0026-895X/07/7101-158–168$20.00
MOLECULAR PHARMACOLOGY
Vol. 71, No. 1
27623/3158669
Mol Pharmacol 71:158–168, 2007
Printed in U.S.A.

In Silico Prediction of Pregnane X Receptor Activators by Machine Learning Approaches[S]

C. Y. Ung, H. Li, C. W. Yap, and Y. Z. Chen

Bioinformatics and Drug Design Group, Department of Pharmacy and Department of Computational Science, National University of Singapore, Singapore (H.L., C.W.Y., Y.Z.C.); and Department of Biochemistry, the Yong Loo Lin School of Medicine, National University of Singapore, Singapore (C.Y.U.)

Received June 6, 2006; accepted September 26, 2006

ABSTRACT

Pregnane X receptor (PXR) regulates drug metabolism and is involved in drug-drug interactions. Prediction of PXR activators is important for evaluating drug metabolism and toxicity. Computational pharmacophore and quantitative structure-activity relationship models have been developed for predicting PXR activators. Because of the structural diversity of PXR activators, more efforts are needed for exploring methods applicable to a broader spectrum of compounds. We explored three machine learning methods (MLMs) for predicting PXR activators, which were trained and tested by using significantly higher number of compounds, 128 PXR activators (98 human) and 77 PXR nonactivators, than those of previous studies. The recursive feature-selection method was used to select molecular descriptors relevant to PXR activator prediction, which are consistent with conclusions from other computational and structural studies. In a 10-fold cross-validation test, our MLM systems correctly predicted 81.2 to 84.0% of PXR activators, 80.8 to 85.0% of hPXR activators, 61.2 to 70.3% of PXR nonactivators, and 67.7 to 73.6% of hPXR nonactivators. Our systems also correctly predicted 73.3 to 86.7% of 15 newly published hPXR activators. MLMs seem to be useful for predicting PXR activators and for providing clues to physicochemical features of PXR activation.

Pregnane X receptor (PXR) is a nuclear receptor known to be activated by structurally diverse xenobiotics and endogenous compounds (Lehmann et al., 1998; Jones et al., 2000; Ekins, 2004). PXR plays important roles in the metabolism of xenobiotics and drug-drug interactions by regulating the expression of metabolizing enzymes such as cytochrome P450 enzymes (CYP3A4, CYP2B6, and CYP2C8/9), and glutathione-*S*-transferases (Kliewer et al., 2002). It also regulates the expression of important drug transporters such as P-glycoprotein and multidrug resistance proteins (Ekins, 2004; Xie et al., 2004). Therefore, drugs capable of activating PXR may have significant impact on their own metabolism, transport, and interaction with other drugs. Identification of PXR activators is important for analyzing metabolism and pharmacokinetic profiles of drug candidates and for detecting potential drug-drug interactions.

Most of the drug metabolism prediction efforts have been directed at the development of tools for predicting cytochrome P450 substrates and inhibitors (Ekins et al., 2000; Doniger et al., 2002). However, significantly fewer works have been devoted to the development of tools for identifying PXR activators. So far, experimental high-throughput screening assays have been used for detecting PXR binding ligands (Jones et al., 2000), and computational pharmacophore (Ekins and Erickson, 2002; Schuster and Langer, 2005) and quantitative structure and activity relationship (QSAR) (Jacobs, 2004) models have been developed for predicting PXR activators. Because of the importance of PXR in drug metabolism and drug-drug interactions, more efforts are needed to explore additional methods for predicting a broader spectrum of PXR activators than those covered by existing studies.

We explored machine learning methods (MLMs) for predicting PXR and human PXR (hPXR) activators. PXRs show a high amount of sequence diversity in their ligand-binding domains (Moore et al., 2002), resulting in marked differences in ligand selectivity of PXRs across species, which is likely to have evolutionary significance in cross-species difference in adaptation to toxic compounds (Krasowski et al., 2005). Some

Article, publication date, and citation information can be found at http://molpharm.aspetjournals.org.
doi:10.1124/mol.106.027623.
[S] The online version of this article (available at http://molpharm.aspetjournals.org) contains supplemental material.

ABBREVIATIONS: PXR, pregnane X receptor; MLM, machine learning method; QSAR, quantitative structure and activity relationship; h, human; SVM, support vector machine; PNN, probabilistic neural network; k-NN, k nearest neighbor; RFE, recursive feature elimination; DI, diversity index; CAS, Chemical Abstracts Service; SR12813, 4-[2,2-bis(diethoxyphosphoryl)ethenyl]-2,6-di-*tert*-butyl-phenol.

compounds are known to activate mouse but not human PXR and vice versa. Therefore, it is more relevant to develop prediction systems for hPXR activators. Nonetheless, prediction systems for PXR and hPXR activators were developed in this work for facilitating the search of broader spectrum of activators, particularly those of species frequently used in drug toxicity tests.

