Significance Analysis and Multiple Pharmacophore Models for Differentiating P-Glycoprotein Substrates

Wu-Xiong Li,* Leping Li, John Eksterowicz, Xuefeng Bruce Ling, and Mario Cardozo

Amgen Inc., 1120 Veterans Boulevard, South San Francisco, California 94080

Received August 2, 2007

P-glycoprotein (Pgp) mediated drug efflux affects the absorption, distribution, and clearance of a broad structural variety of drugs. Early assessment of the potential of compounds to interact with Pgp can aid in the selection and optimization of drug candidates. To differentiate nonsubstrates from substrates of Pgp, a robust predictive pharmacophore model was targeted in a supervised analysis of three-dimensional (3D) pharmacophores from 163 published compounds. A comprehensive set of pharmacophores has been generated from conformers of whole molecules of both substrates and nonsubstrates of P-glycoprotein. Four-point 3D pharmacophores were employed to increase the amount of shape information and resolution, including the ability to distinguish chirality. A novel algorithm of the pharmacophore-specific *t*-statistic was applied to the actual structure−activity data and 400 sets of artificial data (sampled by decorrelating the structure and Pgp efflux activity). The optimal size of the significant pharmacophore set was determined through this analysis. A simple classification tree using nine distinct pharmacophores was constructed to distinguish nonsubstrates from substrates of Pgp. An overall accuracy of 87.7% was achieved for the training set and 87.6% for the external independent test set. Furthermore, each of nine pharmacophores can be independently utilized as an accurate *marker* for potential Pgp substrates.

INTRODUCTION

P-glycoprotein (Pgp),[1−3] the product of the multidrug resistance (MDR) genes, is a member of the ATP-binding cassette superfamily of active transporter proteins. It is located in the plasma membrane of mammalian cells with a molecular weight of 170 kDa and is assumed to consist of two homologous halves joined by a linker region, each half containing six transmembrane α-helix segments and a consensus nucleotide binding domain. Pgp is normally expressed at many physiological barriers,[4] including the apical membranes of the epithelia, the luminal surface of the small intestine, colon, capillary endothelial cells of the brain, and kidney proximal tubules. Besides expelling xenobiotic and cytotoxic endogenous chemical agents, the Pgp-mediated efflux pump can efficiently transport a wide variety of clinically important drugs leading to multidrug resistance and changes in pharmacokinetics. Substrates transported by Pgp can be as diverse as cancer therapeutics (doxorubicin and paclitaxel), HIV protease inhibitors (amprenavir and indinavir), cardiac drugs (digoxin and quinidine), and chemicals from many other drug classes. The understanding of Pgp-mediated drug efflux can have implications for improving blood−brain-barrier penetration of central-nervous-system (CNS) drugs, designing chemotherapeutical anticancer drugs, enhancing renal and biliary excretion of substrate drugs, and minimizing Pgp-related drug−drug interactions.

Pgp-mediated drug efflux has been one of the major obstacles to the success of cancer therapeutics, as high expression of Pgp is observed in many cancer cells.[5] One approach to overcome the undesired MDR phenotype is the use of MDR reversal agents that inhibit Pgp transport. Another approach to circumvent MDR is to identify potential Pgp substrates early in the drug discovery process and to select drug candidates that are less likely to be transported by Pgp. The transport activity assessment of MDR reversal agents (i.e., Pgp inhibitors) or Pgp substrates can be achieved experimentally through *in vitro* or *in vivo* assays, or computationally through simulations based on *in silico* models of quantitative structure−activity relationships (QSAR).

In general, three types of *in vitro* assays have been utilized to screen the binding activity of substrates and inhibitors to Pgp.[3] They are (1) transport assays on confluent cell monolayers (e.g., monolayer efflux assay), (2) accumulation and efflux assays using fluorescent probes (e.g., Calcein-AM assay), and (3) ATPase assays that monitor the ATPase activity of Pgp proteins. The monolayer efflux assay, whereby the apparent permeability ratio of basolateral-to-apical direction (B → A) to apical-to-basolateral direction (A → B) of the drug is compared with one in the presence of a Pgp inhibitor, is currently the definitive way for identifying Pgp substrates. However, the monolayer efflux assay is labor-intensive and very low throughput. In addition, it only provides concordant results for those compounds with apparent passive permeability between 20 and 300 nm/s and a mass recovery rate exceeding 50%.[6] For compounds exhibiting high passive permeability (>300 nm/s), the Calcein-AM assay is recommended to detect the percentage inhibition of fluorescence response relative to a positive control that is considered to give a maximum response, while the ATPase assay measures changes in basal ATPase activity in the presence and absence of potential substrates. Both Calcein-AM and ATPase assays offer higher throughput and

* Corresponding author. Tel: (650) 813-1192. E-mail: li5xiong@yahoo.com.

10.1021/ci700284p CCC: $37.00
Published on Web 10/23/2007

are readily automated. However, both tend to underestimate the substrate activity of compounds with low permeability. Moreover, the ATPase assay also suffers from high intra- and interassay variability.

The ultimate determination of the impact of Pgp-mediated efflux on drug pharmacokinetic properties requires *in vivo* examinations based on transgenic or mutant animal models.[3] Transgenic animal models have been produced through gene knockout by removing or silencing genes that express Pgp, while mutant animal models are naturally deficient in the expression of a drug efflux transport proteins (e.g., Pgp). Measuring the CNS uptake of a compound that penetrates the blood−brain barrier is one such way that is frequently utilized. The concentration at half time or area under the curve (AUC) of a compound in brain, blood, or plasma can be analyzed from animal models with comparison to wild-type animals. For a substrate of Pgp, the concentration (or AUC) ratio of brain to blood (or brain to plasma) in animal models can be observed to increase significantly when compared to those in wild-type animals.

The *in vitro* and *in vivo* assays described above are costly, laborious, and time-consuming. They are considered more appropriate for use in the later stages of lead optimization when the candidate compounds exhibit adequate potency and other superior pharmacokinetic properties. On the other hand, *in silico* models for Pgp substrate/inhibitor prediction can provide a means of rapid and cost efficient assessment early on during the lead optimization process. X-ray crystal structures of the transmembrane of Pgp are not yet available at atomic resolution. Consequently, the existing *in silico* models are ligand-based[7−15] and are derived primarily from structure−activity relationships, structural recognition elements, and multiple pharmacophores of Pgp substrates. In contrast to most other transport proteins that recognize a few structurally similar substrates, Pgp recognizes a broad range of pharmacologically and structurally diverse compounds. Multiple SAR studies of related series have revealed the amphiphilic nature of Pgp substrates: the presence of aromatic rings, hydrophobic groups, and nitrogen atoms or hydrogen-bond acceptors.

From an analysis of 3D structures for a large diverse set of drugs, Seelig[13] proposed a general pattern for Pgp substrate recognition comprising two or three electron-donor (or hydrogen-bonding acceptor) groups with a fixed spatial separation of 2.5 ± 0.3 (as a type-I pattern) or 4.6 ± 0.6 Å (as a type-II pattern), respectively.

