Classification of Substrates and Inhibitors of P-Glycoprotein Using Unsupervised Machine Learning Approach

Yong-Hua Wang,[†] Yan Li,[‡] Sheng-Li Yang,[†] and Ling Yang*[,†]

Lab of Pharmaceutical Resource Discovery, Dalian Institute of Chemical Physics, Graduate School of the Chinese Academy of Sciences, #457 Zhongshan Road, Dalian 116023, China, and School of Chemical Engineering, Dalian University of Technology, #158 Zhongshan Road, Dalian 116012, China

Received February 4, 2005

P-glycoprotein (P-gp), a drug efflux pump, affects the bioavailability of therapeutic drugs and plays a potentially important role in clinical drug−drug interactions. Classification of candidate drugs as substrates or inhibitors of the carrier protein is of crucial importance in drug development. Accurate classification is difficult to achieve due to two major factors: i. The extreme diversity of substrates and the presence of multiple binding sites complicate the understanding of the mechanisms behind and hinder the development of a true, conclusive quantitative structure−activity relationship (QSAR) for P-gp substrates. ii. Both inhibitors and substrates interact with the same binding site of P-gp, as a result, it is not surprising that both share many common structural features. In this work, an unsupervised machine learning approach based on the Kohonen self-organizing maps (SOM) was explored, which incorporated a predefined set of physicochemical descriptors encoding the key molecular properties capable of discerning a substrate from an inhibitor. The SOM model can discriminate between substrates and inhibitors with an average accuracy of 82.3%. The current results show that the SOM-based method provides a potential in silico model for virtual screening.

INTRODUCTION

P-glycoprotein (P-gp), a transmembrane protein, has been discovered in various resistant tumor cells as well as many normal tissues.[1,2] P-gp actively transports a wide variety of anticancer drugs out of the cell in an energy-dependent manner, resulting in an obstacle in chemotherapeutic treatment of cancer. An increased expression of P-gp is associated with multidrug resistance (MDR); therefore, many efforts are directed to find out whether a compound is pumped out by the carrier protein or has P-gp MDR reversal effects.

P-gp confers resistance to a large number of xenobiotics, such as HIV protease inhibitors, steroids, immunaosuppressants, calcium cannel blockers, dopamine antagonists, peptides, phospholipids, and cholesterol.[2] It is noteworthy to mention that the P-gp substrates are chemically and structurally different compounds. Recent researches demonstrated there are different binding sites at P-gp to bind substrates,[3] although the type and number of the binding sites is still a question of debate. The structural and functional variety of P-gp substrates raises the question how P-gp broad substrate specificity can be explained, which also complicates understanding and hinders developing a true, conclusive quantitative structure−activity relationship (QSAR) for P-gp substrates.[4]

In the last years a number of studies have been undertaken to develop MDR reversing compounds with potential clinical significance. According to their pharmacological and chemical classes, the P-gp MDR reversals can be divided into many classes that include analeptics, antidepressants, antimalarials, calcium channel blockers, steroid hormones, and many others.[5,6] Although the P-gp inhibitors, similar to the P-gp substrates, cover a large chemical and structural diversity, basically the mechanisms of them to reverse P-gp MDR possibly involve three aspects, i.e., (a) by blocking drug binding site, or (b) by interfering ATP hydrolysis, or (c) by altering integrity of cell membrane lipids.[4] On one hand, any reasonable (Q)SAR analysis relies on a real or putative presentation about the interaction mechanism. On the other hand, up to now those compounds that have been clearly demonstrated to be related with the above mechanisms are not enough, and the interaction of different molecules with P-gp involves diverse complex mechanisms. All these become barriers to build a general QSAR model for the whole P-gp inhibitors.

Intuitively, P-gp substrate and inhibitor are different and should belong to two different classes of molecules. Theoretically, by a careful selection of relevant structural and biological data, the two classes of compounds should be clearly distinguished from each other. However, this question is not so simple. It has been suggested that most inhibitors were supposed to work by blocking P-gp substrate binding sites.[4] Therefore, it is not surprising that these compounds should share multiple similar or overlapping chemical specificities, resulting in difficulties to differentiate them. Questions become even more complicated considering the fact that some inhibitors (verapamil, diltiazem, etc.) bind to the protein and are transported by it, while others (progesterone, etc.) although binding to P-gp cannot be effluxed.[6] These complex factors make it more difficult to decide how a compound interacts with P-gp and what mechanisms are involved. However, due to the pharmacological significance of P-gp, identification of potential P-gp substrates and

* Corresponding author e-mail: yling@dicp.ac.cn.
[†] Graduate School of the Chinese Academy of Sciences.
[‡] Dalian University of Technology.

10.1021/ci050041k CCC: $30.25 © 2005 American Chemical Society
Published on Web 04/20/2005

inhibitors is of importance, which is of advantage not only to the development of MDR reversals but also to eliminating drug candidates that are P-gp substrates. Therefore, great efforts have been made to determine whether a compound is a P-gp substrate or inhibitor.[6] In silico methods as potentially facile and economic alternative to in vitro methods are now emerging. The use of computational models over their traditional counterparts is preferred for many reasons such as ease of use, speed, and relatively low cost.