MLMs have been used for predicting compounds of different pharmacological classes (Doniger et al., 2002; Xue et al., 2004b; Yap and Chen, 2005). The most widely used MLMs in these studies are support vector machines (SVMs) (Burges, 1998), probabilistic neural network (PNN) (Specht, 1990), and k nearest neighbor (k-NN) (Johnson and Wichern, 1982). These methods have consistently exhibited good prediction performance for compounds of diverse structures. Moreover, a feature selection method can be incorporated into these methods for selecting molecular descriptors most relevant to the prediction of compounds with specific pharmacological property (Xue et al., 2004a,b; Li et al., 2005a,b).

PXR activators are structurally diverse partly because PXR ligand binding domain is highly flexible (Watkins et al., 2001). Nonetheless, certain common physicochemical characteristics can be found at the binding site. For instance, the binding site is largely hydrophobic but contains a few polar residues capable of both donating and accepting hydrogen bonds (Watkins et al., 2001). These and other distinguished binding-site features probably define the common structural and physicochemical properties of the compounds that can bind and activate PXR, which can be exploited by using MLMs to distinguish PXR activators and nonactivators. Several molecular descriptors of PXR activators have been used for deriving QSAR (Jacobs, 2004) and pharmacophore models (Ekins and Erickson, 2002; Schuster and Langer, 2005). It is likely that not all of the molecular descriptors related to PXR activation have been included in previous studies because of the limited coverage of compounds and the number of other relevant descriptors. Therefore, feature selection methods (Xue et al., 2004a,b; Li et al., 2005a,b) may be applied for finding additional molecular descriptors relevant to PXR activation. The use of a higher number of relevant molecular descriptors also serves to improve the performance of MLMs.

In this work, PXR and hPXR activator prediction systems were developed by using SVM, PNN, and k-NN, which were trained and tested by using a significantly higher number of compounds than those used in the previous studies. A comprehensive literature search was conducted to collect a diverse set of literature-reported PXR activators and nonactivators. A popular feature selection method, recursive feature elimination (RFE) (Guyon et al., 2002; Xue et al., 2004a,b; Li et al., 2005a,b), was used to extract molecular descriptors associated with PXR activation. The performance of these systems were tested by using 10-fold cross-validation and an independent set of 15 newly published experimental PXR activators (Lemaire et al., 2006).

Materials and Methods

Collection of PXR Activators and Nonactivators

Figure 1 illustrates the procedure for searching and selecting PXR activators, hPXR activators, and the corresponding nonactivators. PXR activators were selected based on the criterion that they have been reported to show potent activation to at least one PXR ortholog regardless of its effect on other PXR orthologs. A total of 128 PXR activators were collected from literature, which were used as the activator data set for predicting PXR activators irrespective of host species. There are 98 PXR activators reported to activate hPXR, which were used as the activator data set for predicting hPXR activators. The first data set is of higher statistical significance because of the higher number of compounds included. Compared with the largest data set of 53 compounds used in the previous studies (Ekins and Erickson, 2002; Jacobs, 2004; Schuster and Langer, 2005), our data sets contain a significantly higher number of compounds and are more diverse in structures, as shown by the computed structural diversity index as will be described.

PXR nonactivators include known PXR antagonists and PXR nonbinders reported in the literature. Moreover, compounds explicitly reported to not activate PXR-regulated gene expression of CYP3A4 were further considered to be implicated PXR nonactivators if they satisfied the subsequent criterion that they have not been reported to induce the expression of other PXR-regulated drug-metabolizing enzyme genes such as CYP2B6 and CYP2C8/9. These PXR nonactivators and implicated PXR nonactivators were used as the nonactivator data set for predicting PXR activators irrespective of host species. The hPXR nonactivator data set include all compounds in the PXR nonactivator data set plus known nonhuman PXR activators.

The 2D and 3D structure of each compound was generated by using ChemDraw (http://www.cambridgesoft.com/) and DS ViewerPro 5.0 (http://www.accelrys.com/), respectively, and geometrical optimization was conducted subsequently. The optimized 3D structure of each compound was manually inspected to ensure that the chirality of each chiral agent is properly generated and is consistent with that described in the literature. For those compounds with transactivation activities but without a reported active enantiomer, the