Ekins et al.[8,9] utilized HIPHOP or HYPOGEN methods in Catalyst[16] to build pharmacophore QSAR models for qualitatively ranking inhibitors that inhibit Pgp-mediated substrate transport. A single substrate pharmacophore was produced by overlaying likely common structures of verapamil and digoxin, followed by fitting vinblastine to this substrate model. Pharmacophore alignment of these compounds revealed multiple hydrophobic and hydrogen-bond acceptor features as important characteristics of Pgp substrates.

Penzotti et al.[12] reported an ensemble model of 100 pharmacophores, consisting of a set of two-, three-, and four-point pharmacophores, to discriminate between Pgp substrates and nonsubstrates. This pharmacophore ensemble was derived from a full ensemble of conformers of 144 compounds and developed through a pairwise comparison of both substrates and nonsubstrates at the class level in terms of the relative information content conveyed by each pharmacophore. A compound matching at least 20 pharmacophores in the ensemble was considered to be a potential Pgp substrate. The model offered an overall classification accuracy of 80% for the training set, but only 63% for a hold-out test set.

de Cerqueira Lima et al.[17] recently explored a combinatorial QSAR modeling of Pgp substrates which employs four typical techniques (k-nearest neighbor classification, decision tree, binary QSAR, and support vector machines) and four sets of commonly used descriptors that include 381 molecular connectivity indices, 173 atom pair descriptors, 72 VolSurf descriptors, and 189 molecular operation environment descriptors, respectively. Among 16 combinations, the model using support vector machines and atom pair descriptors yielded the best predictive accuracy of 81% for their test set. While these types of methods can be powerful and accurate, the use of a large number of descriptors (often >100) and resulting complex models may be not easy for the medicinal chemists to visually or physically interpret. Svetnik et al.[18] investigated boosting tree or bagging tree techniques for Pgp substrate classification, each of which consisted of a sequence of about 100 tree classifers based on 1522 binarized atom pair descriptors. Though this type of method can show higher accuracy than other QSAR methods, such tree ensemble models make predictions by a consensus and do not offer chemists a single straightforward relationship between compound activity and the structural descriptors.

This study aims to develop a global SAR model with improved accuracy that is capable of representing the structural diversity of Pgp substrates. A pharmacophore approach was chosen as it offers a 3D model that can be intuitive to visualize and simple to interpret. This can significantly assist chemists involved in the lead optimization stage of a project.

The strategy we employed to enhance prediction accuracy applies supervised machine learning techniques to the 3D pharmacophore descriptors. In this study, only four-point pharmacophores were used. The reasoning is as follows: (1) Mason et al.[19] observed that it was important to move from three-point to four-point pharmacophores to increase the amount of shape information and resolution, including the ability to distinguish chirality (a fundamental requirement for many ligand−receptor interactions); (2) preliminary model development showed that moving from three-point to four-point pharmacophores resulted in a significant increase in accuracy; (3) the exclusive use of four-point pharmacophores rather than a mixture of two-, three-, and four-point pharmacophores avoids the problematic issue of correlated (or dependent) descriptors.[20]

Unlike traditional pharmacophore approaches[21] where activity prediction models were optimized by perturbing and annealing the chemical features and locations of a very small set of top-ranking pharmacophores, our approach consisted of three sequential procedures: (1) the exhaustive enumeration of all possible pharmacophore configurations for both substrate and nonsubstrate compounds (similar to Penzotti et al.'s work), (2) the identification of a statistically significant optimal ensemble of pharmacophores able to differentiate Pgp substrates from nonsubstrates, and (3) the

development of a simple *in silico* model for Pgp transport activity prediction. In the second procedure, millions of pharmacophores were evaluated for Pgp efflux activity through a frequency analysis of pharmacophore occurrence and then a pharmacophore-specific *t*-statistic. A set of the top-ranked, significant pharmacophores was selected as the basis for the construction of a classification tree. The derived pharmacophores and classification tree were then evaluated for predictive performance against an independent test set of 97 uncorrelated known drugs.

The intention of this article is to present a pragmatic *in silico* model, which has been performing extremely well in our own internal lead optimization projects. It should be noted, however, that some of the top-ranking pharmacophores make use of a nontraditional chemical feature (a ring projection point without the corresponding aromatic ring feature). These pharmacophores function very well and are predictive but perhaps should be best interpreted as QSAR descriptors rather than traditional pharmacophores.

The paper is organized as follows. The Materials and Methods section provides a description of the data set and details of the pharmacophore generation and modeling methodologies. The Results and Discussion section provides details of the performance of the significant pharmacophores and classification model and also a comparison of this type of analysis to ranking by information content. A summary is provided and future work is discussed in the final section.

METHODS AND MATERIALS

Data Set. Available data from the literature were used for the training and test sets consisting of 163[12] and 97[6,22] compounds, respectively. All acid and base functional groups were neutralized.

The 163 training set compounds were obtained from Penzotti et al.'s work[12] and consisted of 91 substrates and 72 nonsubstrates (Supplementary Table A, Supporting Information). CONCORD[23] was used to create single 3D initial geometry from the starting SMILES (Simplified Molecular Input Line Entry System) strings.

The test compounds were compiled from two sources. Mahar Doan et al.[22] measured Calcein AM inhibition and apparent permeability ratios on MDR1-MDCKII cells for 93 marketed drugs. Gombar et al.[10] developed a QSAR Pgp model using 58 compounds whose structures were obtained from the Derwent World Drug Index database.[24] From comparison of these two collections to each other and the training set and by removing compounds with a molecular weight above 700 Da, 97 unique compounds were obtained (Supplementary Table B, Supporting Information).

Pharmacophore Generation. Low-energy conformers of each compound were generated[25] using Catalyst and stored in Catalyst binary databases. For each compound, the maximum number of conformers generated was limited to 100, with a maximum conformational energy cutoff of 20 kcal/mol. The average number of stored conformers per compound was 52.7 for the training set and 48.2 for the test set.

Pharmacophore generation for the training and test sets was conducted with Cerius2.[26] Four-point pharmacophores were enumerated using the chirality flag, the minimal feature separation requirement of 3.0 Å, eight distance bins covering a length of 20 Å, and six feature types: hydrogen-bond acceptor (HBA), hydrogen-bond donor (HBD), hydrophobic (HYD) including aromatic rings and aliphatic chains, negative ionizable (NEGI), positive ionizable (POSI), aromatic ring centroid (RING), and aromatic ring projection point (RNGP). A total of 12.6 million potential four-point pharmacophores were generated from the training set. The mapping of a pharmacophore to any conformer of a compound turns that bit "on" in the index string where each index identifies a unique pharmacophore in Cerius2.

As mentioned previously, nontraditional pharmacophores were considered here, consisting of aromatic ring projection points without requiring that the corresponding aromatic feature also be present. The analysis showed that these pharmacophores are still statistically significant. The presence of this feature in specific cases may imply that the nature of the ligand/receptor interaction is important but not the directionality.

