Journal of
Medicinal
Chemistry

© Copyright 2002 by the American Chemical Society

Volume 45, Number 9 April 25, 2002

Letters

A Computational Ensemble Pharmacophore Model for Identifying Substrates of P-Glycoprotein

Julie E. Penzotti, Michelle L. Lamb,
Erik Evensen, and Peter D. J. Grootenhuis*

Deltagen Research Laboratories,
4570 Executive Drive, Suite 400,
San Diego, California 92121

Received January 17, 2002

Abstract: P-glycoprotein (P-gp) functions as a drug efflux pump, mediating multidrug resistance and limiting the efficacy of many drugs. Clearly, identification of potential P-gp substrate liability early in the drug discovery process would be advantageous. We describe a multiple-pharmacophore model that can discriminate between substrates and nonsubstrates of P-gp with an accuracy of 63%. The application of this filter allows large virtual libraries to be screened efficiently for compounds less likely to be transported by P-gp.

Introduction. P-glycoprotein (P-gp), encoded by the highly conserved MDR (multidrug resistance) genes, is a 170 kDa member of the ATP-binding cassette superfamily of membrane transporters. As a membrane transport protein, P-gp is distinctive in that it transports a wide variety of xenobiotic and cytotoxic endogenous chemical agents out of the cell at the expense of ATP hydrolysis. Substrates transported by P-gp include chemically and mechanistically unrelated drugs such as cancer therapeutics doxorubicin and paclitaxel, HIV protease inhibitors amprenavir and indinavir, cardiac drugs digoxin and quinidine, and chemicals from many other drug classes. P-gp is abundant in cells with a protective barrier function, including the luminal membrane of the endothelial cells comprising the blood-brain barrier and the apical membrane of mucosal cells in the intestine. The effect of P-gp mediated drug efflux on limiting intestinal absorption and oral bioavailability and on tissue distribution (e.g., brain penetration) can have implications for the efficacy of drug regimens that include P-gp substrates.

P-gp mediated drug efflux is one of the major obstacles in the success of cancer therapeutics, as high expression of P-gp is observed in many cancer cells.[1] Many cytotoxic chemotherapeutic agents, such as anthracyclines or vinca alkaloids, induce the up-regulation and overexpression of P-gp. High levels of P-gp result in a lower intracellular accumulation of drug and an increase in efflux. As a result of exposure to an “inducer”, the cells become resistant to a broad spectrum of structurally and mechanistically dissimilar cytotoxic drugs. This phenomenon is known as multidrug resistance. P-gp overexpression may also play a significant role in the development of resistance to antibiotics[2] and in diminished efficacy of HIV drugs.[3,4] One promising approach to overcoming the MDR phenotype employs compounds that inhibit P-gp transport as MDR reversal (MDRR) agents. Several studies have investigated the structure−activity relationships for MDRR agents to derive models that may be useful in designing new MDRR agents.[5,6] Another approach to circumvent MDR is to identify the potential for P-gp activity in compounds early in the drug discovery process and to select drug candidates that are less likely to be transported by P-gp. Knowledge of the factors that determine P-gp substrate specificity is crucial for this second approach.

The molecular mechanism of P-gp mediated transport is not well understood. In part, this is due to the lack of an atomic resolution structure for this transmembrane protein composed of two homologous halves, each containing six putative transmembrane α-helices.[7−9] In the absence of a high-resolution structure for P-gp, the structural features for recognition must be derived from analyses of the chemicals transported by P-gp. In contrast to other transport proteins, which recognize specific classes of compounds such as peptides or carbohydrates, P-gp exhibits a very broad specificity in substrate recognition. A number of structure−activity studies for related series of compounds have identified structural properties reflecting the amphiphilic nature of compounds that interact with P-gp: the presence of aromatic ring structures,[10] hydrophobicity, in general,[11] and nitrogens or hydrogen bond acceptor groups.[12] A

* To whom correspondence should be addressed. Tel: (858) 625-6401. Fax: (858) 625-6487. E-mail: pgrootenhuis@combichem.com.