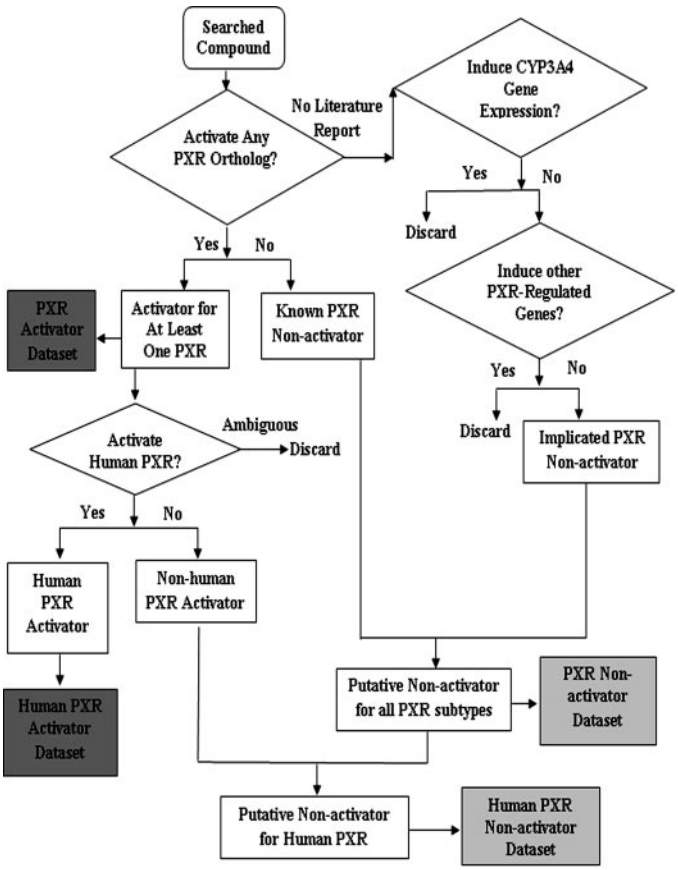

Fig. 1. A flowchart of the procedure for searching and selecting PXR activators, hPXR activators, and the corresponding nonactivators in this work.

default enantiomer structure in the chemical database such as PubChem (http://pubchem.ncbi.nlm.nih.gov/) and ChemFinder (http://www.chemfinder.com/) was straightforwardly used.

Determination of Structural Diversity

Structural diversity of a collection of compounds can be measured by using the diversity index (DI) value, which is the average value of the similarity between pairs of compounds in a data set (Perez, 2005):

$$DI = \frac{\sum_{i=1}^{N} \sum_{j=1, i \neq j}^{N} sim(i,j)}{N(N-1)} \quad (1)$$

where $sim(i,j)$ is a measure of the similarity between compounds i and j, and N is the number of compounds in the data set. The structural diversity of a data set increases with decreasing DI value. In this work, $sim(i,j)$ is computed by using the Tanimoto coefficient (Willett et al., 1998):

$$sim(i,j) = \frac{\sum_{d=1}^{l} \mathbf{x}_{di}\mathbf{x}_{dj}}{\sum_{d=1}^{l} (\mathbf{x}_{di})^2 + \sum_{d=1}^{l} (\mathbf{x}_{dj})^2 - \sum_{d=1}^{l} \mathbf{x}_{di}\mathbf{x}_{dj}} \quad (2)$$

where l is the number of descriptors computed for the molecules in the data set.

Construction of Training and Testing Sets

PXR and hPXR activators and nonactivators were divided into training and testing sets in a manner suitable for conducting 10-fold cross-validation study. For instance, the 128 PXR activators and 77 PXR nonactivators were each randomly divided into 10 subsets of approximately equal size. Nine of the subsets were used as the training set, and the remaining subset was used as the testing set for PXR activators and nonactivators, respectively. This process was repeated 10 times such that every subset is used as the test set once. The same procedure was applied to the 98 hPXR activators and 77 hPXR nonactivators for constructing the training and testing sets of the hPXR activator prediction systems. An additional set of 15 experimentally determined PXR activators (14 of which are structurally dissimilar in our data set by visual inspection) obtained from a newly published article (Lemaire et al., 2006) was used as the independent set for further evaluation of the performance of our prediction systems.

Molecular Descriptors

Molecular descriptors are quantitative representations of structural and physicochemical features of molecules, which have been extensively used in the structure-activity relationship (Fang et al., 2001), QSAR (Jacobs, 2004) and other machine learning studies of pharmaceutical agents (Doniger et al., 2002; Zernov et al., 2003; Xue et al., 2004b; Yap and Chen, 2004). A total of 199 molecular descriptors were used in this work. These descriptors were selected from more than 1000 descriptors described in the literature by eliminating those descriptors that are obviously redundant or unrelated to the prediction of pharmaceutical agents (Xue et al., 2004b; Li et al., 2005b). The resulting 199 molecular descriptors include 18 descriptors in the class of simple molecular properties, 28 descriptors in the class of molecular connectivity and shape, 97 descriptors in the class of electrotopological state, 31 descriptors in the class of quantum chemical properties, and 25 descriptors in the class of geometrical properties. They were computed from the 3D structure of each compound by using our own designed molecular descriptor computing program. A feature selection method, recursive feature elimination (described below), was used for eliminating those descriptors that are redundant or have no significant contribution to PXR activator prediction (Guyon et al., 2002).

Feature Selection Method

The RFE method (Guyon et al., 2002) was used in this work as the feature selection method for selecting molecular descriptors associated with PXR activation. RFE has gained popularity due to its effectiveness for improving prediction performance and for discovering informative features associated with drug activity (Guyon et al., 2002), pharmacokinetic, and toxicological properties (Xue et al., 2004a,b). Each of the compounds studied is represented by a vector $\boldsymbol{x}_i$, with its molecular descriptors (or features) as its components. The task of selecting appropriate molecular descriptors to a particular compound classification problem can be conducted by ranking and selecting those with meaningful contributions to the classification of the studied compounds.