Pharmacophore Significance Analysis. In traditional pharmacophore approaches, typically a small number of compounds are chosen by hand for the training set, and a small number of initial conformational alignments are generated. Using a methodology like HIPHOP/HYPOGEN from the Catalyst package, the pharmacophore SAR model can then be iteratively refined for the best activity prediction by varying the location and identity of the chemical features, followed by a simulated annealing approach.[21] Alternatively, as in GASP[27] from the Sybyl package,[28] common feature pharmacophore patterns can be *genetically* uncovered from the local optimization of the conformational overlay of several flexible molecules by mimicking the process of evolution.[27] However, the goal of this work is to seek a predictive model across a training set of 163 diverse compounds containing potentially millions of pharmacophore hypotheses. This is beyond the capacity of the above methods. With inspiration from Tibshirani et al.'s work[29] on microarray data analysis, a significance analysis of pharamacopores (SAP) for activity model generation was designed and implemented to elucidate comprehensive patterns from large sets of pharmacophores.

The method was conceived to unveil 3D structural patterns of compounds in different activity classes, and such patterns are assumed to be pharmacophore-specific. To discover these patterns, we adopt a two-class *t*-statistic based on the ratio of change in pharmacophore occurrences to standard deviation in compound classes for that pharmacophore. A full implementation of the method has been documented by Tibshirani's group and is available at the following Web site: http://www-stat.stanford.edu/~tibs/SAM/index.html. A concise description of relevant procedures along with our customization is given below.

The ranking score d_j for pharmacophore j is computed as

$$d_j = \frac{\langle x_j^S \rangle - \langle x_j^N \rangle}{s_j + s_0}$$

where $<x_j^S>$ and $<x_j^N>$ are defined as the average occurrences for pharmacophore j in Pgp substrate class S and nonsubstrate class N, respectively. s_j is the standard deviation of occurrence measurements:

$$s_j = \sqrt{\frac{1/n_1 + 1/n_2}{n_1 + n_2 - 2}\left[\sum_{\mathrm{S}}(x_j^{\mathrm{S}} - \langle x_j^{\mathrm{S}}\rangle)^2 + \sum_{\mathrm{N}}(x_j^{\mathrm{N}} - \langle x_j^{\mathrm{N}}\rangle)^2\right]}$$

where x_j^{S} and x_j^{N} are the booleans (1 or 0) of mapping pharmacophore j to a substrate or a nonsubstrate, $\sum_{\mathrm{S}}$ and $\sum_{\mathrm{N}}$ are summations of the occurrence measurements in classes S and N, respectively, and n_1 and n_2 are the numbers of compounds in classes S and N. The constant s_0 is a fudge factor, and its value was chosen to minimize the coefficient of variation of d_j.

Pharmacophores with scores d_i greater than a threshold are considered potentially significant. Meanwhile, a fraction of pharmacophores can be called significant by chance. To estimate the called-by-chance fraction, control data are required to assign statistical significance to the activity effect of each pharmacophore. We generated a large number of controls by computing the ranking scores from 400 sets of permutations of class labels, S and N, with replacement across all compounds. Such permutations of class labels have been assumed to decorrelate activities to the original compounds and were utilized for significance verification of *in silico* SAR modeling.[8,9]

To measure significant changes in pharmacophore occurrences, pharmacophores are sorted in decreasing order of their d_i values, and indices j of d_j are accordingly reassigned to consecutive ascending numbers starting from 1. After reindexing, d_j is the jth largest true score. For each permutation set, artificial ranking scores d_i^p are also calculated, and the pharmacophores are again sorted and reindexed such that d_i^p is the ith largest score for permutation p. The expected ranking score, d_i^E, is computed as $d_i^E = \sum_p d_i^p/400$.

To identify potentially significant changes of pharmacophore occurrences, the true score d_i is compared to the expected score d_i^E. Though for most of them $|d_i| > |d_i^E|$, a smaller portion of pharmacophores can have their displacements $|d_i - d_i^E|$ exceeding a greater threshold (labeled Δ), conveying a higher statistical significance. To determine the number of falsely significant pharmacophores, a horizontal cutoff is defined as the smallest d_i among the significant pharmacophores that contribute to activity and the least negative d_i among the significant pharmacophores that contribute to inactivity. The number of called-by-chance pharmacophores corresponding to each permutation is computed by counting pharmacophores whose artificial scores exceed the horizontal cutoff. The number of falsely significant pharmacophores is estimated at a specific percentile of the numbers of called-by-chance pharmacophores from all 400 permutations. By varying Δ, the number of significant pharmacophores can be changed as well as the number of the falsely called pharmacophores. The number of significant pharmacophores may be optimized when the number of the falsely called pharmacophores is minimized as Δ increases in a bottom-up manner.

Classification Tree. Conventional SAR studies for Pgp have been performed on the basis of classical QSAR principles which were designed for transporters or receptors which naturally bind *one* specific substrate or analog series from an aqueous environment. In these types of studies, the same binding mechanism or mode is often assumed for all modeled conformations, and solvent or membrane effects are considered negligible. Seelig et al.[1] suggested that the classical QSAR methods were not adequate to describe the action mechanism of Pgp as the protein transports not one specific compound but many diverse substrates. This is consistent with the publications from various groups that point to the evidence of multiple drug binding sites for Pgp.[30−32] Furthermore, Pgp differs from other transporters in that it recognizes substrates dissolved in the lipid membrane and not in aqueous solution.[33] The multiple binding mechanisms and membrane partitioning effects present a significant challenge for creating a simplified but accurate QSAR model.

In this study, a tree-based method—classification tree—was taken to circumvent some of the problems associated with traditional QSAR models. The classification tree splits the compounds into different subsets by grouping together the compounds that map to a specific pharmacophore. These pharmacophores can be independent of each other, and this allows for a single tree to account for the potential of multiple binding modes and membrane partitioning effects. In addition, the splitting descriptors are SAP-selected pharmacophores, and their high significance can make the interpretation of the splits less precarious.

The classification tree was constructed using a recursive partition tree (rpart),[34] the algorithm for which has been published by Breiman et al.[35] The Pgp transport activity (or class label) of compounds is input as a target variable, and the optimal set of significant pharmacophores identified through SAP is used as the pool of potential splitting descriptors. The categorical attribute of the target variable (i.e., substrate or nonsubstrate) leads to a model of the classification tree. The root of such a tree starts with a full set of training compounds. Compounds satisfying the specific criterion at each junction are assigned to the right branch, and the others to the left branch. At each junction, the criterion referred to is the presence or the absence of a specific pharmacophore in any sampled conformers of a compound. The classification tree grows by recursively partitioning the compounds. In growing the tree, the classification gain of the Gini index, measuring class purity of the resultant junctions, is maximized at each partitioning step. Across all significant candidates, the pharmacophore conferring the largest gain in the Gini index is selected as the primary criterion to split the compound set at the current junction. If the node has more than 10 compounds, it is eligible to be further split. The minimal size of each child node is set to three compounds. And the resultant tree is back-pruned if its cost complexity decreases by less than 0.02.

RESULTS AND DISCUSSION

Compound Diversity. To illustrate compound diversity in the training set, the Tanimoto pairwise similarity was calculated using Daylight fingerprints. The mean pairwise similarity was 0.21, and the median was 0.17. This indicated that there was significant chemical diversity in the training set and the compounds were not from congeneric series.