To date, the atomic resolution structure of this transmembrane protein is still unavailable, and little is known about the 3D-structure.[7] Efforts have been primarily directed at the development of quantitative structure−activity relationships (QSARs) models[5,7] and identification of pharmacophoric features for prediction of P-gp substrates[5,7,8] or inhibitors.[8,9] The multiple-pharmacophore model presented by Penzotti[10] and co-workers showed promising predictivity of P-gp substrates, achieving a prediction accuracy of 63% for a set of 195 compounds. The mechanism-based 3D-phamacophore models are valuable for identifying the molecular features required for P-gp substrates. Currently, it is still desirable to develop methods that extend the prediction range beyond those agents covered by known pharmacophore models in virtual screening.[11] Moreover, to our knowledge, an in silico model to classify P-gp substrates and inhibitors involving a large amount of molecules is still unavailable.

Artificial neural networks (ANNs) have been widely used in QSAR studies. In this study, an unsupervised algorithm, the Kohonen self-organizing map (SOM)[12] artificial neural network was applied. The SOM has shown promising capability for solving a number of biological classification problems.[13] The reason for choosing this algorithm in this investigation is that the SOM is particularly efficient for systems with multiple mechanisms such as P-gp. For comparison to the unsupervised approach, a supervised back-propagation neural network (BPNN) model was also established. Special attention is paid to build an effective SOM model for classification of potential P-gp substrates and inhibitors, which will be of value in the early stage of drug discovery.

MATERIALS AND METHODS

Data Sets. P-gp substrates were collected from experimental literatures such that each compound has been either described as being transported by P-gp or reported to induce the overexpression of P-gp which directly contributes to MDR. P-gp inhibitors were also collected from the literature reported to reverse MDR, which covered a wide range of structurally diverse compounds. Compounds shared by the both sets such as verapamil, an inhibitor that is also transported by P-gp, called overlapping compounds in this paper, were also included in the data set. In this way, a data set of 206 chemicals including 96 substrates, 78 inhibitors, and 32 overlapping compounds was compiled in this paper (Table 1). These compounds were used to build two independent classification models: the unsupervised SOM and supervised BPNN models. When building the SOM model, all molecules were introduced. Whereas, when building the BPNN model, all overlapping compounds were excluded for reasons discussed in section 3.4. In addition, to avoid overfitting and to improve generalization of the BPNN, we randomly took one-fourth of the 206 compounds as a cross-validation set, one-fourth as a test set, and the remainder as a training set.

Calculation of Molecular Descriptors. Construction of the QSAR models depends on the generation of molecular descriptors. By simply using various molecular modeling tools, it is possible to calculate thousands of these descriptors directly from the structure of any particular molecule. The Molconn-Z program (in the SYBYL software package) is a useful tool to calculate the molecular connectivity indices and electrotopological state (E-state) descriptors. Molconn-Z extends the descriptor set produced by its predecessor Molconn-X, which calculates E-state hydrogen and bond-type descriptors and provides 462 descriptors. Molconn-Z calculates additional 287 descriptors. These structural variables include the following: the molecular connectivity Chi indices, ${}^{m}\chi_{t}$ and ${}^{m}\chi_{t}^{v}$; Kappa shape indices, ${}^{m}k$ and ${}^{m}k_{\alpha}$; E-state indices, ES_i; hydrogen E-state indices, HES_i; atom type and bond type E-state indices; and topological equivalence indices and total topological indices.

Although this program calculates more than 700 molecular descriptors, for the present example, we selected 'standard Molconn-Z descriptors' that satisfied two requirements: (i) they were physically intuitive and directly related to gross structural and/or molecular biophysical properties, and (ii) they were a relatively small number of molecular descriptors but most useful in QSAR modeling (Molconn-Z manual), including 248 descriptors. The rationale for this decision was that the aim of the present study was to develop a reliable QSAR model but using as few as possible indices.

To reduce the descriptor space as well as to find more informative and understandable molecular descriptors, a feature selection method, stepwise discriminant analysis (SDA),[14,15] was used. Eventually, 11 of the most relevant descriptors (Table 2) were applied as input indices in neural network experiments. Among these descriptors, 7 were used in the 6 × 8 SOM and BPNN, and 4 were used in the 4 × 6 SOM.

Neural Network Modeling. Self-organizing maps are a special kind of neural network that can be used for clustering, visualization, and abstraction tasks. In this work, SOM was used to classify and visualize P-gp substrates and inhibitors. We performed the SOM clustering by two steps. In the first step, for the studied molecules a 6 × 8 node architecture was chosen for the model to provide sufficient distribution space, with 7 input descriptors for the network. Subsequently, those compounds that have not been clearly classified in the first step were further evaluated by using another 4 × 6 Kohonen SOM with 4 input descriptors. Figure 1 depicts the flowchart of the modeling procedure. The nodes of SOM were arranged in a rectangular grid, with the "bubble" function used as radial adjustment function. The training parameters were as follows: the starting adjustment radius for the training runs 0.1 and the decay factor 0.001.

To build a back-propagation feed-forward neural network, a three-layer fully connected network with the Levenberg−Marquardt training algorithm[16] was adopted. The optimum 10 neurons were selected in the hidden layer, with one neuron in the output layer. We assigned score values for the P-gp substrates of 1 and for inhibitors of 2. The training was performed with 15 iterations. An internally developed C language program was used for the QSAR analysis.