Descriptor ranking in RFE is based on the magnitude of the change of an objective function of a MLM model upon removing each descriptor (which roughly measures the extent of contribution of each feature to the prediction capability of the model) (Kohavi and John, 1997). The prediction capability of a MLM model is more significantly affected by a greater change in the objective function, and thus the corresponding descriptor is ranked higher. To improve the efficiency of training, this objective function is represented by a cost function J computed from the training set only. When a given feature is removed or its weight is brought to 0, the change $DJ(i)$ in the cost function J is computed by $DJ(i) = [(1\partial^2 \mathbf{J})/(2\partial \mathrm{w}_i^2)] \times (Dw_i)^2$, where w_i is the weight of the feature i, and the change in weight $Dw_i = w_i$ corresponds to the removed descriptor $\mathbf{x}_i$. One or more of descriptors with the smallest $DJ(i)$ can be eliminated in each iteration (Guyon et al., 2002).

Machine Learning Methods

SVM. SVM is illustrated in Fig. 2. A linear SVM constructs a hyperplane separating two different classes of feature vectors with a maximum margin (Vapnik, 1995). This hyperplane is constructed by finding a vector $\mathbf{w}$ and a parameter b that minimizes $\|\mathbf{w}\|^2$, which satisfies the following conditions: $\mathbf{w} \times \mathbf{x}_i + b \geq +1$, for $y_i = +1$ (PXR activators as positive class) and $\mathbf{w} \times \mathbf{x}_i + b \leq -1$, for $y_i = -1$ (PXR nonactivators as negative class). Here $\mathbf{x}_i$ is a feature vector, y_i is the group index, $\mathbf{w}$ is a vector normal to the hyperplane, $|b|/\|\mathrm{w}\|$ is the perpendicular distance from the hyperplane to the origin, and $\|\mathrm{w}\|^2$ is the Euclidean norm of $\mathbf{w}$. A nonlinear SVM projects feature vectors into a high-dimensional feature space by using a kernel function such as $K(\mathbf{x}_i,\mathbf{x}_j) = \mathrm{e}^{-\|\mathbf{x}_j - \mathrm{i}\|^2/2\sigma^2}$. The linear SVM procedure is then applied to the feature vectors in this feature space. After the determination of $\mathbf{w}$ and b, a given vector $\mathbf{x}$ can be classified by using $sign[(\mathbf{w} \times \mathbf{x}) + b]$; a positive or negative value indicates that the vector $\mathbf{x}$ belongs to the positive or negative class, respectively.

k-NN. k-NN is illustrated in Fig. 3. k-NN measures the Euclidean distance between a to-be-classified vector $\mathbf{x}$ and each individual vector $\mathbf{x}_i$ in the training set (Johnson and Wichern, 1982). The Euclidean distances for the vector pairs are calculated using the following formula:

$$D = \sqrt{\|\mathbf{x} - \mathbf{x}_i\|^2} \quad (3)$$

A total of k number of vectors nearest to the vector $\boldsymbol{x}$ are used to determine its class, $f(\boldsymbol{x})$.

$$\hat{f}(\mathbf{x}) \leftarrow \arg\max_{v \in V} \sum_{i=1}^{k} \delta(v, f(\mathbf{x}_i)) \quad (4)$$

where $\delta(a,b) = 1$ if $a = b$ and $\delta(a,b) = 0$ if $a \neq b$, argmax is the maximum of the function, V is a finite set of vectors $\{v_1, \ldots v_s\}$, and

$\hat{f}$ is an estimate of $f(\mathbf{x})$. Here, estimate refers to the class of the majority of the k nearest neighbors.

PNN. As illustrated in Fig. 4, PNN is a form of neural network designed for classification through the use of Bayes' optimal decision rule (Specht, 1990):

$$h_i c_i f_i(\mathbf{x}) > h_j c_j f_j(\mathbf{x}) \tag{5}$$

where h_i and h_j are the prior probabilities, c_i and c_j are the costs of misclassification, and $f_i(\mathbf{x})$ and $f_j(\mathbf{x})$ are the probability density function for classes i and j, respectively. An unknown vector $\mathbf{x}$ is classified into population i if the product of all the three terms is greater for class i than for any other class j (not equal to i). In most applications, the prior probabilities and costs of misclassifications are treated as being equal. The probability density function for each class for a univariate case can be estimated by using the Parzen's nonparametric estimator

$$g(\mathbf{x}) = \frac{1}{n\sigma} \sum_{i=1}^{n} W\left(\frac{\mathbf{x} - \mathbf{x}_i}{\sigma}\right) \tag{6}$$

where n is the sample size, σ is a scaling parameter defining the width of the bell curve that surrounds each sample point, W(d) is a weight function that has its largest value at $d = 0$, and $(\mathbf{x} - \mathbf{x}_i)$ is the distance between the unknown vector and a vector in the training set. The Parzen's nonparametric estimator was later expanded by Cacoullos for the multivariate case:

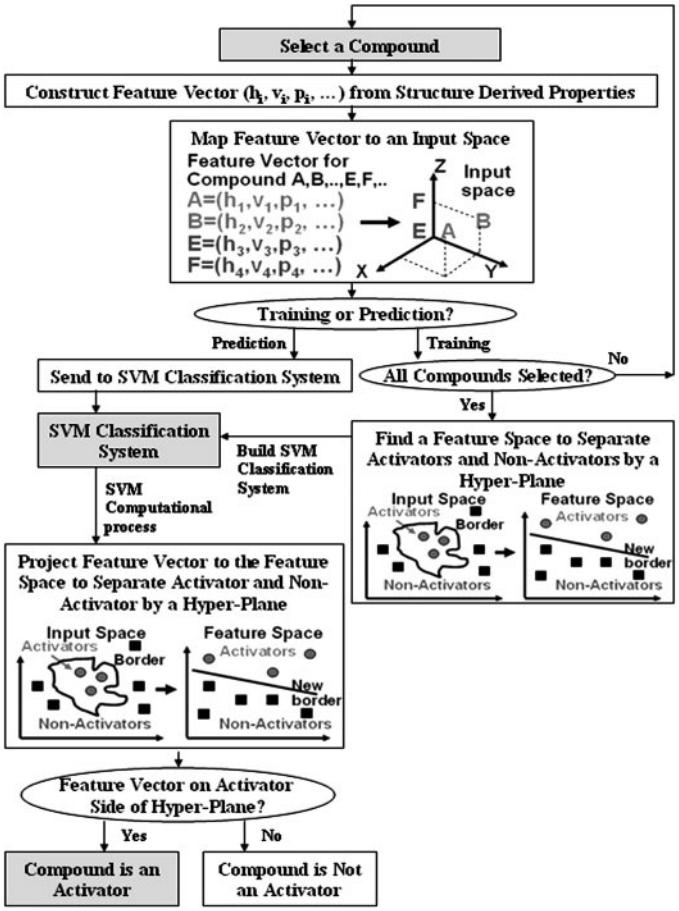

Fig. 2. Schematic diagram illustrates the process of predicting PXR activators by using SVM. A and B, feature vectors of agents with the property; E and F, feature vectors of agents without the property; feature vector (h_i, p_i, v_i,. . .) represents such structural and physicochemical properties as hydrophobicity, volume, etc.

$$g(x_1, \ldots, x_p) = \frac{1}{n\sigma_1 \ldots \sigma_p} \sum_{i=1}^{n} W\left(\frac{x_1 - x_{1,i}}{\sigma_1}, \ldots, \frac{x_p - x_{p,i}}{\sigma_p}\right) \tag{7}$$

The Gaussian function is frequently used as the weight function because it is well-behaved, easily calculated, and satisfies the conditions required by Parzen's estimator. Thus the probability density function for the multivariate case becomes

$$g(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} \exp\left(- \sum_{j=1}^{p} \left(\frac{x_j - x_{ij}}{\sigma_j}\right)^2 \right) \tag{8}$$

The network architectures of PNN are determined by the number of compounds and descriptors in the training set. There are four layers in a PNN. The input layer provides input values to all neurons in the pattern layer and has as many neurons as the number of descriptors in the training set. The number of pattern neurons is determined by the total number of compounds in the training set. Each pattern neuron computes a distance measure between the input and the training case represented by that neuron and then subjects the distance measure to the Parzen's nonparameteric estimator. The summation layer has a neuron for each class, and the neurons sum all the pattern neurons' output corresponding to members of that summation neuron's class to obtain the estimated probability density function for that class. The single neuron in the output layer then estimates the class of the unknown vector $\boldsymbol{x}$ by comparing all the probability density function from the summation neurons and choosing the class with the highest probability density function.

Evaluation of Prediction Performance

As in the case of all discriminative methods (Baldi et al., 2000), the performance of MLMs can be evaluated by the quantity of true

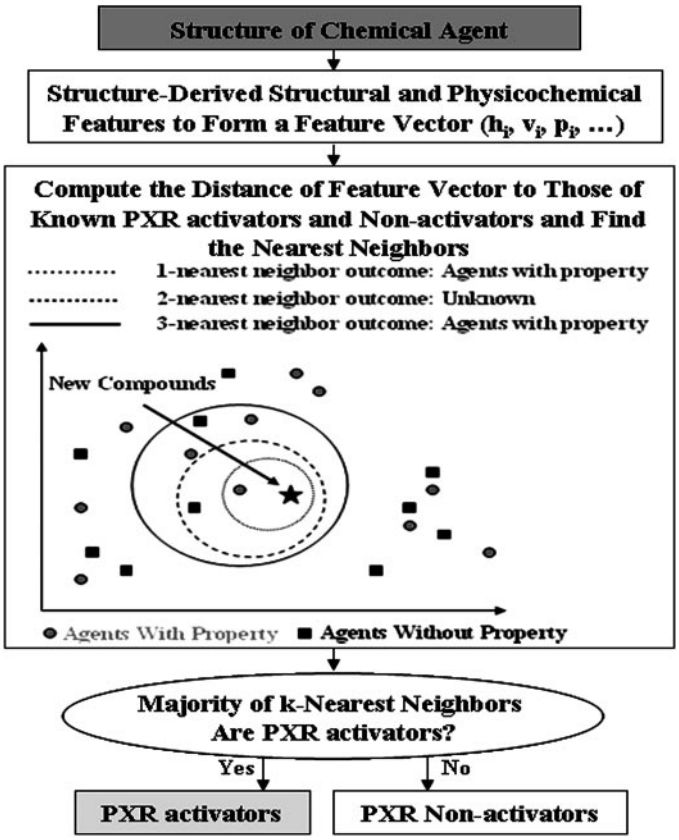

Fig. 3. Schematic diagram illustrating the process of the prediction of PXR activators by using k-NN.