The 97 compounds from the test set are shown in Supplementary Table B (Supporting Information). To ensure diversity for this set, all compounds were compared to each other and to the training set compounds. Within the test set,

the average pairwise similarity is 0.21 with a median value of 0.20—indicating sufficient diversity. When the same similarity metric is used, the structural relationship of the test set to the training set was analyzed by computing the similarity between each test compound and the most structurally similar compound in the training set. The pairwise Tanimoto similarity between such compounds is denoted as the maximal similarity of a test compound to the training set. Among 97 test compounds, 81 compounds have a maximal similarity less than 0.70, while only six compounds exceed a maximal similarity of 0.90.[36] The magnitude of similarity coefficients reveals that the test set is structurally diverse itself and distinct from the training set.

Potential Pharmacophores. Of the 12.6 million possible pharmacophores, 5.48 million are unique, and 3.32 million are present in only a single compound. The most frequent pharmacophore maps to 67 compounds.

The ideal pharmacophore is one that is able to discriminate between actives and inactives and is present in a moderate number of compounds. In contrast, pharmacophores with a very low occurrence (e.g., < 2%) or very high occurrence (e.g., >98%) are assumed to be relatively weak class discriminators and not useful for model generation. In this study, the most frequent pharmacophore appears 67 times, well below the median number ($N = 81$) of training compounds, and thus no advantage can be taken from the cutoff for maximal pharmacophore occurrence. The cutoff for minimal pharmacophore occurrence was set to 15 to help reduce the memory requirements of the calculations, resulting in 63 085 pharmacophores of interest.

Elucidation of Significant Pharmacophores. This reduced set of 63 085 potential pharmacophores was taken forward into the significance analysis. A numeric matrix with a dimension of 63 085 rows by 163 columns was constructed representing the pharmacophore indices and compounds, respectively. Each pharmacophore is inspected against the index string of each compound, with the value of "1" assigned to the matrix element of that index if the pharmacophore is present; otherwise, it is set to "0".

SAP was applied to the above matrix, supervised by class labels for 163 compounds of Pgp substrates versus nonsubstrates. A total of 400 permutations of class labels were generated as null reference data. The true classes of compound activities yielded a real *t*-statistic score while the permutations were used to determine an expected score for each pharmacophore. Figure 1 depicts the scatter plot of real scores of pharmacophores versus their expected scores. The solid line suggests an expectation reference, and each diamond marks a pharmacophore. The pharmacophores contributing to active conformers of Pgp substrates stand in the upper right-hand corner, while the pharmacophores associated with Pgp nonsubstrates sit at the bottom left-hand corner.

Nearly 100% of 63 085 pharmacophores were displaced from the expectation line. The relative statistical significances of these pharmacophores can be determined by the magnitude of the displacements. The largest displacement of ordered real scores to null scores is 0.645. Threshold Δ's were set to 100 intervals equally spaced from 0 to 0.645. At each interval Δ, pharmacophores whose scores exceeded the horizontal cutoff were counted separately for the real data set and 400 permutation sets. The number of called-by-chance pharmacophores was computed at the 95th percentile among 400 permutations. As the threshold increases, the possibility of called-by-chance remains high and nearly unchanged until $\Delta = 0.47$; after that, it decreases in a much faster pace than the real significant pharmacophores. As Δ becomes greater than 0.51, the number of falsely called pharmacophores drops to zero, and this results in an optimal set of 598 pharmacophores.

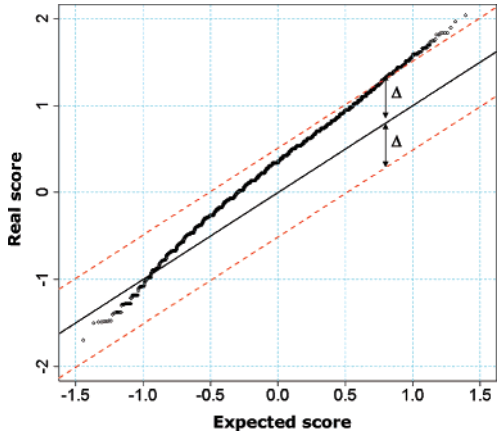

Figure 1. Scatter plot of real *t*-statistic scores versus expected scores for 68 053 pharmacophores. The solid line indicates an expectation reference, while dashed lines show the optimal cutoff ($\Delta = 0.51$) to call the ensemble of 598 significant pharmacophores.

Feature components and spatial arrangements of these significant pharmacophores are summarized in Supplementary Table C (Supporting Information). A total of 83% of the significant pharmacophores contain RING or RNGP features. A total of 67% contain the HYD feature. The HBA feature is present in 86% of the significant pharmacophores, while the HBD feature appears in a lower 28%. The predominance of hydrogen-bond acceptor features rather than hydrogen-bond donors supports the suggestion[13] that the ligand acceptor interactions are the most significant. Furthermore, 218 significant pharmacophores each have two H-bond acceptors, and 23 exhibit the type-II pattern described by Seelig.[13] None relate to the type-I pattern. A total of 45 significant pharmacophores each have three H-bond acceptors; among them, 30 exhibit the type-II pattern, and only one shows the type-I pattern. In the pharmacophore enumeration process, the minimum feature separation requirement of 3.0 Å may prevent the generation of pharmacophores that represent the type-I pattern.

The statistics of chemical features exhibited in pharmacophores of Pgp substrates are mostly in accordance with the findings of other similar studies. Biophysical parameters such as hydrophobicity indices, lipid diffusibility, and hydrogen-bond acceptor strength have been used to characterize structural features of drugs that mediate their interaction with Pgp.[37,38] Another interesting observation is that, although only 35% of the significant pharmacophores contain the POSI feature (generally derived from a basic nitrogen atom), it is present in 11 out of the top 12 most significant pharmacophores—a strong indication of its key role in certain binding modes to Pgp. Actually, it has been addressed by Pearce et al.'s findings that a basic nitrogen atom constituted the common pharmacophore of Pgp.[11] The explanation for this feature presence could be that weak bases can cross