10.1021/jm0255062 CCC: $22.00 © 2002 American Chemical Society
Published on Web 03/26/2002

structure−activity study of propafenone analogues demonstrated a strong correlation between hydrogen bond acceptor strength and P-gp inhibitory potency.[12] Another computational study of 22 diverse drugs revealed that molecular descriptors associated with strong hydrogen bonding strength and high polarizability promote increased P-gp ATPase activity.[13] From an analysis of three-dimensional structures for a larger diverse set of drugs, Seelig identified two specific recognition elements for P-gp composed of hydrogen bond acceptor units with distinct spatial arrangements.[14] Seelig refers to two hydrogen bond acceptors separated by 2.5 ± 0.3 Å as a type I pattern. Type II patterns are formed by two hydrogen bond acceptors separated by 4.6 ± 0.6 Å or three hydrogen bond acceptors separated by 2.5 ± 0.3 Å with a 4.6 ± 0.6 Å separation of the outer two acceptor groups. Molecules with at least one type I or type II unit are predicted by Seelig to be transported by P-gp, and molecules containing one or more type II units are predicted by Seelig to induce overexpression of P-gp.

In this Letter, we describe the automated generation of a computational model, composed of a set or ensemble of two-to-four point pharmacophores, which discriminates between P-gp substrates and nonsubstrates. The model has an overall classification success rate of 80% for the training set and 63% for a hold-out set. The predominant chemical features of the pharmacophore ensemble reflect the amphiphilic nature of P-gp substrates and include combinations of hydrophobic or aromatic groups, hydrogen bond acceptors, and hydrogen bond donors. The ensemble model can be used as a computational filter to rapidly screen large virtual libraries to deselect compounds that are likely to be substrates for P-gp.

Dataset Description. Using the Seelig study as a starting point,[14,15] we assembled a data set of 195 compounds from literature sources and personal communications. Compounds were classified as P-gp substrates if they were reported to be transported by P-gp; this class includes many compounds that are also reported to induce the overexpression of P-gp and thereby contribute to MDR, such as phenobarbital.[16] Compounds were designated P-gp nonsubstrates if they were not transported by P-gp. This set includes compounds such as progesterone and tamoxifen that are noted to be bound but not transported by P-glycoprotein. This binary classification scheme reduces the error associated with the amalgamation of measurements from different labs, assays, cell types, and conditions.

A hold-out data set (∼25%) was selected randomly from the 195 compounds collected from the literature. This test set contained 32 P-gp substrates and 19 nonsubstrates. The remaining 144 compounds including 76 substrates and 68 nonsubstrates were used as the training data set to derive the computational model.

The average pairwise Tanimoto similarity calculated for the Daylight[17] fingerprints of all 195 compounds is 0.18, which reflects the broad chemical diversity of the data set. Likewise, the training data set and the test data set share an average pairwise Daylight similarity of 0.18. Comparing the P-gp substrates to the P-gp nonsubstrates gives an average Daylight similarity of 0.17.

Computational Model Building. In constructing the computational filter, we used pharmacophore-based three-dimensional whole molecule descriptors. Pharmacophore descriptions of molecules and their application to virtual library searching and design have been described elsewhere[18] and will only be summarized here. As an example, one major component of our 3D descriptors is the "four-point pharmacophore", which consists of four chemical features, selected from hydrogen-bond acceptors and donors, hydrophobes, negative and positive charges, and aromatic groups, and the associated six interfeature distances. Similar to Mason and Cheney,[19,20] a "molecular signature" was created for each molecule by generating a full conformational model, followed by mapping the presence or absence of all 2-, 3-, and 4-point pharmacophores that were present in the molecule's conformers into a single bit string. The set of possible pharmacophores was constrained such that each pharmacophore could contain at most two hydrophobic features. This resulted in a signature length of ∼12 million bits.