Table 1. Substrates (Class S), Inhibitors (Class I), and Overlapping Compounds (Class O) of P-Glycoprotein in the Data Set

no.	chemical	class	no.	chemical	class
S1	clarithromycin	S	S49	monensin	S
S2	actinomycind	S	S50	morphine-6-glucuronide	S
S3	adriamycin	S	S51	morphine	S
S4	aldosterone	S	S52	domperidone	S
S5	alpha_methyldigoxin	S	S53	navelbine	S
S6	amoxicillin	S	S54	nortriptyline	S
S7	amprenavir	S	S55	NSC328426	S
S8	beta_acetyldigoxin	S	S56	NSC339281	S
S9	bisantrene	S	S57	NSC359449	S
S10	bunitrolol	S	S58	NSC630176	S
S11	NSC359449	S	S59	NSC640085	S
S12	carbamazepine	S	S60	NSC66490	S
S13	catharanthine	S	S61	ondansetron	S
S14	ciprofloxacin	S	S62	pafenolol	S
S15	colchicine	S	S63	phenoxazine	S
S16	cortisol	S	S64	phenytoin	S
S17	cyclosporin A	S	S65	podophyllotoxin	S
S18	daunorubicin	S	S66	prazosin	S
S19	debrisoquine	S	S67	prednisolone	S
S20	dexamethasone	S	S68	propiconazole	S
S21	dibucaine	S	S69	puromycin	S
S22	digitoxin	S	S70	rapamycin	S
S23	digoxin	S	S71	rhodamine 123	S
S24	docetaxel	S	S72	rifampin	S
S25	eletriptan	S	S73	risperidone	S
S26	endosulfan	S	S74	sparfloxacin	S
S27	enoxacin	S	S75	tacrolimus	S
S28	epirubicin	S	S76	talinolol	S
S29	estradiol	S	S77	taxol	S
S30	etoposide	S	S78	teniposide	S
S31	farnesol	S	S79	topotecan	S
S32	fexofenadine	S	S80	traphenyl-phosphonium bromide	S
S33	Hoechst 33342	S	S81	trimethoprim	S
S34	hydrocortisone	S	S82	vincristine	S
S35	hydroxyrubicin	S	S83	vindesine	S
S36	idopine	S	S84	vindoline	S
S37	indinavir	S	S85	vinorelbine	S
S38	irinotecan	S	S86	doxorubicin	S
S39	K02	S	S87	paclitaxel	S
S40	levofloxacin	S	S88	clotrimazole	S
S41	loratidine	S	S89	dexamethasone	S
S42	L_dopa	S	S90	efavirenz	S
S43	melphalan	S	S91	dellavirdine	S
S44	methylprednisolone	S	S92	nefazodone	S
S45	mithramycin	S	S93	nevirapine	S
S46	mitomycin	S	S94	phenobarbital	S
S47	mitoxantrone	S	S95	rifampin	S
S48	modipine	S	S96	trazodone	S
I1	apigenin	I	I40	lidocaine	I
I2	astemizole	I	I41	loratadine	I
I3	atorvastatin	I	I42	maprotiline	I
I4	atovaquone	I	I43	mefloquine	I
I5	azacyclonol	I	I44	methadone	I
I6	azelastine	I	I45	mifepristone	I
I7	benidipine	I	I46	mitomycin_c	I
I8	broussochalconea	I	I47	nefazodone	I
I9	caffeine	I	I48	NSC665333	I
I10	carvedilol	I	I49	NSC667739	I
I11	chalcone	I	I50	NSC676590	I
I12	chlorzoxazone	I	I51	NSC676591	I
I13	chrysin	I	I52	NSC676597	I
I14	clarithromycin	I	I53	NSC676599	I
I15	cyclosporine	I	I54	NSC676600	I
I16	cyproheptadine	I	I55	NSC68075	I
I17	dehydrosilybin	I	I56	ofloxacin	I
I18	desipramine	I	I57	omeprazole	I
I19	dexrazoxane	I	I58	pantoprazole	I
I20	8-DMA-chrysin	I	I59	piperine	I
I21	8-DMA-galangin	I	I60	8-prenyl-chrysin	I
I22	8-DMA-kaempferide	I	I61	8-prenyl-dehydrosilybin	
I23	doxazosin	I	I62	8-prenyl-galangin	I
I24	erythromycin	I	I63	probenecid	I
I25	felodipine	I	I64	progesterone	I
I26	fentanyl	I	I65	promethazine	I

Table 1 (Continued)

no.	chemical	class	no.	chemical	class
I27	flavone	I	I66	propafenone	I
I28	fluoxetine	I	I67	quercetin	I
I29	fluphenazine	I	I68	quinine	I
I30	fluvoxamine	I	I69	rifluoperazine	I
I31	galangin	I	I70	sertraline	I
I32	8-geranyl-chrysin	I	I71	simvastatin	I
I33	8-geranyl-dehydrosilybin	I	I72	spironolactone	I
I34	imipramine	I	I73	sufentanil	I
I35	7-*O*-isopropylchrysin	I	I74	tamoxifen	I
I36	ivermectin	I	I75	terfenadine	I
I37	josamycin	I	I76	thioridazine	I
I38	ketoconazole	I	I77	trimipramine	I
I39	lansoprazole	I	I78	valinomycin	I
O1	amitriptyline	O	O17	diltiazem	O
O2	cepharanthine	O	O18	FK506	O
O3	cimetidine	O	O19	haloperidol	O
O4	dipyridamole	O	O20	itraconazole	O
O5	loperamide	O	O21	levothyroxin	O
O6	lovastatin	O	O22	midazolam	O
O7	nelfinavir	O	O23	NSC668360	O
O8	pimozide	O	O24	NSC676605	O
O9	ritonavir	O	O25	paroxetine	O
O10	saquinavir	O	O26	propranolol	O
O11	spiperone	O	O27	quinidine	O
O12	alfentanil	O	O28	reserpine	O
O13	amiloride	O	O29	verapamil	O
O14	bromocriptine	O	O30	vinblastine	O
O15	chloroquine	O	O31	amiodarone	O
O16	chlorpromazine	O	O32	emetine	O