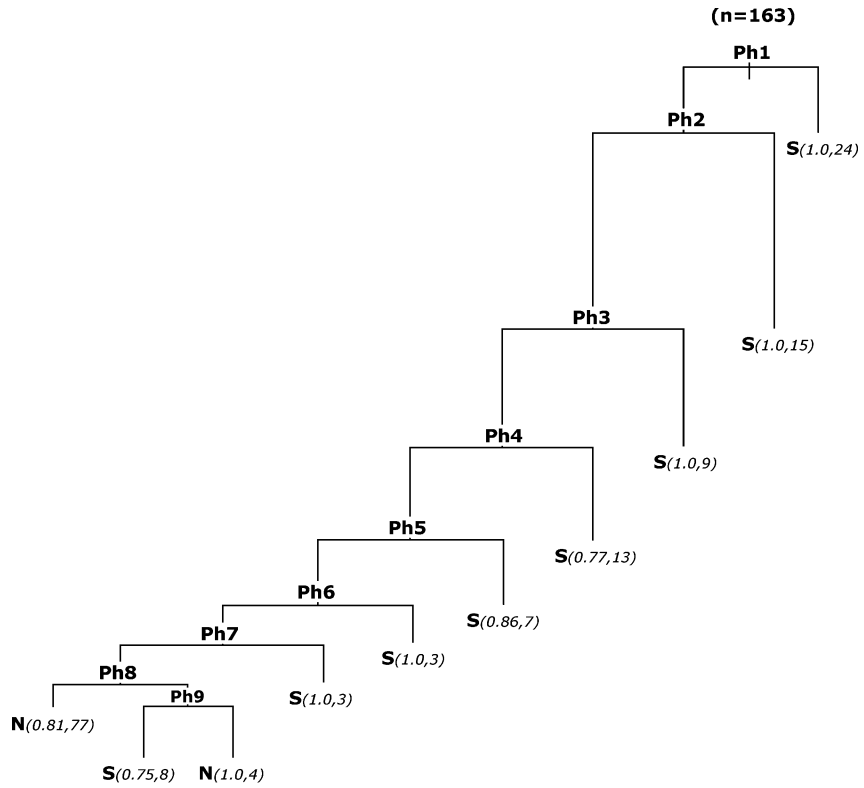

Figure 2. Classification tree for differentiating Pgp substrates derived from 163 training compounds. Each splitting junction has a pharmacophore classifier as indicated. If a compound has the classifier pharmacophore, then it goes to the right branch; otherwise, it goes to the left branch. The text under each leaf node denotes the classification status with **N**, a nonsubstrate class, and **S**, a substrate class of Pgp. The first fraction in parentheses expresses the posterior classification probability, and the second integer corresponds to the number of training compounds in the leaf. The junction height is proportional to its magnitude of the gain of Gini index.

Table 1. Chemical Features and Spatial Arrangements of Nine Pharmacophore Markers of Pgp Substrate[a]

	chemical feature				inter-feature distance (Å)						
Ph label	*f*1	*f*2	*f*3	*f*4	*d*12	*d*13	*d*14	*d*23	*d*24	*d*34	chiral
Ph1	POSI	RNGP	HBA	RING	5	5.59	6.12	5.59	3.54	4.33	−
Ph2	HBA	HBA	HBA	HBD	5	9.01	9.01	7.5	5.59	3.54	+
Ph3	POSI	RNGP	HBA	HYD	5	7.07	6.12	5	3.54	3.54	−
Ph4	HBA	HYD	HYD	HYD	5	9.01	8.29	5.59	4.33	3.54	−
Ph5	RNGP	HYD	HBD	HBA	2.5	5.59	6.12	5	4.33	4.33	−
Ph6	HYD	POSI	RNGP	HBA	5	9.01	10.61	5.59	6.12	4.33	+
Ph7	HBD	HBA	HBA	HYD	5	10.61	12.75	7.91	9.35	3.54	−
Ph8	HBA	HYD	RNGP	HBA	7.5	9.01	10.61	5	4.33	4.33	−
Ph9	HBA	RNGP	HBA	HBD	5	7.07	8.29	5	4.33	4.33	+

Four feature points are labeled as *f*1, *f*2, *f*3, and *f*4. The features are abbreviated as H-bond acceptor (HBA), H-bond donor (HBD), hydrophobic group (HYD), aromatic ring group (RING), aromatic ring projection point (RNGP), and positive ionizable atom (POSI). Distances between *f*1 and *f*2 are labeled as *d*12, and this pattern is followed similarly for remaining distances. In each tetrahedral scheme, the "sign" refers to the sign of the *z* coordinate of feature *f*4: "+" for $z > 0$ and "−" otherwise, given that *f*1 is placed at (0,0,0), *f*2 has $x > 0$, and *f*3 has $y > 0$.

the lipid membrane in the uncharged form and reprotonate in the negatively charged cytosolic leaflet of the membrane.

Classification Tree. Recursive partitioning was applied to the set of 598 discriminate pharmacophores, and the resulting classification tree that best partitions the 163 training compounds contains nine significant pharmacophores, labeled Ph1−Ph9, and is shown in Figure 2. Ph1 and Ph3 were ranked as the top two from the significance analysis, but Ph2 (rank = 29), Ph6 (rank = 59), Ph7 (rank = 64), and Ph9 (rank = 80) come from the top 100, and Ph4 (rank = 530), Ph5 (rank = 498), and Ph8 (rank = 555) are scattered far behind the others.

Table 1 breaks down the components of the significant pharmacophores showing feature type, interfeature distances, and the chirality. All nine pharmacophores contain a hydrogen-bond acceptor feature, and this implies that this is an important recognition feature for Pgp-mediated efflux. The ring projection point appears in six pharmacophores and indirectly shows the importance of aromaticity for efflux. The hydrophobic feature is also present in six pharmacophores. The high representation of these feature types is consistent with the findings from other published studies.[7,11] The positive ionizable feature appears in three pharmacophores, and although it is not prevalent in other pharmacophore studies,[9,12] Pearce et al. showed that a basic nitrogen was an essential feature in his work.[11] Unfortunately, in terms of features and interfeature distances, we failed to align any of these nine pharma-

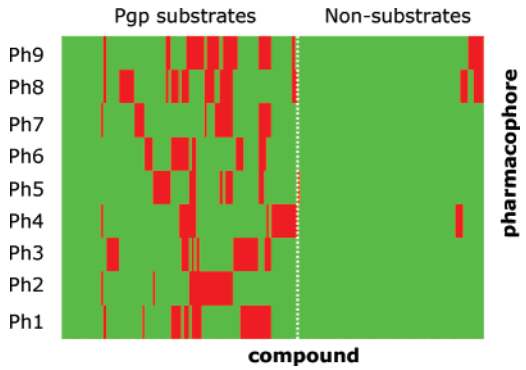

Figure 3. Pharmacophore pattern of Pgp substrates versus nonsubstrates for 163 compounds, with red spots denoting the presence and green the absence of the corresponding pharmacophore in the row to conformers of each compound in the column. Pgp substrates are clustered in the left region and nonsubstrates in the right region, and two regions are separated by a white dashed line. Pharmacophores Ph1−Ph9 are placed bottom-up in the row.

cophores with Ekins and co-workers' models for Pgp inhibitors.[8,9]

Figure 2 shows the optimized classification tree. Given a test compound **A**, its low-energy conformers are first sampled through conformational analysis in Catalyst. The resultant conformers are then mapped to the leading pharmacophore Ph1. If any of the conformers of **A** present a match to Ph1, the compound is classified as a substrate. If there are no matches, the compound progresses down the tree, with every conformer tested against each subsequent pharmacophore for potential classification as a substrate. If none of the nine pharmacophores are matched, then compound **A** would be classified as a non-substrate. The possibility of each assignment is indicated as a decimal fraction in the bracket of each leaf node.

Figure 3 shows a heat map representation of the nine pharmacophores mapped to the 163 training set compounds. The columns correspond to the compounds and the rows to the pharmacophores, with a green block denoting a match; otherwise, the block is colored red. Pgp substrates are clustered in the left region and nonsubstrates in the right region, with the two regions separated by a white dashed line. This figure shows that the model does very well in discriminating between substrates and nonsubstrates.

Ph1 maps to 24 substrates and one example of the alignment are shown in Figure 4. Ph2 maps to 19 substrates, and only four compounds overlap between the two pharmacophores. Ph3 maps to 23 substrates, but 14 of those are mapped by Ph1 as well. And 14 substrates are not identified by any of the pharmacophores. A total of 63 nonsubstrate compounds do not match any of the pharmacophores and thus are correctly classified as such.

