or near-constant values for the training set, associated with low information content, were removed from random and rational descriptors pools.

A binary, dummy variable codifying the category of each compound was used as a dependent variable (class = 1 for substrates and class = −1 for non-substrates). Stepwise forward multiple linear regression was used to select the descriptors from

each random pool that best discriminated the category of the compounds; linear discriminant analysis (LDA) was used to characterize the corresponding linear discriminant functions (dfs). Dfs assume the following general form:

$$\text{df value} = a_0 + \sum_i ai - di$$

where a_0 is the constant and ai is the coefficient associated with molecular descriptor di. Due to the values arbitrarily assigned to substrates and non-substrates, substrates will tend to have positive df values, and non-substrates will tend to assume negative values.

The binary classification scheme reduces the error associated with the process of combining data obtained in different labs and conditions [12]. Multiple regression and discriminant analysis modules from Statistica version 5.1 (Statsoft Ltd., 1996) were used for modeling purposes. Tolerance values no lower than 0.2 were used in order to avoid inclusion of highly correlated pairs of descriptors. The minimum cases to predictors ratio allowed was 10 (10 or more cases in the training set for each descriptor included in the model) in order to reduce chances of over-fitting. Only descriptors with significant coefficients at an α level of 0.05 are allowed into the model. Randomization, leave-group-out (LGO) cross-validation and external validation (predicting the class for the independent 125-compound test set) were used to assess robustness and predictive ability of all models. Forty randomized models were built in the randomization test. In each LGO row, 12 compounds were randomly removed from the training set to the test set and the LGO models were used to assess the category of the removed compounds; this process was repeated until all the compounds in the training set had been removed at least once.

2.3. Combining models

Two important indicators of the performance of a given QSAR model are sensitivity (Se) and specificity (Sp). They are defined by the following expressions:

$$\text{Se} = \frac{\text{TP}}{\text{TP}+\text{FN}}$$
$$\text{Sp} = \frac{\text{TN}}{\text{TN}+\text{FP}}$$

where TP refers to true positives, FN refers to false negatives, TN refers to true negatives and FP refers to false positives. Here, since we are looking for compounds that are not transported by Pgp (Pgp non-substrates) and we want to discard Pgp substrates, the previous expressions may be re-written as follows:

$$\text{Se} = \frac{\text{True non-substrates}}{(\text{True non-substrates} + \text{False substrates})}$$
$$\text{Sp} = \frac{\text{True substrates}}{(\text{True substrates} + \text{False non-substrates})}$$

By modifying the selection threshold from the lowest to the highest score provided by the model, Se and Sp will evolve in opposite ways; consequently, it is not possible to optimize both parameters simultaneously, and a tradeoff has to be found [36]. ROC curves are a widely used tool to assess and compare the performance of different models [36]. They are graphical plots of the sensitivity (true positive rate) versus 1 minus specificity (i.e. 1 less the false positive rate), for a binary classifier system, as its discrimination cutoff value changes. ROC curves provide a rational and user-friendly basis to balance type I and type II errors, selecting optimal models and optimal cutoff values. The area under the ROC curve can be used for general comparison purposes of different models or methodologies: an ideal model will present an area under the ROC curve of 1 (equivalent to perfect classification, i.e. a sensitivity of 1 and a specificity of 1 for a given cutoff value) while random classification is represented by a line of slope 1 and corresponds to an area under the ROC curve of 0.5. Here, we have built ROC curves to compare the performance of the individual models developed and the performance of a three-model ensemble obtained through different, simple data fusion schemes.

It has been pointed out that there is no general rule for balancing errors [36,37]: balancing FP and FN depends on pragmatic considerations that are to be judged by the researcher [38]. We are interested in adopting a conservative attitude and developing highly specific models, i.e. models capable of discarding practically all Pgp substrates. This is strongly related to our background: a small academic research group from a developing country with limited resources to invest in drug acquisition and pharmacologic testing. Therefore, we will prefer Sp over Se: at the risk of losing some valuable scaffolds when applying our topological Pgp affinity models in virtual screening campaigns, we will choose to avoid acquiring or synthesizing a drug candidate that, once sent to pharmacological testing, will prove to be an FP (a drug that was predicted as a non-substrate of Pgp but which is actually transported by Pgp). Therefore, in the light of the fact that most of our models perform better in the classification of non-substrates (which indicates a higher rate of FP, wrongly classified substrates, than FN, wrongly classified non-substrates, see Section 3), and taking into account that Pgp is characterized by broad substrate specificity (probably because of the existence of multiple binding sites in the protein [19,39]), we have chosen to look for combinations of the topological dfs that provide the lowest rate of FP in the external validation. Substrate promiscuity indicates that it might be difficult to obtain a single model capable of identifying the entire set of Pgp substrates. We have combined the models by the very simple strategy of looking for all the possible two-model combinations of the models built from the random pools of descriptors, and then joining the best two-model combination with the models obtained from the rational pool, to give a three-model ensemble. On the basis of our aforementioned decision to prefer highly specific models, we used the FP rate on the 125-compound test set to define the best combinations of models, using zero as a cutoff value and selecting, among those combinations with lowest FP rate, the combination with lowest FN rate. As a first approach, the maximum (MAX operator) value among the three values provided for each compound by the three independent classifiers that compose the three-model ensemble was used to classify each of the test-set compounds as Positive (Non-substrate) or Negative (Substrate). Later, two additional data fusion schemes were also explored: (a) the sum of the three values provided for each compound by the three independent models that compose the ensemble (SUM) and (b) the average of the three values provided for each compound by the three independent models that compose the ensemble (AVE). The strategy of combining different models has been successfully used previously to improve classification [14,17]. Table I summarizes the different stages applied to obtain the best three-model ensemble.