Table 2. Molecular Descriptors Used in Present Work[a]

descriptor	definition[c]
nHaaCH	number of atom group :CH:
nsNH2	number of atom group −NH2
SaaN, SdsCH	electrotopological state (E-state) index values for atom types
Xvch6	connectivity valence chain indices
Xvp3, Xvp8[b]	connectivity valence path indices
SHBint3	E-state descriptors of potential internal H-bond strength
ndsCH[b]	number of atom group =CH-
ndssC[b]	number of atom group =CH<
SHaaCH[b]	:CH atom group

[a] Parameters without superscript used in 6× 8 SOM. [b] Parameters used in 4× 6 SOM. [c] All atom type (group) ID definitions, such as nHaaCH, are made by Molconn-Z program. The atom type symbol consists of symbols for the bonds in the group and symbols for the elements in the group.

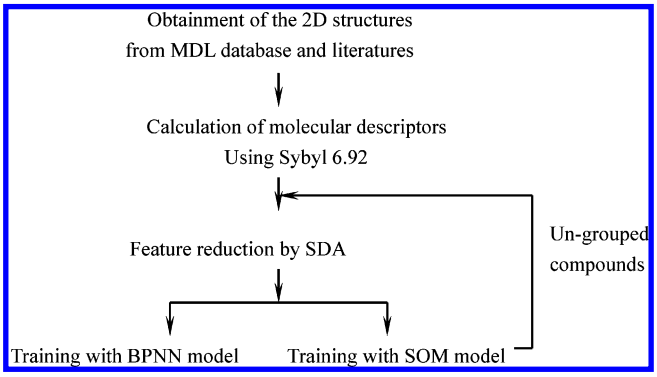

Figure 1. Flowchart of the model building.

RESULTS AND DISCUSSION

Molecular Descriptors. Before generating QSAR models, it is often necessary to perform a variable reduction process on the original set of descriptors. Basically, the objective of variable selection is 3-fold: (a) to provide faster and more cost-efficient predictors; (b) to provide a better understanding of the underlying process that generated the data; and (c) also the most important one is, to eliminate noise (uninformative descriptors) and prevent overfitting or chance correlations. Generally, the feature selection methods perform better than feature recombination methods. And the resulted parameters by feature selection are intuitively more understandable than those parameters by feature recombination methods, such as principal component analysis.[17] For this reason, SDA was used to choose a subset of features from the original feature set. A more detailed SDA method can be referred in Jennrich's paper.[14]

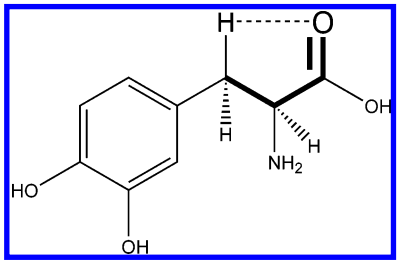

Figure 2. The internal hydrogen bond in L-dopa molecule (P-gp substrate). The dotted line represents the internal hydrogen bond.

The selected set of descriptors reveals the primary differences of the functional groups among compounds, such as *:CH:*, $-NH_2$, and connectivity indices including *Xvch6*, *Xvp3*, and *Xvp8*. E-state arises from the electronic environment of each atom due to its intrinsic electronic properties and the influence of other atoms in the molecule. In this work, the two E-state atom type *SaaN* and *SdsCH* were selected. *SHBint3* is the sum of E-state products for potential internal hydrogen bonds, which considers the occurrence of internal hydrogen bonding when the donor and acceptor are