In addition, an unusual pharmacophore pattern can be observed from the upper-right corner of the heat map (in Figure 3). If a compound has Ph8 and Ph9 both matched but the seven other pharmacophores not matched, it can be classified as a nonsubstrate. Four compounds in the training set, NSC268251, NSC630148, NSC630721, and NSC674508, show such a pattern. Though the pattern is statistically significant in differentiating the Pgp substrate in the model, further investigation is recommended to determine whether or not it could merely be a statistical artifact.

The classification tree using multiple pharmacophores correctly classifies 87.7% of the compounds as substrates or nonsubstrates. The false negative rate is reasonable at 15.4%, as 14 out of 91 Pgp substrates are incorrectly classified as nonsubstrates. The false positive rate is very low at 8.3%, with only six out of 72 nonsubstrates misclassified.

Pharmacophore Markers. As demonstrated in Figure 2, the classification tree displays a unique structure. Its extreme asymmetry and simplicity enable each pharmacophore classifier to be considered as a *marker*. Similar to the concept of a *biomarker* in a clinical diagnostic test, we are able to use these classifiers as pharmacophore markers for Pgp substrates. That is to say, independent from the mapping status of other pharmacophores, the presence of any pharmacophore marker in a 3D molecular structure can detect a substrate of Pgp with high confidence. As in a diagnostic test, the positive predictive value (PPV) or negative predictive value (NPV) is used to estimate the predictive performance of the pharmacophore markers.

In this work, the test criterion is whether or not a given compound maps to a specific pharmacophore. If the compound matches the pharmacophore, the test result is considered positive and the compound is classified as a substrate. The PPV of a pharmacophore marker is the probability that a compound predicted to be a substrate can be confirmed with real efflux assays. Table 2 summarizes the predictive performance of each pharmacophore when employed independently to discriminate Pgp substrates. Three categories of pharmacophore markers can be inferred by their prediction performance and SAP rankings: strong descriptors (Ph1, Ph2, Ph3, Ph6, and Ph7) with a perfect PPV of 1.00, moderate descriptors (Ph4 and Ph5) with a PPV of around 0.90, and weak descriptors (Ph8 and Ph9) with a PPV of around 0.80. The scale from strong to weak denotes the relative confidence, from high to low, of compounds classified as Pgp substrates to be confirmed by Pgp activity assays.

Similarly, the NPV of a pharmacophore can be interpreted as the probability that a compound predicted to be a nonsubstrate is truly a nonsubstrate. But each of the nine pharmacophores has a NPV value of only around 0.50 despite the fact that each has a high PPV. While, individually, the probability of a pharmacophore correctly predicting a nonsubstrate is low, negative mappings to all pharmacophores together can predict a nonsubstrate with a high confidence of 0.81.

Model Evaluation. The actual classes of Pgp activity of the test compounds have been corroborated mainly by multiple experiments reported from *in vitro* assays of monolayer efflux or Calcein-AM and *in vivo* assays in animal models. Compounds having their membrane permeability A → B in the range of 20 to 300 nm/s were assigned an actual class in terms of their efflux ratios, B → A/A → B. This is the same criteria as in Polli et al.'s work;[6] namely, if B → A/A → B > 1.5 and collapses to ∼1.0 in the presence of a Pgp inhibitor, the compounds are then assigned to Pgp substrate; otherwise, they were assigned to Pgp nonsubstrate. For highly permeable compounds A → B > 300 nm/s, the Calcein-AM assay result is the alternative reference for efflux activity, resulting in a Pgp substrate if a positive response (>10% maximum inhibition response) is detected. For very

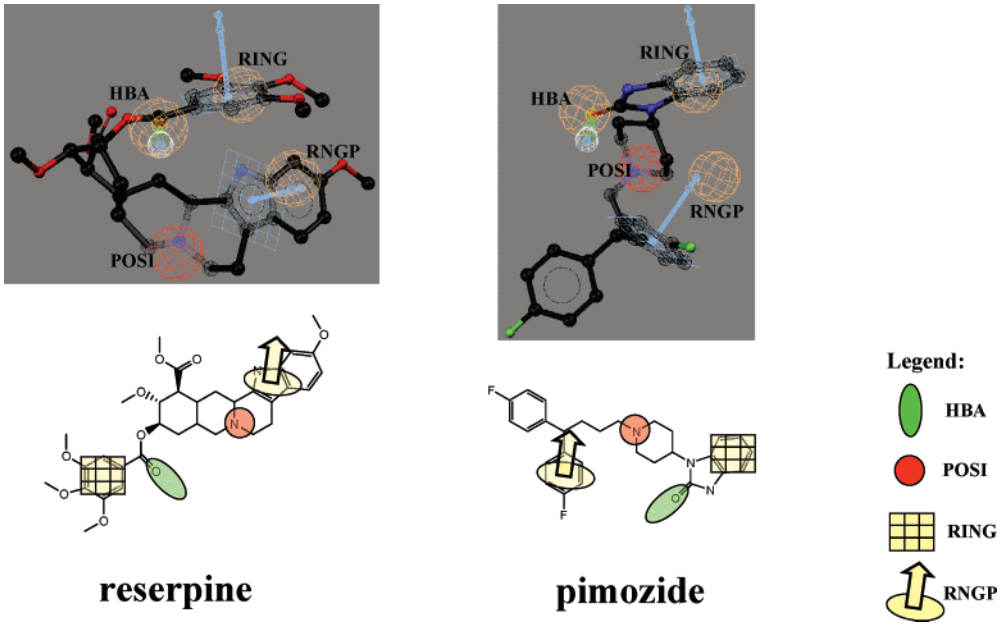

Figure 4. (Top) Alignments in Catalyst of pharmacophore Ph1 to two Pgp substrates: reserpine from training set and pimozide from test set. In both overlays, the positive charge ionizabilty point in red aligns with the aliphatic tertiary amine ($-NR_2$, blue atoms). The light blue arrow points from an aromatic ring centroid to its projection point; the hydrogen-bond acceptor aligns with the carbonyl groups (>C=O, red double bond), and its attached cone indicates a hydrogen-bonding direction. The spheres indicate feature points mapped to Ph1. (Bottom) 2D structures of reserpine and pimozide, and sketches of their alignments with Ph1.

Table 2. The Summary for the Alignments of Pharmacophore Markers to the Training Compounds, and Their Posterior Positive Predictive Values for Pgp Substrate

Ph labels	absence from nonsubstrates	absence from substrates	presence in nonsubstrates	presence in substrates	positive predictive value
Ph1	72	67	0	24	1.00
Ph2	72	72	0	19	1.00
Ph3	72	68	0	23	1.00
Ph4	69	72	3	19	0.86
Ph5	71	75	1	16	0.94
Ph6	72	73	0	18	1.00
Ph7	72	73	0	18	1.00
Ph8	65	66	7	25	0.78
Ph9	66	64	6	27	0.82

low permeable compounds A → B < 20 nm/s, their assignments were taken from the consensus of the above two studies, or from the literature such as Seelig et al.'s work if conflicting assignments exist. There are several special assignments in Supplementary Table B (Supporting Information). The compound metegoline was assigned to Pgp substrate because it showed the largest percentage (84%) of inhibition among all Calcein-AM assays in Mahar Doan et al.'s work. The compound zolmitriptan, with a membrane permeability A → B = 2.52 nm/s, was reportedly crossing the brain−blood barrier[39] and clinically categorized as a nonsubstrate of Pgp.[40] In contrast, the molecule sumatriptan, with a relatively higher membrane permeability A → B = 3.99 nm/s, was reported not to cross the brain−blood barrier in the preclinical studies[41] and exhibited the least CNS-adverse effects among all triptan analogs, which may imply a borderline substrate.