We generated the conformational model for each compound using an in-house program, CONAN.[21,22] In the generation of conformers for the P-gp data set, we allowed for a maximum number of conformers of 100 per stereoisomer. The resulting average number of conformers per molecule was 111.

To select the subset of all pharmacophores that is best able to differentiate substrates and nonsubstrates, the signatures from the 144 structurally diverse training set compounds were systematically analyzed in the context of their associated activity data. We considered each of the 3 million pharmacophores sampled by these molecules to be a separate hypothesis that potentially predicts P-gp activity. Each pharmacophore was ranked on the basis of its ability to discriminate between substrates and nonsubstrates. The ranking criterion was "information content", which was calculated using a previously published formula.[23] The pharmacophores with the greatest information content taken together comprised the "ensemble model" for P-glycoprotein activity. To assess the significance of the information content values, the substrate and nonsubstrate assignments were randomly permuted in 50 trials to determine the (average) information content level expected from random chance correlations in the data set. This test revealed that ∼3450 pharmacophores had values for information content greater than the expected random noise level for information content in the data set. Further, ∼1030 of the pharmacophores had information content values greater than one standard deviation above the mean random noise.

Ensembles of the top 1000, 500, 200, and 100 pharmacophores were further analyzed by plotting the fraction of the compounds in each set (substrate and nonsubstrate) versus the number of pharmacophores from the ensemble matched. Visual inspection of these plots revealed overall similar performance for each ensemble size; for each, it was possible to identify an ensemble filter that would flag a large fraction of the substrate data set and pass a large portion of the nonsubstrate molecules. The final informative ensemble selected to balance these goals contained 100 pharmacophores, and its performance is demonstrated in Figure 1.

Once an ensemble model is developed that has such discriminating behavior, criteria for virtual library search/screening are determined. The number of phar-

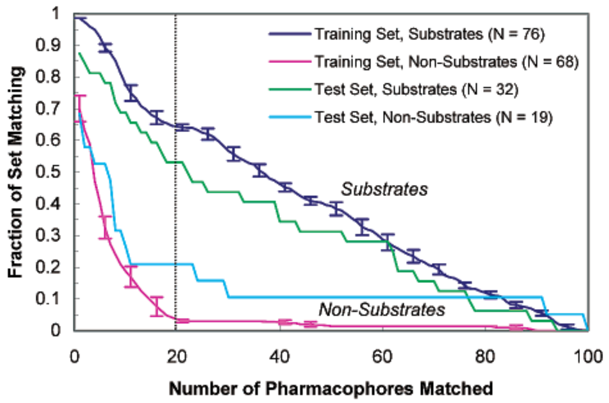

Figure 1. Pharmacophore ensemble model performance graph for the P-gp. The ensemble model is evaluated by plotting the number of pharmacophores matched from the ensemble versus the fraction of molecules (substrate and nonsubstrate). The graph shows the average results of 10 leave-one-out cross-validation trials with the error bars indicating the standard deviation of the trials for the training substrates (blue) and nonsubstrates (magenta). The graph also depicts the performance of the model on the test set, shown for the substrates (green) and nonsubstrates (cyan). A threshold of 20 pharmacophores matched provides a good model for discriminating between P-gp substrates and nonsubstrates.