Table I. Summary of the different stages of the combination scheme used to obtain the three-model ensemble

Stage	Description	Criteria used
1	Generation of models from the random pools of descriptors	30 random subsets of descriptors are derived from a pool of 867 molecular descriptors from Dragon software Linear discriminant analysis is applied, using a tolerance no smaller than 0.2 to avoid inclusion of redundant descriptors No model with less than 75% of overall good classifications on the training set is kept No model with cases to predictor ratio below 10 is kept, to avoid over-fitting Cross-validation and randomization tests are performed External validation is performed through a 125-compound test set
2	Generation of models from the rational pools of descriptors	A 31st pool of descriptors is used to derive more models. This pool is based on previously reported models to identify Pgp substrates. Criteria identical to those used with the random pools are applied
3	Analysis of all possible two-model combinations of the models derived from the random pool of molecular descriptors	To increase Sp at the expense of Se, the combination of models providing smallest FN rate among those combinations with smallest FP rate is selected The MAX value among the two df values provided by the two independent classifiers composing each combination is used to classify each of the 125 test set compounds Whenever the MAX df value is positive, a test set compound is considered a Pgp substrate
4	Generation of the three-model ensemble	The best two-model combination from stage 1 is combined with each of the models derived from the rational pool of molecular descriptors Again, we look for the three-model combination with lowest FP rate. Between two combinations with similar FP rate, we prefer that with lower FN rate The MAX value among the three df values provided by the three independent classifiers composing each combination is used to classify each of the 125 test set compounds Whenever the MAX df value is positive, a test set compound is considered a Pgp substrate
5	Exploration of alternative data fusion schemes	The sum and the average of values from the three independent models are proposed as alternative data fusion schemes to combine the information from the three independent models
6	Selection of the threshold value	ROC curves are built for the three-model ensemble and each data fusion scheme. The curves are used to select the cutoff value to differentiate substrates and non-substrates, preferring Sp over Se; the ROC curves from the ensemble are also compared to the ROC curves of individual models to check the success of the model combination strategy

3. RESULTS

3.1. Models built from the random pools

Among the 30 models obtained from the 30 random pools of descriptors, only one presented an acceptable explanatory power (overall classification of 75% on training set compounds). Details of the model are provided below these lines.

$$\begin{aligned} \text{Df value} &= -4.45 + 0.63 * \text{IC3} - 0.94 * \text{nCN} \\ &\quad + 1.34 * \text{GATS4v} - 0.01\,\text{T(N..Cl)} \qquad (1) \\ F &= 12.13 \quad p < 0.0000 \quad \text{Wilk's } \lambda = 0.71 \end{aligned}$$

Tolerance = 0.5; overall good classifications training set = 78.4%; overall good classifications test set = 77.6%; average overall good classifications test set LGO = 76.8%; average overall good classification in randomized models = 60.8%; cases to descriptors ratio = 31.25. We have kept Dragon's nomenclature for the descriptors. IC3 represents the information content index considering third-order neighborhoods; nCN represents the CN count in the molecule; GATS4v symbolizes Geary autocorrelation of lag 4 weighted by atomic Van der Waals volumes and T(N..Cl) represents the topological distance between nitrogen and chlorine atoms in the molecule.

The performance of this model is quite similar to that of previously developed models [12–23,25] and approaches the upper bound of 85% of good classifications in Pgp models estimated by Zhang *et al.* on the basis of the variability of experimental data on Pgp-related assays [16] (only the work from Huang *et al.* reports a performance of 90%, above the calculated upper bound [24]). The percentage of well-classified training set substrates is 73.1%, while for the non-substrates the percentage rises to 82.2%. The percentage of well-classified test set substrates is 69.2%; for the non-substrates in the test set, the

percentage is 83.6%. Note that the percentages are quite similar in the training and the test sets. The performance of randomization models, as expected, approaches random performance (in fact, a deeper analysis of the randomization models reveals that they tend to classify most of the compounds as non-substrates: the average percentage of well-classified compounds is 89.5% for the non-substrates and only 20.3% for the substrates. In other words, randomization models are highly biased toward predicting non-substrates, and therefore they have almost no classificatory power at all). The difference of classification ability between substrates and non-substrates in both the training and the test sets (the non-substrates' classification is clearly better) has also been observed in other modeling efforts [12,17] and may be a consequence of the wide substrate specificity of the molecular target. It may also be reflecting the uneven distribution of substrates and non-substrates in the training set (non-substrates are more frequent than substrates in the training set). Note that the constant in the model (−4.45) is negative and quite far from zero, which suggests a possible bias toward predicting non-substrates. To investigate whether these observations arise from the intrinsic nature of the problem at hand (Pgp promiscuity) or from a problem in the dataset (over-representation of non-substrates), we have repeated the modeling process on the random pools of descriptors, but for this second modeling round be have set the constant to 0 in the multiple regression and setting the *a priori* classification probability as 'the same for all groups' in the Discriminant Analysis module (in fact, we cannot know for sure whether the distribution of the classes in the datasets mimics distribution in nature). This time we obtained four models with acceptable performance explanatory power:

$$\begin{aligned}\text{Df value} = {} & -0.40 * \text{nCl} - 1.48 * \text{nTB} + 2.70 * \text{MATS3p} \\ & + 0.067 * \text{GGI2} \\ F = 12.40 \quad & p < 0.0000 \quad \text{Wilk's } \lambda = 0.71\end{aligned} \tag{2}$$