The predicted and experimental results for the test set are given in Supplementary Table B (Supporting Information). In this set, 64 compounds were classified as Pgp nonsubstrates, and 55 of them were experimentally confirmed as true negatives, resulting in an accuracy of NPV = 85.9%. Nine substrates including cetirizine, chloroquine, daunorubicin, labetolol, mequitazine, methysergide, nalbuphine, protriptyline, and acrivastine were misclassified as nonsubstrates. Meanwhile, 33 compounds were classified as Pgp substrates, and 30 of them were confirmed as true positives, resulting in an accuracy of PPV = 90.9. Illustrated in Figure 4, test compound pimozide and training compound reserpine are dissimilar in their 2D molecular structure, but both contain pharmacophore Ph1. Reserpine is a substrate and pimozide was correctly predicted as a substrate as well. Three nonsubstrates including doxylamine, oxprenolol, and warfarin were misclassified as substrates of Pgp. The overall success rate is up to 87.6% for the independent test.

Comparison to Ranking by Information Content. To clarify the performance of the SAP method, SAP-derived ranking has been compared to another ranking via information content.[20,42] To address this question, the same number of 63 085 four-point pharmacophores were ranked by information content[43] and examined individually against the ranking from SAP. The overlap between the methods was determined for a series of ensemble sizes ranging from N = 1 to 1000. For example, for a size of N = 100, there are 49 pharmacophores present in both the SAP ensemble and the information-content ensemble; therefore, the overlapping rate can be computed as 0.49. It turns out that both rankings yield the same top two pharmacophores. As depicted in Supplementary Figure A (Supporting Information), when the ensemble size grows from N = 4 to N = 30, the overlapping rate averages to about 0.65; from N = 30 to 180, it averages to 0.52 and can drop as low as 0.46 at N = 95; when N > 200, it shows a tendency to increase with an average of 0.68 and rises to 0.80 at N = 1000. In addition, different rankings for Ph2 and Ph4 through Ph9 are identified as 11, 717, 266, 30, 25, 1322, and 457, respectively. As a whole, both pharmacophore rankings are consistent but not identical.

CONCLUSION

In this study, a 3D pharmacophore approach was combined with a statistical analysis methodology to develop a model that differentiates Pgp substrates from nonsubstrates. Instead of merely estimating the statistical significance of pharamcophores, permutations of activity classes of compounds were utilized to compute the probability of called-by-chance pharmacophores, and further to determine the optimal size of the pharmacophore ensemble. The analysis revealed that hydrogen-bond acceptors, aromatic rings, and hydrophobic groups are essential features for substrate activity and that a positive ionizable feature can also play a distinct role. A simple and characteristic classification tree that comprises only nine unique pharamcophores was developed from the optimal ensemble of pharmacophores and achieved an overall success rate of 87.7% for the training set and 87.6% for the independent test set. In addition, the nine pharmacophores from the classification tree also performed well individually in identifying Pgp substrates. Furthermore, the Pgp model has also demonstrated in-house success. A few hundred compounds from more than 10 internal projects were evaluated by the model, and 82% of the compounds flagged for Pgp activity were confirmed in either *in vitro* or *in vivo* assays.[44]

The Pgp pharmacophore model developed here exhibits a particularly high degree of accuracy. Moreover, the model also performs well on the independent internal data set, further supporting the robust nature of the methodology. We speculate that the exhaustive enumeration of the four-point pharmacophore hypotheses and the robustness of the SAP method are pivotal to success. The principle of the approach is general and can be applicable to multiclass classification as well as a regression analysis of quantitative targets such as IC_{50} assays or affinity measurements of ligand−receptor binding.

ACKNOWLEDGMENT

The authors thank Dr. Yaxiong Sun for his valuable discussion on model development; gratefully acknowledge Drs. Nigel Walker, Julio Medina, and Steve Young for their critical reading and comments on the manuscript; and thank our colleagues Dr. Zheng Pan, Epic Ding, Jane Liu, and Walter Pan for their enormous support.

Supporting Information Available: The overlapping percentage of a significant pharmacophore ensemble obtained from information content theory versus from SAP analysis was plotted versus the ensemble size (Supplementary Figure A). The training set, compound names, and their actual classification status and predicted status of Pgp substrate (Supplementary Table A). The test set, compound names, predicted status, and experiment-inferred status of Pgp substrate (Supplementary Table B). Chemical features and spatial arrangements of 598 significant pharmacophores (Supplementary Table C). This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Seelig, A.; Landwojtowicz, E.; Fisher, H.; Blatter, X. L. Towards P-glycoprotein structure-activity relationships. In *In drug bioavailability/estimation of solubility, permeability and absorption*; Waterbeemd, L. a. A., Ed.; Wiley/VCH: Weinheim, Germany, 2003; pp 461−492.

(2) Stouch, T. R.; Gudmundsson, O. Progress in understanding the structure-activity relationships of P-glycoprotein. *Adv. Drug Delivery Rev.* **2002**, *54*, 315−328.

(3) Zhang, Y.; Bachmeier, C.; Miller, D. W. In vitro and in vivo models for assessing drug efflux transporter activity. *Adv. Drug Delivery Rev.* **2003**, *55*, 31−51.

(4) Loscher, W.; Potschka, H. Drug resistance in brain diseases and the role of drug efflux transporters. *Nat. Rev. Neurosci.* **2005**, *6*, 591−602.

(5) Cordon-Cardo, C.; O'Brien, J. P.; Boccia, J.; Casals, D.; Bertino, J. R.; Melamed, M. R. Expression of the multidrug resistance gene product (P-glycoprotein) in human normal and tumor tissues. *J. Histochem. Cytochem.* **1990**, *38*, 1277−1287.

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

(7) Osterberg, T.; Norinder, U. Theoretical calculation and prediction of P-glycoprotein-interacting drugs using MolSurf parametrization and PLS statistics. *Eur. J. Pharm. Sci.* **2000**, *10*, 295−303.

(8) Ekins, S.; Kim, R. B.; Leake, B. F.; Dantzig, A. H.; Schuetz, E. G.; Lan, L. B.; Yasuda, K.; Shepard, R. L.; Winter, M. A.; Schuetz, J. D.; Wikel, J. H.; Wrighton, S. A. Three-dimensional quantitative structure-activity relationships of inhibitors of P-glycoprotein. *Mol. Pharmacol.* **2002**, *61*, 964−973.

(9) Ekins, S.; Kim, R. B.; Leake, B. F.; Dantzig, A. H.; Schuetz, E. G.; Lan, L. B.; Yasuda, K.; Shepard, R. L.; Winter, M. A.; Schuetz, J. D.; Wikel, J. H.; Wrighton, S. A. Application of three-dimensional quantitative structure-activity relationships of P-glycoprotein inhibitors and substrates. *Mol. Pharmacol.* **2002**, *61*, 974−981.