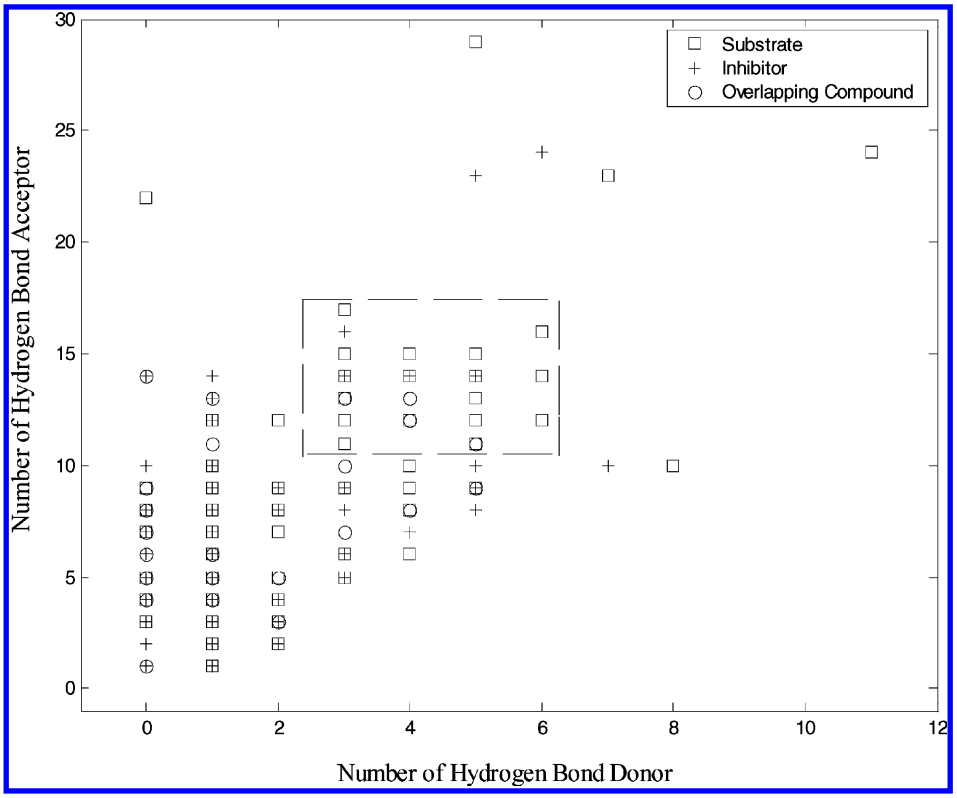

Figure 3. Clustering of P-gp substrates, inhibitors, and overlapping compounds based on number of molecular hydrogen bond acceptors and donors. Plus, square and circle indicate P-gp inhibitor, substrate, and overlapping compounds, respectively.

separated by 3 bonds along a path. Figure 2 depicts the internal hydrogen bond (H-bond) in a molecule. When analyzing the *SHBint3* descriptor, some new interesting facts come out. About 34% (33/96) of the substrate compounds have internal hydrogen bonds with an *SHBint3* mean value of 28.5. While, only 6.4% (5/78) of inhibitors have internal H-bonds with the mean value of 10.14. As for the overlapping compounds, 25% (8/32) have internal H-bonds and the mean value is 21.12. The comparison concludes that P-gp substrates are more likely to have the internal H-bonding than inhibitors. And to some extent, the overlapping compounds are more likely to show the characteristics of the internal H-bond.

The *SHBint3* has describes the H-bond in a molecule in spatial distance. Whereas the role of the number of H-bonds (including nHBa, the number of H-bond acceptors, and nHBd, the number of H-bond donors) in affecting compounds with P-gp activities still needs to be concerned, since they are very important molecular chemical features for P-gp substrates and inhibitors.[6] Interestingly, the two descriptors, i.e., *nHBa* and *nHBd* were not selected out by the SDA procedure to distinguish between the two classes of compounds in the present work. Based on a relatively large data set collected, the mean values of *nHBa* and *nHBd* for the three sets were calculated, resulting in 2.7 and 9.1 for substrates, 1.7 and 6.7 for inhibitors, and 1.6 and 7.7 for overlapping compounds, respectively. It seems that P-gp inhibitors and overlapping compounds are quite similar to each other, while they are both distinct from substrates in the number of H-bonds. Substrates tend to have a greater number of H-bonds than inhibitors and overlapping compounds. However, a clustering in terms of the two descriptors for all the data sets may offer a plainer picture for the comparisons (Figure 3). Figure 3 illustrates that most of the molecules are heavily overlapped, and only a small amount of compounds appearing in the dotted square are distinguished from others such as P-gp substrates. The above results imply that conditions of H-bond acceptors or donors may not be the essential difference between substrates and inhibitors. Therefore, to solely use *nHBa* and *nHBd* to represent the characteristics of substrate and inhibitor may lead to improper conclusions.

The P-gp substrates and inhibitors are so structurally diverse that it is difficult to identify the common structural elements they share. Like a basic nitrogen can be outlined in several MDR reversals, but for a wide collection of inhibitors, compounds that possess basic nitrogen atoms in structures only account for a very small proportion.[6] Possibly, no conserved elements of molecular recognition can be found. However, using a combination of structural features such as the different atom types, particularly, the E-state values for atom type and internal hydrogen bonding, the P-gp substrates and inhibitors may be proper distinguished.

Kohonen Self-Organizing Map. The SOM can be efficiently used in data visualization due to its ability to approximate the probability density of input data and to represent them in two dimensions. The unified distance matrix (U-matrix) method by Ultsch[18] was used to visualize the structure of the SOM. First, the matrix of distances (U-matrix) between the weight vectors of adjacent units of a two-dimensional map is formed. Second, the matrix is visualized as a gray-level image.[19] The lighter the color between two map units, the smaller the relative distance between them. We used a Kohonen net with a 2D organization of the network units (neurons).