Table 1. Classification Success When at Least 20 Pharmacophores in the Ensemble Model Are Matched

data set	no. in set	% classified as substrates/ nonsubstrates	no. classified correctly/ incorrectly	% correctly classified overall
		Training Set		
substrates	76	64/36	49/27	
nonsubstrates	68	4/96	66/2	
all	144		115/29	80
		Test Set		
substrates	32	53/47	17/15	
nonsubstrates	19	21/79	15/4	
all	51		32/19	63

macophores, or fraction, of the ensemble matched that provides the greatest separation between the substrates and nonsubstrates is chosen as the threshold for a virtual filter. Using Figure 1, the threshold for the P-gp ensemble was set at 20 pharmacophores matched. When this threshold is applied to a pool of compounds, those matching at least 20 of the 100 pharmacophores in the ensemble are likely to be P-gp substrates. A large fraction of the substrates is identified correctly by this filter, while a large fraction of the nonsubstrates is not “flagged”. With this model, all compounds that match fewer than 20 pharmacophores in the ensemble would be considered as synthetic candidates because they are predicted to be less likely to be transported by P-gp.

Results and Discussion. As demonstrated in Figure 1 and further summarized in Table 1, the model correctly classifies 80% of the compounds as substrates or nonsubstrates. The false negative rate of incorrectly classifying nonsubstrates as substrates is very low with 96% of the nonsubstrates correctly identified. The false positive rate is 36%, as 27 of the 76 P-gp substrates are incorrectly classified as nonsubstrates. The error bars in Figure 1 represent the results of 10 trials of leave-one-out cross-validation on the training set, for which the average pairwise pharmacophore overlap (identity) in each ensemble is 82.5%.

We further analyzed the performance and generalization of this computational filter by applying it to a diverse test set, composed of 32 P-gp substrates and 19 nonsubstrates. Using the threshold of matching at least 20 of the 100 pharmacophores of the ensemble, 53% of the P-gp substrates were correctly identified by the computational filter, while only 21% of the nonsubstrates were incorrectly predicted to be P-gp substrates. To further assess whether the ensemble model described above generalizes well, a cross-validation approach was used in which 144 compounds from the full data set were randomly selected for training in each of 10 trials. The performance plot mimics that shown in Figure 1, with average classifications as substrate for the “training set” substrates (61%) and nonsubstrates (6%), and “hold-out” substrates (53%) and nonsubstrates (19%), comparable to that reported in Table 1.

The overall prediction success for the ensemble performance on the hold-out set is 63%, and the success rate is 79% for the nonsubstrates. Greater success with the classification of nonsubstrates than P-gp compounds overall was also reported by Stouch and co-workers, who developed a QSAR model for P-gp substrate activity.[24] On the basis of the performance with test and training sets, the model could be used to design a library with fewer compounds that are P-gp liabilities by filtering a virtual library to remove 50−60% of the compounds likely to be P-gp substrates, without significant loss of likely nonsubstrates.

While four-point pharmacophores are typically the dominant element of our ensemble models, three-point pharmacophores also represent a large component of the top 100 most informative pharmacophores presumably because of the higher diversity present among P-gp substrates. The ensemble contains 53 four-point pharmacophores, 39 three-point pharmacophores, and 8 two-point pharmacophores. The informative pharmacophores contain combinations of four of the six possible feature types: hydrogen bond acceptors, hydrogen bond donors, hydrophobic groups, and aromatic rings. As also reported by others,[12,14] the hydrogen bond acceptor or electron donor group is observed to be an important chemical feature associated with P-gp substrate activity. Eighty-eight of the 100 pharmacophores composing the ensemble contain at least one hydrogen bond acceptor feature. Furthermore, 22 pharmacophores contain two acceptors and nine pharmacophores contain three hydrogen bond acceptors. Contained in these pharmacophores are examples of the type I and type II recognition units composed of two hydrogen bond acceptors separated by a distances of 2.5 or 4.6 Å described by Seelig[14] (Figure 2).