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

(11) Pearce, H. L.; Safa, A. R.; Bach, N. J.; Winter, M. A.; Cirtain, M. C.; Beck, W. T. Essential features of the P-glycoprotein pharmacophore as defined by a series of reserpine analogs that modulate multidrug resistance. *Proc. Natl. Acad. Sci. U.S.A.* **1989**, *86*, 5128−5132.

(12) Penzotti, J. E.; Lamb, M. L.; Evensen, E.; Grootenhuis, P. D. A computational ensemble pharmacophore model for identifying substrates of P-glycoprotein. *J. Med. Chem.* **2002**, *45*, 1737−1740.

(13) Seelig, A. A general pattern for substrate recognition by P-glycoprotein. *Eur. J. Biochem.* **1998**, *251*, 252−261.

(14) Raub, T. J. P-glycoprotein recognition of substrates and circumvention through rational drug design. *Mol. Pharm.* **2006**, *3*, 3−25.

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

(16) *Catalyst*, version 4.9; Accelrys Software Inc.: San Diego, CA, 2005.

(17) de 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.

(18) Svetnik, V.; Wang, T.; Tong, C.; Liaw, A.; Sheridan, R. P.; Song, Q. Boosting: an ensemble learning tool for compound classification and QSAR modeling. *J. Chem. Inf. Model.* **2005**, *45*, 786−799.

(19) Mason, J. S.; Morize, I.; Menard, P. R.; Cheney, D. L.; Hulme, C.; Labaudiniere, R. F. New 4-point pharmacophore method for molecular similarity and diversity applications: overview of the method and applications, including a novel approach to the design of combinatorial libraries containing privileged substructures. *J. Med. Chem.* **1999**, *42*, 3251−3264.

(20) Bradley, E. K.; Beroza, P.; Penzotti, J. E.; Grootenhuis, P. D.; Spellmeyer, D. C.; Miller, J. L. A rapid computational method for lead evolution: description and application to alpha(1)-adrenergic antagonists. *J. Med. Chem.* **2000**, *43*, 2770−2774.

(21) Kurogi, Y.; Guner, O. F. Pharmacophore modeling and three-dimensional database searching for drug design using catalyst. *Curr. Med. Chem.* **2001**, *8*, 1035−1055.

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

(23) Pearlman, R. S. Rapid generation of high quality approximate 3-dimension molecule structures. *Chem. Des. Auto. News* **1987**, 2.

(24) Derwent World Drug Index. http://www.derwent.com (accessed Mar 2005).

(25) Smellie, A.; Teig, S. L.; Towbin, P. Poling: promoting conformational coverage. *J. Comput. Chem.* **1995**, *16*, 171−187.

(26) *Cerius2*, version 4.11; Accelrys Software Inc.: San Diego, CA, 2006.

(27) Jones, G.; Willett, P.; Glen, R. C. A genetic algorithm for flexible molecular overlay and pharmacophore elucidation. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 532−549.

(28) *Sybyl*, version 7.3.1; Tripos Inc.: St. Louis, MO, 2006.
(29) Tusher, V. G.; Tibshirani, R.; Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 5116−5121.
(30) Shapiro, A. B.; Fox, K.; Lam, P.; Ling, V. Stimulation, of P-glycoprotein-mediated, drug transport by prazosin and progesterone. Evidence for a third drug-binding site. *Eur. J. Biochem.* **1999**, *259*, 841−850.
(31) Shapiro, A. B.; Ling, V. Positively cooperative sites for drug transport by P-glycoprotein with distinct drug specificities. *Eur. J. Biochem.* **1997**, *250*, 130−137.
(32) Martin, C.; Berridge, G.; Higgins, C. F.; Mistry, P.; Charlton, P.; Callaghan, R. Communication between multiple drug binding sites on P-glycoprotein. *Mol. Pharmacol.* **2000**, *58*, 624−632.
(33) Raviv, Y.; Pollard, H. B.; Bruggemann, E. P.; Pastan, I.; Gottesman, M. M. Photosensitized labeling of a functional multidrug transporter in living drug-resistant tumor cells. *J. Biol. Chem.* **1990**, *265* (7), 3975−80.
(34) R-project. http://www.r-project.org (accessed Mar 2005).
(35) Breiman, L.; Friedman, J.; Olshen, R. A.; Stone, C. J. *Classification and Regression Trees*; Wadsworth: Belmont, CA, 1983.
(36) These six compounds were identified as doxorubicin, indinavir, mitoxantrone, perphenazine, saquinavir, and taxol, each related to a distinct training molecule.
(37) Ecker, G.; Huber, M.; Schmid, D.; Chiba, P. The importance of a nitrogen atom in modulators of multidrug resistance. *Mol. Pharmacol.* **1999**, *56*, 791−796.
(38) Chiba, P.; Ecker, G.; Schmid, D.; Drach, J.; Tell, B.; Goldenberg, S.; Gekeler, V. Structural requirements for activity of propafenone-type modulators in P-glycoprotein-mediated multidrug resistance. *Mol. Pharmacol.* **1996**, *49*, 1122−1130.
(39) Rolan, P. E.; Martin, G. R. Zolmitriptan: a new acute treatment for migraine. *Expert Opin. Invest. Drugs* **1998**, *7*, 633−652.
(40) Calson, S. E. Presentation: Migraine. http://www.npgs.org/downloads/Migraine.ppt (accessed Sep 2007).
(41) Humphrey, P. P.; Feniuk, W.; Marriott, A. S.; Tanner, R. J.; Jackson, M. R.; Tucker, M. L. Preclinical studies on the anti-migraine drug, sumatriptan. *Eur. Neurol.* **1991**, *31*, 282−290.
(42) Srinivasan, J.; Castellino, A.; Bradley, E. K.; Eksterowicz, J. E.; Grootenhuis, P. D.; Putta, S.; Stanton, R. V. Evaluation of a novel shape-based computational filter for lead evolution: application to thrombin inhibitors. *J. Med. Chem.* **2002**, *45*, 2494−2500.
(43) As used in ref 42, the information content, I, is calculated with the following equation: $I = -1/N(N_a \log\{N_a\}/\{N\} + N_i \log\{N_i\}/\{N\}) + N_p/N(p_{ap} \log p_{ap} + p_{ip} \log p_{ip}) + N_n/N(p_{an} \log p_{an} + p_{in} \log p_{in})$ where N is the total number of compounds (substrates and nonsubstrates); N_a is the total number of substrates; p_{ap} and p_{ip} are the fractions of substrates and nonsubstrates, respectively, that have positive predictions (N_p); N_i is the total number of nonsubstrates; and p_{an} and p_{in} are the fractions of substrates and nonsubstrates, respectively, that have negative predictions (N_n). There are two terms in the information content equation: the first represents the uncertainty as to whether a molecule is a substrate, and the second accounts for the uncertainty as to whether a molecule is a substrate given whether it fits pharmacophore model.
(44) The true positive rate for Pgp substrates is 87%, and the true negative rate for Pgp nonsubstrate is 70%.

CI700284P