Tolerance = 0.35; overall good classifications training set = 76.8%; overall good classifications test set = 63.2%; average overall good classifications test set LGO = 63.9%; average overall good classifications in randomized models = 52.0%; cases to descriptors ratio = 31.25. nCl represents the number of chlorine atoms; nTB represents the number of triple bonds; MATS3p is the Moran autocorrelation of lag 3, weighted by atomic polarizabilities and GGI2 is the Galvez' topological charge index of second order. The percentage of well-classified training set substrates by model (2) is 71.2%, while for the non-substrates the percentage rises to 80.8%. The percentage of well-classified test set substrates is 67.3%; for the non-substrates in the test set, the percentage is 60.3%.

$$\begin{aligned}\text{Df value} = {} & 0.41081 * \text{EEig14d} - 1.38061 * \text{nCN} \\ & - 1.26139 * \text{nSO2N} - 0.01946 * \\ & \text{T(N..Cl)} - 0.03033 * \text{T(O..Br)} - 0.00000188 * \text{VRA1} \\ F = 9.69 \quad & p < 0.0000 \quad \text{Wilk's } \lambda = 0.61\end{aligned} \tag{3}$$

Tolerance = 0.5; overall good classifications training set = 76.0%; overall good classifications test set = 79.2%; average overall good classifications test set LGO = 77.1%; average overall good classifications in randomized models = 54.7%; cases to descriptors ratio = 20.8. EEig14d represents the Eigenvalue 14 from edge adjacency matrix weighted by dipole moments, nCN is the number of ciano groups, nSO2N is the number of sulphonamides, T(N..Cl) represents the sum of topological distances between N..Cl, T(O..Br) is the sum of topological distances between O and Br and VRA1 corresponds to Randic-type eigenvectors based index from adjacency matrix. For model (3), the percentage of well-classified training set substrates is 65.4%, while for the non-substrates the percentage rises to 83.6%. The percentage of well-classified test set substrates is 76.1%; for the non-substrates in the test set, the percentage is 83.6%.

$$\begin{aligned}\text{Df value} = {} & -0.64 * \text{Yindex} + 0.11\,\text{GGI2} - 0.75 * \text{S-108} \\ & - 1.28 * \text{nSO2N} - 0.45 * \\ & \text{Cl-089} + 0.54 * \text{N-069} + 0.56 * \text{C-033} + 0.37 * \text{nRSR} \\ F = 9.23 \quad & p < 0.0000 \quad \text{Wilk's } \lambda = 0.67\end{aligned} \tag{4}$$

Tolerance = 0.35; overall good classifications training set = 80.0%; overall good classifications test set = 75.2%; average overall good classifications test set LGO = 75.3%; average overall good classifications in randomized models = 58.3%; cases to descriptors ratio = 15.6. Y index stands for Balaban Y index; S-108 represents the number of R = S; Cl-089 represents the number of Cl attached to sp2 carbons; N-069 represents a primary aromatic amine or a primary amine bonded to a halogen atom; C-033 symbolizes X–CH...X fragments (X being halogen) and nRSR represents the number of sulfurs. The percentage of well-classified training set substrates for model (4) is 73.1%, while for the non-substrates the percentage is 84.9%. The percentage of well-classified test set substrates is 65.4%; for the non-substrates in the test set, the percentage is 82.2%.

$$\begin{aligned}\text{Df value} = {} & -0.55 * \text{Yindex} + 0.17\,\text{GGI2} - 0.61 * \text{S-108} \\ & - 1.23 * \text{nSO2N} - 0.38 * \text{Cl-089} \\ & + 0.55 * \text{N-069} + 0.59 * \text{C-033} + 0.51 * \text{nRSR} \\ & - 27.70 * \text{JGI8} - 0.34 * \text{nCONR2} \\ F = 8.21 \quad & p < 0.0000 \quad \text{Wilk's } \lambda = 0.58\end{aligned} \tag{5}$$

Tolerance = 0.20; overall good classifications training set = 79.2%; overall good classifications test set = 72.8%; average overall good classifications test set LGO = 76.1%; average overall good classifications in randomized models = 60.9%; cases to descriptors ratio = 12.5. JGI8 stands for Galvez mean topological charge index of eighth order and nCONR2 is the number of tertiary (aliphatic) amines. The percentage of well-classified training set non-substrates for model (5) is 82.2%, while for the substrates the percentage is 75.0%. The percentage of well-classified test set substrates is 80.8%; for the non-substrates in the test set, the percentage is 67.1%.

Note that in one of these later four models (model (2)), the performance of the classification of the test set substrates is better than that of the non-substrates, which suggests that the 'set on zero' strategy has been successful to avoid bias toward good classification of non-substrates.

3.2. Models built from the rational pool

In this case, since only one rational pool was designed, after the first model was obtained through the stepwise forward procedure, we removed from the pool the molecular descriptor that entered the model in the first step-forward step, and

repeated the stepwise forward procedure. We systematically repeated this procedure (removal of the first descriptor added to the model from the rational pool of descriptors and a new, subsequent stepwise forward round) until no significant correlation between the dependent variable and the remaining descriptors of the pool was found. We obtained 28 models, of which two seemed relevant:

$$\begin{aligned} \text{Df value} &= 14.1616 + 0.0029 * \text{ATS8e} - 0.9421 * \text{GATS3e} \\ &\quad + 0.3471 * \text{nCaH} - 14.621 * \text{MATS2m} \\ &\quad + 0.2427 * \text{nCO} - 1.3575 * \text{MATS6p} \\ F &= 7.23 \quad p < 0.0000 \quad \text{Wilk's } \lambda = 0.73 \end{aligned} \tag{6}$$