positives (TP; true PXR activators), true negatives (TN; true nonactivators), false positives (FP; false PXR activators), and false negatives *FN* (false nonactivators). Sensitivity [$SE = TP/(TP + FN)$] and specificity [$SP = TN/(TN + FP)$] are the prediction accuracy for PXR activators and nonactivators, respectively. The overall prediction accuracy (Q) and Matthews' correlation coefficient (C) (Matthews, 1975) are used to measure the overall prediction performance:

$$Q = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

$$C = \frac{TP \times TN - FN \times FP}{\sqrt{(TP + FN)(TP + FP)(TN + FN)(TN + FP)}} \tag{10}$$

Computational Parameters and Performance Evaluation

There is only one parameter to be optimized in training each of the SVM, k-NN, and PNN classification systems. The classification speed of these MLM-based prediction systems is in the order of a few thousands to hundreds of thousands of compounds per second (Li et al., 2005a). The classification speed of SVM is usually 25 to 55% faster than that of k-NN and PNN because SVM typically uses 45 to 75% of the training set as support vectors for classification, whereas k-NN and PNN use the whole training set.

MLMs generally require a sufficient number of samples to develop a classification system. Irrelevant molecular descriptors may reduce the performance of these classification systems (Kohavi and John, 1997; Xue et al., 2004a,b; Li et al., 2005a). SVM has been found to be the least sensitive to data over-fitting, even in cases when a large number of redundant and overlapping molecular descriptors are used (Vapnik, 1995). This is because SVM is based on the structural risk minimization principle, which minimizes both training error and generalization error simultaneously.

SVM, k-NN, and PNN do not explicitly provide information about the importance of each molecular descriptor. For SVM, this problem is further compounded when kernel function is used because there is no simple method to inversely map the solution back into the input space. Incorporation of feature selection methods (Li et al., 2005b; Yap and Chen, 2005) and regression methods (Yap and Chen, 2004) have been frequently used for extracting important molecular descriptors from these machine learning-based prediction systems.

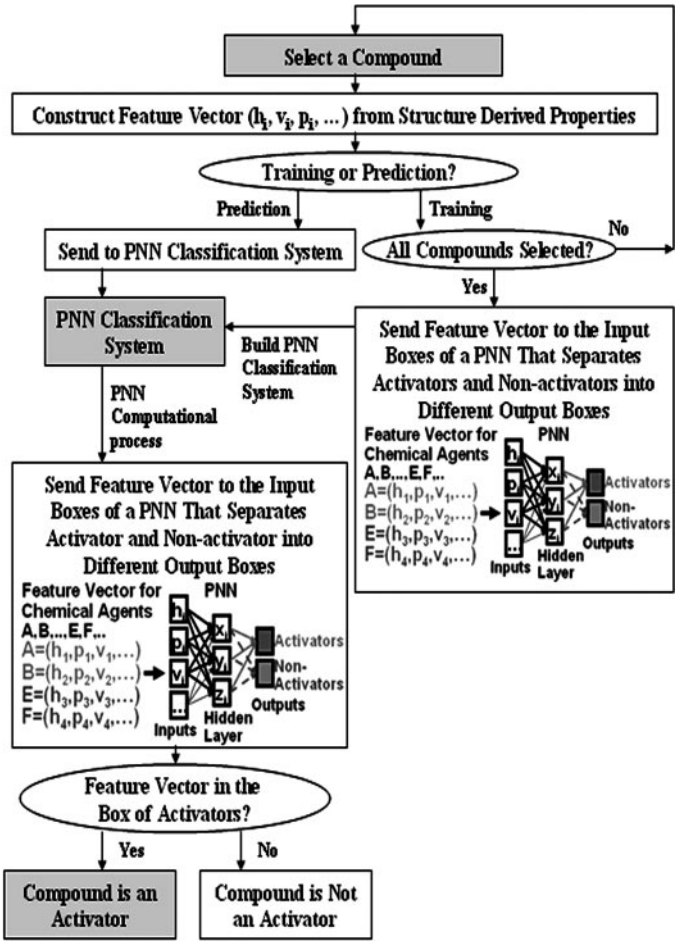

Fig. 4. Schematic diagram illustrates the process of predicting PXR activators by using PNN.

Results

Promiscuity Nature of PXR Activator Structures and the Selected Molecular Descriptors for Classifying PXR Activators. Table 1 gives the computed DI value of PXR activators and those of several groups of compounds possessing various different activities or properties. PXR activators are structurally more diverse not only than some of the well-known promiscuous binder groups such as estrogen receptor agonists and P-glycoprotein substrates, but also than some of the compound groups involved in multiple mechanisms such as human intestine absorbing agents. Figure 5 shows the structures of selected PXR activators, which are indicative of the extent of structural diversity of PXR activators. The DI value of our data set is 0.535, which is smaller than that of 0.605 of the largest data set of other PXR activators studies (Schuster and Langer, 2005). Therefore, our data set is structurally more diverse than those of other studies of PXR activators.