A hydrogen bond donor is present in 97 of the 100 pharmacophores composing the ensemble model, including the pharmacophore with the highest information content (Figure 2c). While chemical features that may function as a hydrogen-bond donor such as a basic nitrogen have previously been reported to be associated with P-gp activity,[10] the predominance of hydrogen bond donors in the most informative pharmacophores is in contrast to several other studies[13,14] that suggest that the hydrogen bonding strength is due to the presence of hydrogen bond acceptors. A study of pesticides by Bain et al. suggested that P-gp transport substrates have high molecular weights, lower log K_{ow} values, and greater hydrogen-bonding potential due to hydrogen

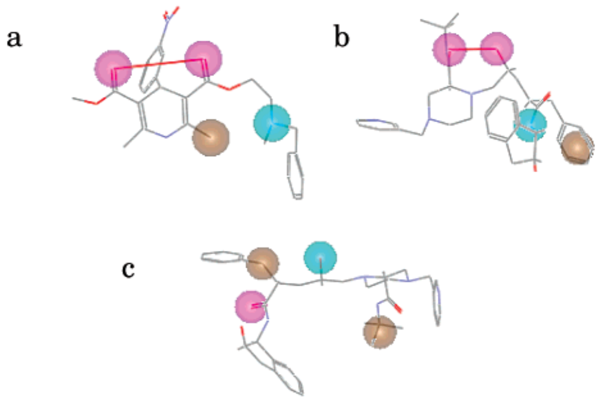

Figure 2. Four-point pharmacophores from the ensemble model, mapped onto conformations of nicardipine (a) and indinavir (b and c). The pharmacophore in part a contains two hydrogen bond acceptors separated by ~4.8 Å (red line), consistent with the Seelig's type II pattern definition. The pharmacophore shown in part b contains an example of a type I pattern, two hydrogen bond acceptors separated by ~2.5 Å (red line). The four-point pharmacophore mapped in part c represents the top scoring pharmacophore in the ensemble. The features of the pharmacophores are colored: hydrogen bond acceptor (magenta), hydrogen bond donor (cyan), and hydrophobe (brown).

bond donors and not acceptors.[25] In the model presented here, half of the pharmacophores (50 of 100) in the ensemble contain the feature combination: hydrogen bond acceptor, donor, and one or two hydrophobic groups. These results support the importance of both hydrogen bond donors and hydrogen bond acceptors in substrate binding to P-gp.

Conclusion. In this Letter we describe the construction and validation of a computational model for recognizing substrates for P-gp. The model consists of an ensemble of 100 two-, three-, and four-point pharmacophores that together are able to capture the various chemotypes that may interact via multiple binding sites and binding modes with P-gp. Cross-validation of the training set and filtering of a hold-out set of compounds that are chemically dissimilar to the training set demonstrated that the model correctly classifies 50−60% of the P-gp substrates and greater than 80% of the nonsubstrates. We anticipate that the computational model described in this Letter may be applied rapidly and routinely to filter likely P-gp substrates from virtual libraries.

Acknowledgment. The authors thank Dr. Susan Bates of the National Institutes of Health for kindly providing data for additional nonsubstrates.

Supporting Information Available: Three tables with the 100 pharmacophores of the ensemble model for P-gp activity and the substrates and nonsubstrates that comprise the data set. This information is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Di Pietro, A.; Dayan, G.; Conseil, G.; Steinfels, E.; Krell, T.; Trompier, D.; Baubichon-Cortay, H.; Jault, J. P-glycoprotein-mediated resistance to chemotherapy in cancer cells: using recombinant cytosolic domains to establish structure−function relationships. *Braz. J. Med. Biol. Res.* **1999**, *32*, 925−939.

(2) Putman, M.; van Veen, H. W.; Konings, W. N. Molecular properties of bacterial multidrug transporters. *Microbiol. Mol. Biol. Rev.* **2000**, *64*, 672−693.

(3) Delph, Y. *P-Glycoprotein: a tangled web waiting to be unraveled*; http://www.aidsinfonyc.org/tag/science/pgp.html, 2000.

(4) Kim, R. B.; Fromm, M. F.; Wandel, C.; Leake, B.; Wood, A. J. J.; Roden, D. M.; Wilkinson, G. R. The drug transporter P-glycoprotein limits oral absorption and brain entry of HIV-1 protease inhibitors. *J. Clin. Invest.* **1998**, *101*, 289−294.