Tolerance = 0.50; overall good classifications training set = 77.6%; overall good classifications test set = 73.6%; average overall good classifications test set LGO = 76.9%; average overall good classifications in randomized models = 55.4%; cases to descriptors ratio = 20.8. ATS8e corresponds to Broto–Moreau autocorrelation of a topological structure of lag 8 weighted by atomic Sanderson electronegativities, GATS3e corresponds to Geary autocorrelation of a topological structure of lag 3 weighted by atomic Sanderson electronegativities, nCaH stands for the number of unsubstituted aromatic Cs, MATS2m represents the Moran autocorrelation of lag 2 weighted by atomic masses, nCO represents the number of ketones and MATS6p represents the Moran autocorrelation of lag 6 weighted by atomic polarizabilities.

This model correctly classifies 67.3 and 84.9% of the training set substrates and non-substrates, in that order, and 61.5 and 82.2% of the test set substrates and non-substrates, respectively.

$$\begin{aligned} \text{Dfvalue} &= 20.61 + 0.12\,\text{nHAcc} - 22.95 * \text{MATS2m} \\ &\quad + 0.042\,\text{mlogP}^2 + 0.38 * \text{nCaH} + 0.22 * \text{nCO} \\ F &= 7.77 \quad p < 0.0000 \quad \text{Wilk's } \lambda = 0.75 \end{aligned} \tag{7}$$

Tolerance = 0.50; overall good classifications training set = 79.2%; overall good classifications test set = 72.8%; average overall good classifications test set LGO = 76.1%; average overall good classifications in randomized models = 60.9%; cases to descriptors ratio = 12.5. nHAcc corresponds to the number of H-bond acceptors; mlog P^2 is the square of Moriguchi's logarithm of the octanol–water partition coefficient. This model correctly classifies 61.5 and 91.8% of the training set substrates and non-substrates, in that order, and 65.4 and 78.1% of the test set substrates and non-substrates, respectively. Since the intercepts here are positive and quite large, we assumed no improvement may be observed by forcing the intercept to take zero value.

3.3. Models combination

As expected, combination of different models according to the scheme presented in Section 2 resulted in a significant increment of Sp (less FP) and a concomitant reduction of Se (increase of FN). The optimal combination of models (the one that raises the most the good classification in the substrate category of the test set—which is our priority—and drops the least the good classification of the non-substrates) was the one combining models (2), (4) and (7), which resulted in 90.4% of good classifications among the test set substrates and 52.1% of good classifications among the test set non-substrates when zero was used as a cutoff value and the MAX data fusion scheme was considered. A brief summary of the performance and features of the selected individual models plus the performance of the best ensemble is presented in Table II. Figures 1 and 2 present the distribution of values that the dfs (1)–(7) and the three-model ensemble assume for both the training and the test sets. The three data fusion schemes (MAX, SUM and AVE) explored for the ensemble are presented. Figure 3 presents the ROC curves of the individual models and the ensemble (considering the three different data fusion schemes), for both the training and the test sets.

4. DISCUSSION

Results indicate that our individual models present similar performance compared to previously reported models developed to recognize Pgp substrates. The overall classification of the individual models, when zero value is used as a threshold to differentiate substrates from non-substrates, is around 80%. As can be appreciated in Table II and in Figures 1 and 2, most of the

Table II. Summary of the results and features of the seven selected models, and the most specific ensemble of topological models

Model	Overall % of good classified training set	Overall % of good classified test set	Well-classified substrates in the training set	Well-classified non-substrates in the training set	Well-classified substrates in the test set	Well-classified non-substrates in the test set	Tolerance	Cases to predictors ratio
1	78.4	77.6	76.1	82.2	69.2	83.6	0.5	31.25
2	76.8	63.2	71.2	80.8	67.3	60.3	0.35	31.25
3	76.0	79.2	65.3	83.6	76.1	83.6	0.5	20.8
4	80.0	75.2	73.1	84.9	65.4	82.2	0.35	15.6
5	79.2	72.8	75.0	82.2	67.1	80.8	0.2	12.5
6	77.6	73.6	67.3	84.9	61.5	82.2	0.5	20.8
7	79.2	72.8	65.1	91.8	65.4	78.1	0.5	12.5
2 + 4 + 7 (ensemble, MAX, cutoff = 0)	71.2	68.8	92.3	56.2	90.4	52.1	—	—