A total of 83 molecular descriptors, listed in Supplementary Table S2, were selected by the RFE method from a set of 199 molecular descriptors. These descriptors include simple molecular descriptions such as count of atom types (nhyd, nhal, nhet, ncocl, nnitro), ring (nring) and rotatable bonds (nrot), molecular connectivity and geometry (3χC, 4χPC, 5χCH, 6χCH, 1χv, 2χv, 3χvP, 3χvC, 4χvPC, 6χvCH, dis1, dis2, dis3, etc.), molecular flexibility (ϕ), electrotopological states or Estates [S car, S het, S hal, S(1), S(5), S(12), S(13), S(16), S(18), Tcent, Tradi, Tdiam, Tiwie, etc.], molecular surface area (polar molecular surface area, Sapc, Sanc, Sapcw, Sancw, Svpc, etc.), molecular shape (Rugty, Gloty), hydrophobicity (Shpl, Shpb, Hiwpl, Hiwpb, Hiwpa), and quantum chemical descriptors (εa, εb, μ, η, SN, IP, A, μ cp, χ en, ω, etc.).

Some features of these RFE-selected descriptors such as hydrophobicity, hydrogen bond acceptors, molecular globularity, and some Volsurf descriptors are also consistent with the structural features or descriptors described or used in the previous pharmacophore and QSAR studies of PXR activators. Pharmacophore models have shown that hydrophobic and hydrogen bond acceptors (HBAs) are important features for PXR activators (Ekins and Erickson, 2002; Schuster and Langer, 2005). In a QSAR study (Jacobs, 2004), hydrogen bond acceptors, dispersion forces, molecular globularity, and some VolSurf descriptors were found to be the key positive correlated variables for constructing the PXR QSAR model for predicting PXR activators.

The number of selected descriptors in this study is substantially larger than the 22 to 39 molecular descriptors selected in the prediction of compounds of various other drug activities or properties (Xue et al., 2004a,b; Li et al., 2005a,b).

An examination of the selected descriptors shows that most of the "extra" set of descriptors is from the electrotopological, connectivity, and quantum chemical classes. As shown in Fig. 5, apart from the usual chemical structures, a substantial number of PXR activators contain highly complex multiaromatic rings, or highly flexible chain-like structures, or

TABLE 1

DI for the compounds in several chemical groups, and the number of molecular descriptors selected by RFE for predicting each group of compounds by using a MLM classification system

These chemical groups are arranged in descending order of structural diversity.

Chemical Group	No. of Compounds	DI Value	No. of Molecular Descriptors Selected by RFE
Blood-brain barrier penetrating agents (Li et al., 2005b)	276	0.430	37
Genotoxic agents (Li et al., 2005a)	229	0.441	39
FDA-approved drugs	1121	0.495	
CYP3A4 inhibitors	233	0.505	
PXR activators (this work)	128	0.535	83
CYP2C9 inhibitors	167	0.541	
Negative chemical ionization diversity set	1804	0.544	
CYP3A4 substrates	362	0.547	
CYP2C9 substrates	144	0.552	
P-glycoprotein substrates (Xue et al., 2004b)	116	0.555	22
CYPD6 inhibitors	180	0.575	
CYP2D6 substrates	198	0.588	
Human intestine absorbing agents (Xue et al., 2004a)	131	0.596	27
PXR activators in Schuster and Langer's pharmacophore model (Schuster and Langer, 2005)	53	0.605	
Estrogen receptor agonists (Li et al., 2006)	243	0.618	31