In the present work, the projection onto the Kohonen map was conducted twice. The first one was a 6×8 Kohonen map projected by the whole data set of 206 molecules (Figure

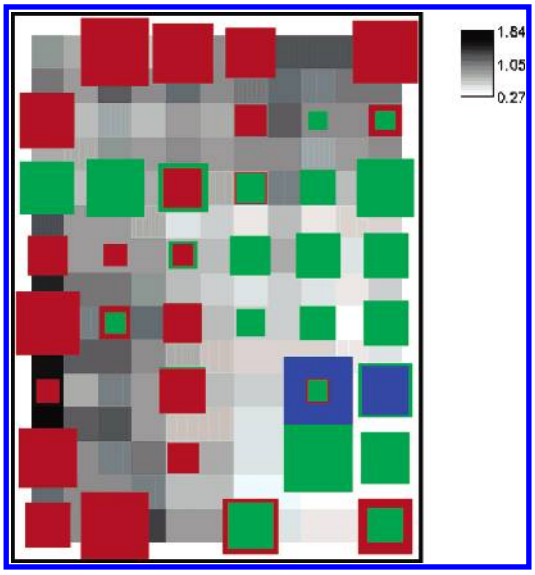

Figure 4. Visualization of the 6 × 8 SOM of P-gp substrates and inhibitors using U-matrix. Every hexagon corresponds to a map neuron in the same position. The graph shows the distribution of P-gp substrates, inhibitors, and overlapping compounds on the map unit. The red, green, and blue colors represent substrates, inhibitors, and overlapping compounds, respectively. The red-green mixed-up color represents the compounds that no specific assignment can be made. The blue-green mixed-up color represents the overlapping compounds.

4). The second one is a 4 × 6 SOM responsible for the unclassified compounds in the first step (Figure 5). Table 3 has summarized the statistical results of the two clustering.

Figure 4 depicts that P-gp substrates (in red color) distribute irregularly throughout the map in several subgroups, whereas the inhibitors (in green color) mainly concentrate on the middle-right side of the map. The places occupied by the substrates are relatively large and scattered compared to those of the inhibitors, which reflect the broad substrate specificity of P-gp. Conversely, the region of the inhibitors is relatively focused, which, to some extent, shows that P-gp inhibitors share more chemical and biological similarities with each other. We suggest that the physicochemical properties of a molecule falling into the red or the green regions of the Kohonen map are consistent with the molecule's ability to become a P-gp substrate or inhibitor.

To have a clear analysis of this issue, one example was offered. For instance, the first left-top neuron was occupied by nine substrates (shown in the Supporting Information, Figure 1). When the descriptors of the nine substrates were investigated, three descriptors, *SdsCH*, *Xvch*, and *Xvp3*, have similar values, and the other four descriptors, i.e., *nHaaCH*, *nsNH2*, *SaaN*, and *nHssNH*, have the same value of zero (data not shown). This example clearly shows the power of the SOM model in predicting the biological activity of a novel compound. When an unknown possesses similar values with the example in the set of descriptors, it will appear on the left-top neuron on the map. Therefore, we can directly imply that this compound should have a similar property with the nine substrates on the site.

Several overlaps are observed as shown in mixed-up colors (the red-green and the green-blue) in Figure 4. The phenomenon indicates that some of the compounds do possess the properties of both substrate and inhibitor, which is consistent with experimental reports. Since the overlapping

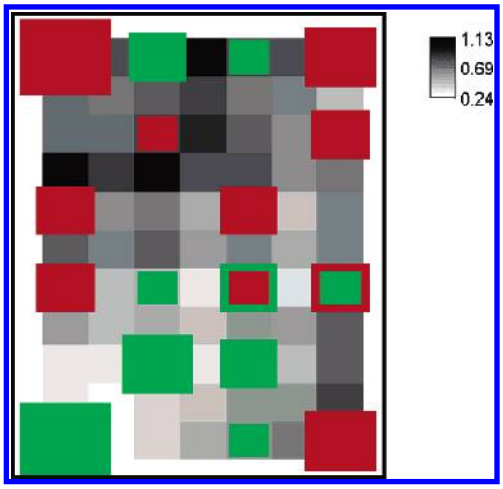

Figure 5. Visualization of the 4 × 6 SOM of P-gp substrates and inhibitors using the U-matrix. The hexagon in a certain position corresponds to the same map neuron. The red and green colors represent substrates and inhibitors, respectively. The mixed-up color for the compounds represents that no specific assignment can be made.

compounds belong to both substrate and inhibitor, they seem to be more special compared to those pure P-gp substrates and inhibitors. Originally we supposed that all the overlapping compounds should cluster in a third concentrated area that is separate from both substrate and inhibitor area. Whereas, from this projection, we found that only compounds in the green-blue neurons have the potential to be overlapping compounds (see the Supporting Information, Figure 1), and these neurons only covered several overlapping compounds, most of which were quite scattered on the map. Two possible reasons for this phenomenon are currently discussed. (i) Although these compounds are seemingly biologically similar, their real interaction mechanisms with P-gp may be quite different. As we know, verapamil and vinblastine have been recognized as both P-gp substrates and inhibitors.[6] However, Ayesh et al.[20] suggested that, to one of the two binding sites at P-gp, vinblastine, mefloquine, and tamoxifen bind preferentially; to the other, verapamil and dipyridamole bind. (ii) This proposed set of descriptors might not be capable of separating overlapping compounds from substrates and inhibitors. Due to the difficulty in obtaining sufficient data for overlapping compounds, it is still hard to develop a proper model to characterize P-gp overlapping compounds. Thus, further studies are still required in both experiment and theory.

On the map, a certain number of compounds (24% and 19% for substrates and inhibitors, respectively, a total of 43 molecules) fell into the area (the red-green mixed-up area), for which no specific assignment could be made. Therefore, additional criteria are needed for assessing the potential of these compounds to be P-gp substrates or inhibitors. The 43 ungrouped compounds were taken into a further classification step. They were treated with the same procedure as that of the initial 206 compounds, using a 4 × 6 SOM and four input parameters selected, i.e., *Xvp8*, *ndsCH*, *ndssC*, and *SHaaCH*. These four descriptors represent the types of atom group in a molecule and show the biggest differences in this set of compounds. The 4 × 6 SOM clustering result was shown in Figure 5, in which the red color is for substrates, and the green for inhibitors. The two groups are separated pretty well on the map.