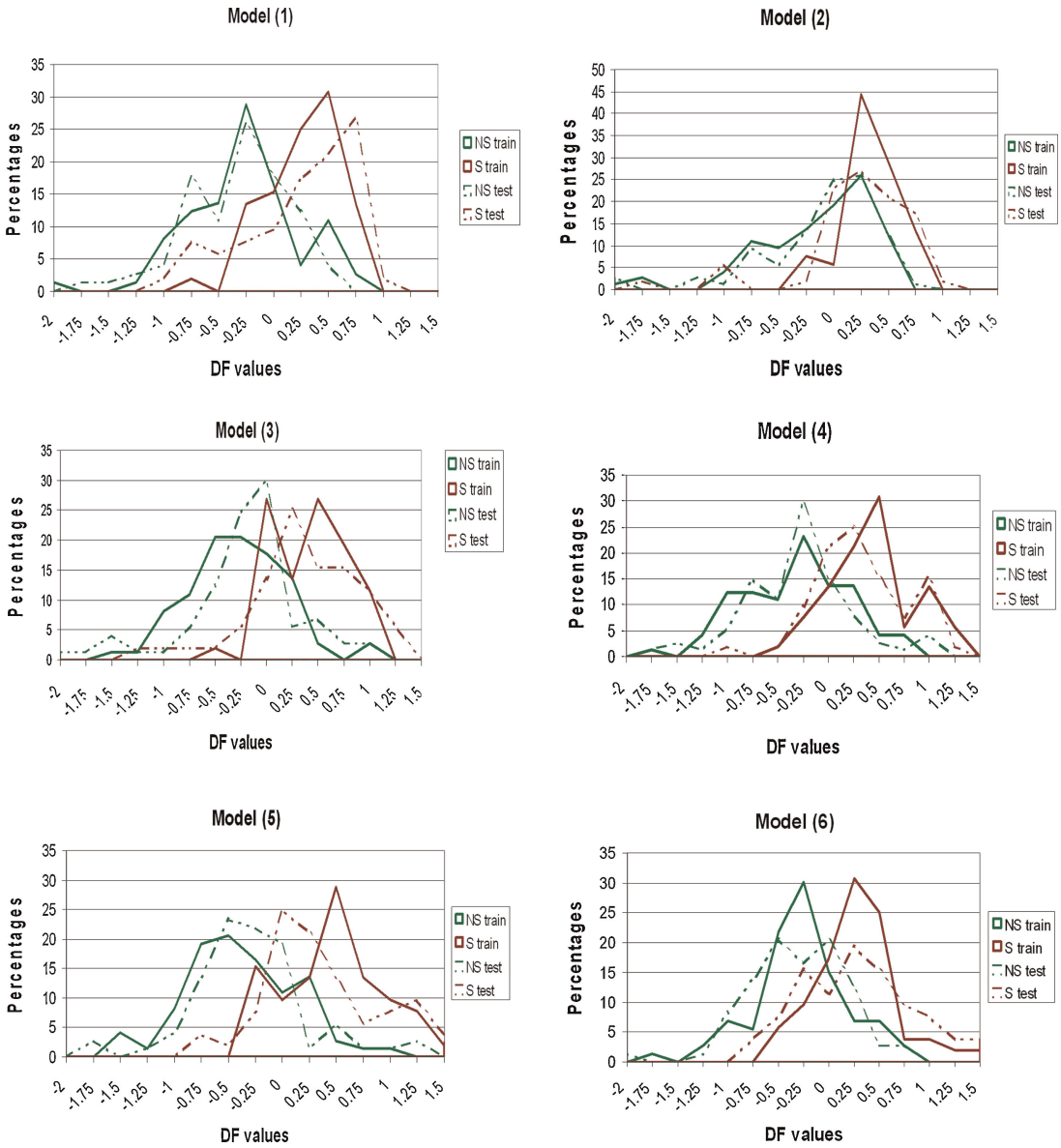

Figure 1. Distribution of values that the discriminant functions (1)–(6) assume for the training and test set compounds (continuous and discontinuous lines, in that order). NS train stands for training set Pgp non-substrates; S train represents training set Pgp substrates; NS test denotes test set non-substrates and S test represents test set substrates.

models showed a similar performance on the training and test sets, i.e. explanatory power of the models is similar to the predictive power, or, in other words, no over-fitting was observed. Note that the plots corresponding to the distribution diagrams of the training and test sets superpose fairly well for all the models, and in particular for the three-model ensemble using AVE data fusion approach. Average performance of the LGO cross-validation models was also similar to the performance on the 125-compound test set used for external validation purposes. Moreover, the average performance of the models obtained through the randomization of the dependent variable always drops considerably compared to that of the actual model (what is more, the performance of the randomized models in some cases is artificially increased due to the fact that many randomized models are biased toward predicting non-substrates, non-substrates being over-represented in the dataset). The strategy of forcing the intercept to take zero value in the stepwise procedure seemed to be efficient when building models from the random pools of descriptors: more relevant models were detected and, in one of them, the classification of the test set substrates outperformed the classification of the test set non-substrates. Some of the bias of the models toward predicting non-substrates may then be related to the uneven distribution of the substrates and non-substrates in the dataset, while substrate promiscuity may explain the difference in classification success across the two considered categories. All the models present good tolerance values, with six out of the seven presented models with tolerance above 0.35, which indicates very low paircorrelation among the included descriptors. The performance of the models is quite near the upper bound calculated by Zhang (85% of accuracy) from the evidence of high variability in Pgp affinity experimental data. Only the Support Vector Machine model from Huang seems to overcome the limit calculated by Zhang. Only one of the models (model (2)) presents significant