(5) Klopman, G.; Shi, L. M.; Ramu, A. Quantitative structure−activity relationship of multidrug resistance reversal agents. *Mol. Pharmacol.* **1997**, *52*, 323−334.

(6) Bakken, G. A.; Jurs, P. C. Classification of multidrug-resistance reversal agents using structure-based descriptors and linear discriminant analysis. *J. Med. Chem.* **2000**, *43*, 4534−4541.

(7) Rosenberg, M. F.; Callaghan, R.; Ford, R. C.; Higgins, C. F. Structure of the multidrug resistance P-glycoprotein to 2.5 nm resolution determined by electron microscopy and image analysis. *J. Biol. Chem.* **1997**, *272*, 10685−10694.

(8) Chang, G.; Roth, C. B. Structure of MsbA from *E. coli*: A homology of the multidrug resistance ATP binding cassette (ABC) transporters. *Science* **2001**, *293*, 1793−1800.

(9) Higgins, C. F.; Linton, K. J. The xyz of ABC transporters. *Science* **2001**, *293*, 1782−1784.

(10) Zamora, J. M.; Pearce, H. L.; Beck, W. T. Physical-chemical properties shared by compounds that modulate multidrug resistance in human leukemic cells. *Mol. Pharmacol.* **1988**, *33*, 454−462.

(11) 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.

(12) 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.

(13) 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.

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

(15) Compounds from the Seelig 1998 dataset classified as “borderline substrates” as well as α-factor pheromone, FK506, the polymer Triton X-100, SDB-ethylenediamine, and methybenzolreserpate were not included in this study.

(16) Schuetz, E. G.; Beck, W. T.; Schuetz, J. D. Modulators and substrates of P-glycoprotein and cytochrome P4503A coordinately up-regulate these proteins in human colon carcinoma cells. *Mol. Pharmacol.* **1996**, *49*, 311−318.

(17) *Daylight Chemical Information Software*; Daylight Chemical Information, Inc.: Irvine, CA, info@daylight.com.

(18) Bradley, E. K.; Beroza, P.; Penzotti, J. E.; Grootenhuis, P. D. J.; Spellmeyer, D. C.; Miller, J. L. A rapid computational method for lead evolution: description and application to α_1-adrenergic antagonists. *J. Med. Chem.* **2000**, *43*, 2770−2774.

(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) Mason, J. S.; Cheney, D. L. Ligand−receptor 3-D similarity studies using multiple 4-point pharmacophores. *Pac. Symp. Biocomput.* **1999**, 456−467.

(21) Teig, S. L.; Smellie, A. S. Combichem, I. Method and apparatus for conformationally analyzing molecular fragments. WO9859306, 1998.

(22) Smellie, A. S.; Stanton, R. V.; Henne, R. M.; Teig, S. L. Conformational analysis by intersection: CONAN. *J. Comput. Chem.* Submitted.

(23) Srinivasan, J.; Castellino, A.; Bradley, E. K.; Eksterowicz, J. E.; Grootenhuis, P. D. J.; Putta, S.; Stanton, R. V. Evaluation of a novel shape-based computational filter for lead evolution: Application to thrombin inhibitors. *J. Med. Chem.* Submitted.

(24) Stouch, T. R.; Gudmundson, O.; Ge, S. E. Prediction of PGP transporter activity using calculated molecular properties. 221st National Meeting of the American Chemical Society, 2001. *Abstr. Pap. − Am. Chem. Soc.* **2001**, *BTEC-037*.

(25) Bain, L. J.; McLachlan, J. B.; LeBlanc, G. A. Structure−activity relationships for xenobiotic transport substrates and inhibitory ligands of P-glycoprotein. *Environ. Health Perspect.* **1997**, *105*, 812−818.

JM0255062