Table 3. Classification Results from the Kohonen Neural Network Modeling

		predicted membership				
step	class	substrate	inhibitor	overlap	ungrouped compound	total
I	substrate	61(63.5%)	9(9.4%)	2(2.1%)	24(25%)	96 (100%)
	inhibitor	8(10.3%)	46(59%)	5(6.4%)	19(24.3%)	78 (100%)
	overlap	12(37.5%)	13(40.6%)	5(15.6%)	2(6.3%)	32 (100%)
II	substrate	17(70.8%)	2(8.3%)	0(0%)	5(20.9%)	24(100%)
	inhibitor	2(10.5%)	12(63.2%)	0(0%)	5(26.3%)	19(100%)
	overlap	1(50%)	1(50%)	0(0%)	0(0%)	2(100%)
sum	substrate	78(81.3%)	11(11.5%)	2(2.1%)	5(5.1%)	96 (100%)
	inhibitor	10(12.8%)	58(74.4%)	5(6.4%)	5(6.4%)	78 (100%)
	overlap	13(40.6%)	14(43.8%)	5(15.6%)	0((0%)	32 (100%)

Table 4. Fraction of Correct Classification of Compounds Using Supervised Back-Propagation Neural Network

randomization	compound category	training set (50%)	validation set (25%)	test set (25%)
rand 1	substrates	84.1%	46.7%	47.6%
	inhibitors	77.8%	30.5%	28.5%
rand 2	substrates	81.4%	45.8%	44.7%
	inhibitors	75.1%	29.0%	28.1%
rand 3	substrates	88.1%	44.0%	47.4%
	inhibitors	72.9%	30.1%	30.4%
average	substrates	84.5%	45.5%	46.6%
	inhibitors	75.3%	29.9%	29.0%

In summation of the two steps, 11 substrates were misclassified as inhibitors and 5 substrates were ungrouped, which resulted in a classification accuracy of 83.3% for substrates. 10 inhibitors were misclassified and 5 were ungrouped, which resulted in a classification accuracy of 80.8% for inhibitors. The results were defined by the compounds' localization in the corresponding areas of the Kohonen map (Table 3).

The categories of compounds have distinctly different localizations on the Kohonen map, which will be helpful for predicting whether a compound is a P-gp substrate or inhibitor. Our observations reveal that the difference between human P-gp substrates and inhibitors can be described by a combination of specific physicochemical features. Using an unsupervised SOM neural network, P-gp substrates and inhibitors can be pretty well distinguished from each other.

Back-Propagation Neural Network. As a comparison to the SOM model, we believe it is useful to evaluate the ability of an alternative classification algorithm, the supervised learning, to discriminate these compounds using the same set of molecular descriptors. Such a comparison is particularly interesting as the back-propagation method was used in many applications of neural networks in chemistry. In this work, BPNN was attempted to model only the classification between substrates and inhibitors, in which the overlapping compounds were not included. Therefore, 174 compounds (96 substrates and 78 inhibitors) were used to derive the BPNN model. The networks were trained with the same 7 molecular descriptors as used in the 6×8 SOM model. Additional validation and test sets were conducted to avoid overtraining when building the models. Considering the borderline scoring value of 1.5, we found that the net average classified 46.6% of substrates and 29.0% of inhibitors (Table 4). Retraining using randomly selected multiple training, validation, and test sets caused no big difference in predictivity. The network stopped with 15 training epochs, and no significant overfitting was found (see the Supporting Information, Figure 3). The discriminative power of the trained network was moderate, as demonstrated by the distribution of the P-gp substrates and inhibitors shown in Table 4.

Comparison of Unsupervised SOM and Supervised BPNN. The choice between the supervised or the unsupervised approaches depends on the specific problem and available data. As discussed above, we are encountering a complex system with multiple mechanisms, thus it is hard to build the mechanisms or the intrinsic relationships between activities and molecular properties of P-gp substrates or inhibitors. Furthermore, due to limitation of information provided by the data in advance, it appears to be difficult to choose the supervised method. However, the unsupervised Kohonen self-organizing map is adequate to deal with this kind of problems as mentioned in the Introduction section.

In this work, the extreme diversity of P-gp substrates and inhibitors caused the moderate level of discrimination of supervised BPNN (Table 3). Such diversity resulted in several distinct clusters in the investigated property space corresponding to separate areas on the Kohonen map. The SOM neural network classified objects into more than two classes and not only into one out of several predefined ones. As a result, to roughly predetermine the substrate as one class and the inhibitor as another to train a supervised ANN will be a problem. This is also the reason the overlapping compounds were excluded in the BPNN modeling in this work, for they could not be simply considered as a third class. Therefore, the unsupervised approach is more suitable for the situation like this.

The self-organizing map is a neural network algorithm that is based on unsupervised learning. The unsupervised neural network method is more flexible due to its many possible outputs and has no limitations of the accurate ascription problem of a compound. In addition, SOMs can also be used to discover some underlying structures of the data. However, the kind of structure we are looking for is very different from, say, principal component analysis or vector quantization. SOM is a topology-preserving map because there is a topological structure imposed on the nodes in the network. The 2D-projection of the combined data set of all molecules onto the Kohonen map will help us to gain more understanding of the relationships between chemical structures and activities. Overall, our studies show that the unsupervised SOM provides with more accurate discrimination between the studied compound categories than the supervised learning BPNN model.