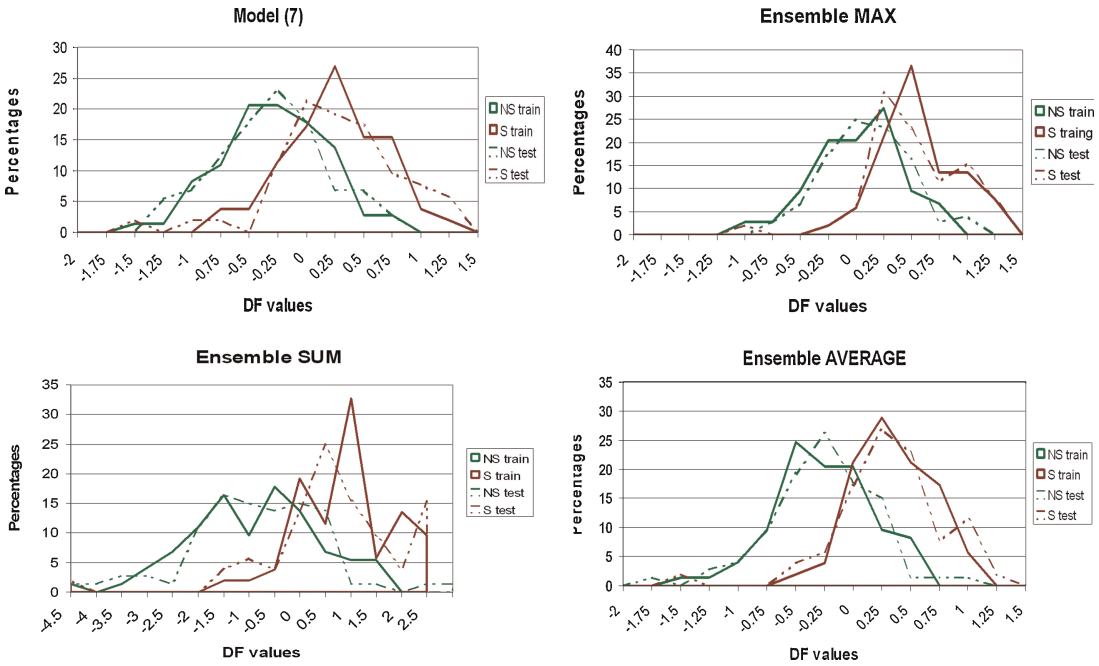

Figure 2. Distribution of values that the discriminant function (7) and the three-model ensemble assume for the training and test set compounds. The three data fusion schemes are considered for the ensemble.

differences between the performance on the training and test sets, indicating a considerable degree of over-fitting. Nevertheless, model (2) appears in several of the ensembles of models that showed best performance, and in the best combination, suggesting that the set of substrates wellclassified by model (2) is complementary to that wellclassified by other models.

The combination of three models resulted in an improved classification of the test set substrates, increasing Sp at the expense of Se when zero was used as a cutoff value to discriminate substrates from non-substrates (note that the balance between Sp and Se can be chosen from observations of ROC curves, i.e. the cutoff value may be rationally optimized from ROC curves on the basis of a particular user's needs, which are background-dependent). The best combination includes one of the models derived from the rational pool of descriptors, which was designed from analysis of previous reports on *in silico* models to recognize Pgp substrates.

Several conclusions may be drawn from the observations of ROC curves and visual comparison of areas under the curves. The AVE and SUM data fusion schemes performed much better than the MAX scheme. The ROC curves built from the performance of the individual models and the ensemble on the training set indicate that the AVE and the SUM ensembles performed quite better than most of the individual models (models (1), (2), (4), (6) and (7)) while models (3) and (5) performed slightly better than the ensemble no matter what threshold value was adopted. Nevertheless, when analyzing the ROC curves built from the test set distribution diagram (which are, in fact, the curves that help us to estimate the predictive capability, i.e. the classificatory power on an independent set of compounds), it is clear that the ensembles outperform the individual models. For example, performances of models (4), (5) and (6) are quite similar to the performance of the AVE and SUM ensembles in the high-specificity region of the curve, but the performance of those individual models is quite lower than that of the ensembles in the high-sensitivity region. In contrast, model (1) outperforms the

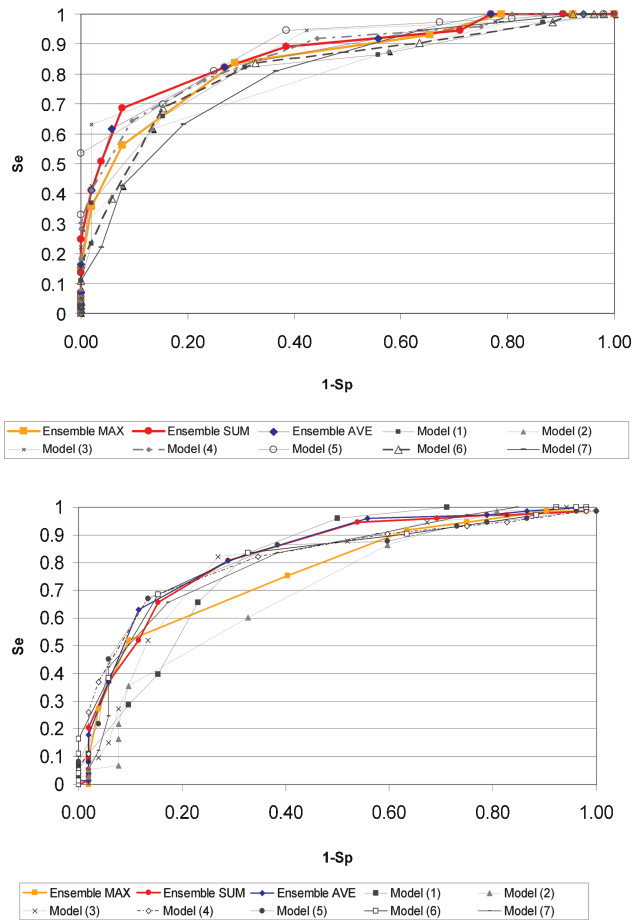

Figure 3. ROC curves comparing the performance of the individual models (1)–(7) to the performance of the three-model ensemble, considering explanatory capability (upper plot, corresponding to training set) and predictive capability (lower plot, corresponding to test set).

ensembles in the high-sensitivity region, but its performance drops considerably below the performance of the ensemble in the high-specificity region. In other words, when using AVE and SUM data fusion schemes, the ensemble performs considerably well on the test set through all the 1-Sp range. We believe that the use of an ensemble of models is better justified from the biochemical point of view, since Pgp has wide substrate specificity and multiple binding sites, suggesting the potential limitations of individual QSAR models to predict the wide range of Pgp substrates.

5. CONCLUSIONS

A set of topological dfs capable of identifying Pgp substrates has been built, whose performance compares quite well with the performance of almost all previously reported models. Remarkably, our models only include low-dimensional descriptors, which makes them adequate for the virtual screening of increasingly large virtual chemical repositories (previous conformational analysis is not required). To the present, relatively few studies exist on exclusively topological *in silico* models to identify potential Pgp substrates. Although easy-to-interpret pharmacophore and grid-based approaches might be more suitable for drug design purposes, topological descriptors look more adequate for virtual screening campaigns.

The ensemble of topological models allowed us to increase specificity, an essential parameter in our current background (limited budget that urges us to make the most efficient possible use of our funding, i.e. to optimize chemical synthesis and pharmacological tests by incorporating rational approaches at the beginning of the drug development project and by increasing, as much as possible, the probability of success at biological testing). ROC curves suggest that the ensemble performs consistently well through all the 1-sp range, in contrast with individual models that tend to perform well in either the high-sensitivity or the high-specificity regions of the curves. Using the average or the sum of the three models for data fusion purposes seems to be a better general strategy than to use the maximum value provided by ensemble models.

Throughout this study we have used previous advances in the field of Pgp substrates modeling in several ways: (a) for comparison purposes (using previous reports as a reference of performance); (b) to develop the ensemble of topological models (by imitating past similar, successful approaches based on combination of pharmacophores or combinations of different kind of models) and (c) to design what we have called 'a rational pool' of descriptors on the basis of the more prominent features of previously reported models. One of the models obtained from the rational pool is present in the best combination of models, merged with other two models obtained from random combinations of Dragon descriptors. In this way, we have combined 'de novo' models (obtained in a rather stochastic way) with already available information from previous studies. Keeping in mind recent advances on the crystallization of Pgp, future efforts should focus on combination of ligand-based and structure-based approaches.

Acknowledgements

AT and EAC are members of the CONICET. LBB is member of the Facultad de Ciencias Exactas, Universidad Nacional de La Plata (UNLP). This research was supported in part through grants from CONICET PIP N011220090100311/510 and Incentivos UNLP.

REFERENCES

1. Schuster D, Laggner C, Langer T. Why drugs fail—a study on side effects in nximahemical entities. *Curr. Pharm. Des.* 2005; **11**: 3545–3559.
2. Kola I, Landis J. Can the pharmaceutical industry reduce attrition rates? *Nat. Rev. Drug Discov.* 2004; **3**: 711–715.
3. Hop C, Cole M, Davidson R, Duignan D, Federico J, Janisewski J, Jenkins K, Krueger S, Lebowitz R, Liston T, Mitchell W, Snyder M, Steyn S, Soglia J, Taylor C, Troutman M, Umland J, West M, Whalen K, Zelesky V, Zhao S. High throughput ADME screening: practical considerations, impact on the portfolio and enabler of *in silico* ADME models. *Curr. Dru. Metab.* 2008; **9**: 847–853.
4. Fromm M. Importance of P-glycoprotein at blood-tissue barriers. *Trends Pharmacol. Sci.* 2005; **25**: 423–429.
5. Taft R. Drug excretion. In *Pharmacology: Principles and Practice*, Hacker M, Messer WS, Bachmann KA (eds). Academic Press: San Diego, USA, 2009; 175–201.
6. Siddiqui A, Kerb R, Weale M, Brinkmann U, Smith A, Goldstein D, Wood N, Sisodiya S. Association of multi-drug resistance in epilepsy with a polymorphism in the drug-transporter gene. *N. Engl. J. Med.* 2003; **348**: 1442–1448.
7. Volk H, Burkhardt K, Potschka H, Chen J, Becker A, Löscher W. Neuronal expression of the drug efflux transporter P-glycoprotein in the rat hippocampus after limbic seizures. *Neurosciene* 2004; **123**: 751–759.
8. Bauer B, Hartz A, Fricker G, Miller D. Pregnane X receptor up-regulation of P-glycoprotein expression and transport function at the blood-brain barrier. *Mol. Pharmacol.* 2004; **66**: 413–419.
9. Thomas H, Coley H. Overcoming multi-drug resistance in cancer: an update on the clinical strategy of inhibitin p-glycoprotein. *Cancer Control* 2003; **10**: 159–165.
10. Kim R, Fromm M, Wandel C, Leake B, Wood A, Roden D, Wilkinson G. The drug transporter P-glycoprotein limits oral absorption and brain entry of HIV-1 protease inhibitors. *J. Clin. Invest.* 1998; **101**: 289–294.
11. Löscher W, Potschka H. Role of drug efflux transporters in the brain for drug disposition and treatment of brain diseases. *Prog. Neurobiol.* 2005; **76**: 22–76.
12. Penzotti J, Lamb M, Evensen E, Grootenhuis P. A computational ensemble pharmacophore model for identifying substrates of P-glycoprotein. *J. Med. Chem.* 2002; **45**: 1737–1740.
13. Xue Y, Yap C, Sun L, Cao Z, Wang J, Chen Y. Prediction of P-glycoprotein substrates by a support vector machine approach. *J. Chem. Inf. Comput. Sci.* 2004; **44**: 1497–1505.
14. Cerqueira Lima P, Golbraikh A, Oloff S, Xiao Y, Tropsha A. Combinatorial QSAR modeling of P-glycoprotein substrates. *J. Chem. Inf. Model.* 2006; **46**: 1245–1254.
15. Cianchetta G, Singleton R, Zhang M, Wildgoose M, Giesing D, Fravolini A, Cruciani G, Vaz R. A pharmacophore hypothesis for P-glycoprotein substrate recognition using GRIND-based 3D-QSAR. *J. Med. Chem.* 2005; **48**: 2927–2935.
16. Zhang L, Balimane P, Johnson S, Chong S. Development of an *in silico* model for predicting efflux substrates in Caco-2 cells. *Int. J. Pharm.* 2007; **343**: 98–105.
17. Li W, Li L, Eksterowicz J, Ling X, Cardozo M. Significance analysis and multiple pharmacophore models for differentiating P-glycoprotein substrates. *J. Chem. Inf. Model.* 2007; **47**: 2429–2438.
18. Chang C, Bahadduri P, Polli J, Swaan P, Ekins S. Rapid identification of P-glycoprotein substrates and inhibitors. *Drug Metab. Dispos.* 2006; **34**: 1876–1984.
19. Ekins S, Kim R, Leake B, Dantzig A, Schuetz E, Lan L, Yasuda K, Shepard R, Winter M, Schuetz J, Wikel J, Wrigton S. Application of three-dimensional quantitative structure-activity relationships of P-glycoprotein inhibitors and substrates. *Mol. Pharmacol.* 2002; **61**: 974–981.
20. Wang R, Kuo C, Lien L, Lien E. Structure-activity relationship: analyses of p-glycoprotein substrates and inhibitors. *J. Clin. Pharm. Ther.* 2003; **28**: 203–228.
21. Boccard J, Bajot F, Di Pietro A, Rudaz S, Boumendjel A, Nicolle E, Carrupt P. A 3D linear solvatation energy model to quantify the affinity of flavonoid derivatives toward p-glycoprotein. *Eur. J. Pharm. Sci.* 2009; **36**: 254–264.