CONCLUSION

In this work, we set out to explore an in silico model for the classification of P-glycoprotein substrates and inhibitors. The model is based on an unsupervised Kohonen learning approach and a preselected set of molecular descriptors. Using this method, 83.3% of the substrates and 80.8% of the inhibitors have been classified correctly.

Meanwhile, the developed neural network model also proved to be an effective tool for visualization of P-gp substrates and inhibitors. The position of a compound on the Kohonen map will determine how it interacts with P-gp even in cases when no unambiguous rule-based prediction can be made.

Due to the multiple outputs and no limitation of the accurate ascription problem of a compound as well as the unique topology-preserving property the unsupervised SOM method exhibits superiority to the supervised back-propagation method. The direct comparison between the two methods reconfirms the capability of the SOM algorithm in making a classification between P-gp substrates and inhibitors. In a word, the SOM model allows for the development of an automated computational algorithm for early assessment of possible P-glycoprotein substrates and inhibitors with high accuracy, which will aid in further drug design.

ACKNOWLEDGMENT

The authors thank the 973 Program (2003CCA03400) and the 863 Program (2003AA223061) of Ministry of Science and Technology of China.

Supporting Information Available: Two figures with the labeled compounds projected on the SOM unit and one figure of the modeling of BPNN. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Yu, D. K. The contribution of P-glycoprotein to pharmacokinetic drug-drug interactions. *J. Clin. Pharmacol.* **1999**, *39*, 1203−1211.

(2) Safa, A. R. Identification and characterization of the binding sites of P-glycoprotein for multidrug resistance-related drugs and modulators. *Curr. Med. Chem. Anti-Canc. Agents*. **2004**, *4*, 1−17.

(3) 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.

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

(5) 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.

(6) Wiese, M.; Pajeva, I. K. Structure−activity relationships of multidrug resistance reversers. *Curr. Med. Chem.* **2001,** *8*, 685−713.

(7) Schmitt, L.; Tampe, R. Structure and mechanism of ABC transporters. *Curr. Opin. Struct. Biol.* **2002**, *12*, 754−760.

(8) Li, Y.; Wang, Y.; Yang, L.; Zhang S. W.; Liu C. H.; Yang S. L. Comparison of steroid substrates and inhibitors of P-glycoprotein by 3D-QSAR analysis. *J. Mol. Struct.* **2005**, *733*, 111−118.

(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. Three-dimensional quantitative structure−activity relationships of inhibitors of P-glycoprotein. *Mol. Pharmacol.* **2002**, *61*, 964−973.

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

(11) Xue, Y.; Yap, C. W.; Sun, L. Z.; Cao, Z. W.; Wang, J. F.; Chen, Y. Z. Prediction of P-glycoprotein substrates by a support vector machine approach. *J. Chem. Inf. Comput. Sci.* **2004**, Web Release Date: May 8.

(12) Kohonen, T. *Self-Organization and Associative Memory*, 2nd ed.; Springer-Verlag: Berlin, 1987.

(13) Korolev, D.; Balakin, K. V.; Nikolsky, Y.; Kirillov, E.; Ivanenkov, Y. A.; Savchuk, N. P.; Ivashchenko, A. A.; Nikolskaya, T. Modeling of human cytochrome p450-mediated drug metabolism using unsupervised machine learning approach. *J. Med. Chem.* **2003**, *46*, 3631−3643.

(14) Jennrich, R. I. *Stepwise discriminant analysis, Statistical Methods for Digital Computers*; Enslein, K., Ralston, A., Wilf, H. S., Eds.; John Wiley & Sons: New York, 1977; pp 77−95.

(15) Sookie, L.; Cho J. H.; Kim J. H.; Kim, K. R. Capillary electrophoretic profiling and pattern recognition analysis of urinary nucleosides from thyroid cancer patients. *Anal. Chim. Acta* **2003**, *486*, 171−182.

(16) Hagan, M.; Menhaj, M. Training feedforward networks with the Marquardt algorithm. *IEEE Trans. Neural Networks* **1994**, 989−993.

(17) Huang, K.; Velliste, M.; Murphy, R. F. Feature reduction for improved recognition of subcellular location patterns in fluorescence microscope images. *Proc. SPIE* **2003**, *4962*, 307−318.

(18) Ultsch, A.; Siemon, H. P. Kohonen's self-organizing feature maps for exploratory data analysis. *Proceedings of 1990 Int. Neural Network Conference (INNC'90)*; 1990; pp 305−308.

(19) Iivarinen, J.; Kohonen, T.; Kangas, J.; Kaski, S. Visualizing the clusters on the self-organizing map. In *Proceedings of the Conference on Artificial Intelligence Research in Finland*; Carlsson, C., Järvi, T., Reponen, T., Eds.; Finnish Artificial Intelligence Society: Helsinki, Finland, 1994; pp 122−126.

(20) Ayesh, S.; Shao, Y. M.; Stein, W. D. Cooperative, competitive and noncompetitive interactions between modulators of P-glycoprotein. *Biochim. Biophys. Acta* **1996**, *1316*, 8−18.

CI050041K