22. Zalloum H, Taha M. Development of predictive in silico model for cyclosporine- and aureobasidin-based P-glycoprotein inhibitors employing receptor surface analysis. *J. Mol. Graph. Model.* 2008; **27**: 439–451.
23. Müller H, Pajeva I, Globisch C, Wiese M. Functional assay and structure-activity relationships of new third-generation P-glycoprotein inhibitors. *Bioorg. Med. Chem.* 2008; **16**: 2448–2462.
24. Huang J, Ma G, Muhammad I, Cheng Y. Identifying P-glycoprotein substrates using a support vector machine optimized by a particle swarn. *J. Chem. Inf. Model.* 2007; **47**: 1638–1647.
25. Cabrera M, González I, Fernández C, Navarro C, Bermejo M. A topological substructural approach for the prediction of P-glycoprotein substrates. *J. Pharm. Sci.* 2006; **95**: 589–606.
26. Bohacek R, Mc Martin C, Guida W. The art and practice of structure-based drug design: a molecular modeling perspective. *Med. Res. Rev.* 1996; **16**: 3–50.
27. Schneider G, So S. *Adaptive Systems in Drug Design*. Landes Bioscience: Georgetown, USA, 2003.
28. Talevi A, Gavernet L, Bruno-Blanch L. Combined virtual screening strategies. *Curr. Comput. Aided Drug Des.* 2009; **5**: 23–37.
29. Seddon A, Curnow P, Booth P. Membrane proteins, lipids and detergents: not just soap opera. *Biochim. Biophys. Acta* 2006; **1666**: 105–117.
30. Aller S, Ward A, Weng Y, Chittaboina S, Zhuo R, Harrell P, Trinh Y, Zhang Q, Urbatsch I, Chang G. Structure of P-glycoprotein reveals a molecular basis for poly-specific drug binding. *Science* 2009; **323**: 1718–1722.
31. Mahar Doan K, Humphreys J, Webster L, Wring S, Shampine L, Serabjit-Singh C, Adkison K, Polli J. 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.
32. Talevi A, Bellera C, Castro E, Bruno-Blanch L. Optimal partition of datasets of QSPR studies: a sampling problem. *MATCH Commun. Math. Comput. Chem.* 2010; **63**: 585–599.
33. Srinivas E, Narashima Murthy J, Raghu Ram Rao R, Narahari Sastry G. Recent advances in molecular modeling and medicinal chemistry aspects of phospho-glycoprotein. *Curr. Drug Metab.* 2006; **7**: 205–217.
34. Raub T. P-glycoprotein recognition of substrates and circumvention through rational drug design. *Mol. Pharm.* 2005; **3**: 3–25.
35. Wiese M, Pajeva I. Structure-activity relationships of multidrug resistance reversers. *Curr. Med. Chem.* 2001; **8**: 685–713.
36. Triballeau N, Acher F, Brabet I, Pin J, Bertrand H. Virtual screening workflow development guided by the 'Receiver Operating Characteristic' Curve approach. Application to high-throughput docking on metabotropic glutamate receptor subtype 4. *J. Med. Chem.* 2005; **48**: 2534–2547.
37. Neyman J, Pearson E. On the problem of the most efficient tests of statistical hypothesis. *Philos. Trans. R. Soc. Lond. A* 1933; **231**: 289–337.
38. Hubbard R, Bayarri M. Confusion over measures of evidence (p's) versus errors (alphas's) in classical statistical testing. *Am. Stat.* 2003; **57**: 171–178.
39. Shapiro A, Fox K, Lam P, Ling V. Stimulation of P-glycoprotein-mediated transport by prazosin and progesterone. Evidence for a third site. *Eur. J. Biochem.* 1999; **259**: 841–850.