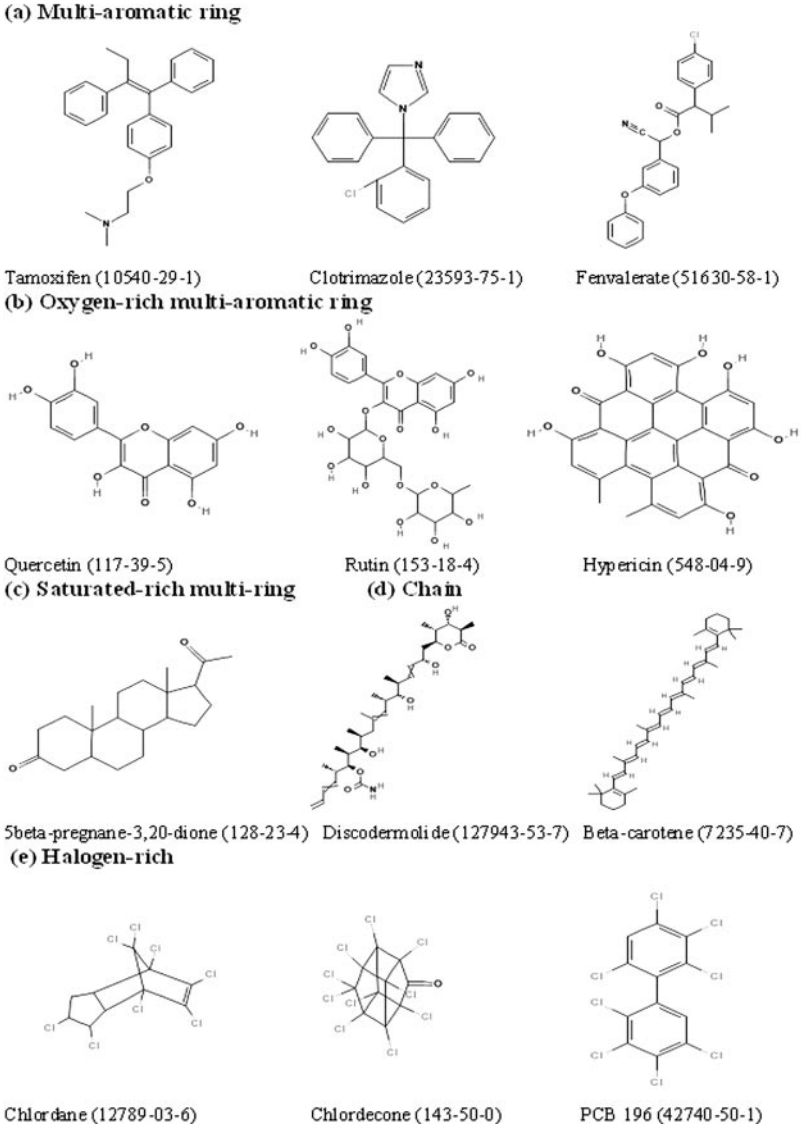

Fig. 5. Structure of selected PXR activators of different structural features. The CAS number of each compound is also given.

halogen-rich structures. These structural features coupled with highly diverse structural frameworks are probably the primary reasons for the need of the "extra" set of electrotopological, connectivity, and quantum mechanical descriptors in distinguishing PXR activators.

Performance of MLMs for Predicting PXR Activators. Table 2 gives the prediction performance of the three MLMs, with and without the use of the RFE feature-selection method, for predicting PXR and hPXR activators and nonactivators based on a 10-fold cross-validation study. The parameters of the PXR SVM, k-NN, and PNN systems are $\delta = 1$, $k = 1$, and $\delta = 0.3$, respectively. Those of the hPXR systems are $\delta = 1$, $k = 3$, and $\delta = 0.2$, respectively. The use of the RFE feature-selection method helps to improve the overall prediction performance of the PXR MLM systems from an accuracy level of 72.6 to 74.0% to that of 75.4 to 77.4% and that of the hPXR systems from an accuracy level of 72.5 to 74.9% to that of 75.0 to 79.6%. All of the MLM systems seem to show good performance. When considering overall prediction accuracies, PNN and SVM perform better than k-NN.

Our classification systems were further evaluated by using 15 newly published hPXR activators (Lemaire et al., 2006) whose structures are shown in Fig. 6. These include five herbicides (pretilachlor, metolachlor, oxadiazon, alachlor, and isoproturon), six fungicides (bupirimate, fenarimol, propiconazole, fenbuconazole, prochloraz, and imazalil), and four insecticides (toxaphene, permethrin, fipronil, and diflubenzuron). As shown in Table 3, 86.7, 73.3, and 73.3% of these activators were correctly predicted by the SVM, PNN, and k-NN PXR prediction systems, and 66.7, 66.7, and 53.3% were correctly predicted by the corresponding hPXR prediction systems, respectively. One possible reason for the lower accuracies of the hPXR systems is that they were trained by using compounds structurally more different from the newly published hPXR activators than some PXR activators in the training set of PXR prediction systems. As shown in supplementary Table S3, the Euclidean distance between the 15 newly published hPXR activators and the 28 PXR activators outside the hPXR data set is closer than that of the 98 hPXR activators. One activator, fenbuconazole, was incorrectly predicted by all of our PXR and hPXR systems. One possible reason for misclassifying this compound is that it contains a cyano group (-C≡N) that may not be adequately represented by existing molecular descriptors.

TABLE 2

Performance of three machine learning methods (k-NN, PNN, and SVM) for predicting PXR and hPXR activators and nonactivators determined by a 10-fold cross-validation study

The results are expressed in SE (sensitivity or prediction accuracy for PXR activators), SP (specificity or prediction accuracy for PXR non-activators), Q (overall accuracy), and C (Matthews' correlation coefficient).

Molecular Descriptors and Method	SE	SP	Q	C
	%	%	%	
All Species				
All descriptors				
k-NN	81.7	57.5	72.6	0.410
PNN	81.9	60.9	74.0	0.446
SVM	81.0	62.2	73.9	0.441
RFE-selected descriptors				
k-NN	84.0	61.2	75.4	0.473
PNN	82.8	68.4	77.4	0.526
SVM	81.2	70.3	77.1	0.528
Human				
All descriptors				
k-NN	80.6	62.4	72.5	0.448
PNN	80.7	63.8	73.2	0.461
SVM	77.8	71.4	74.9	0.504
RFE-selected descriptors				
k-NN	80.8	67.7	75.0	0.499
PNN	85.0	68.7	77.7	0.559
SVM	84.4	73.6	79.6	0.598

Discussion

Our selected descriptors are consistent with the molecular binding features derived from the study of the binding site of the ligand-free and drug-bound PXR receptor structures (Watkins et al., 2001). It has been reported (Watkins et al., 2001) that molecular flexibility, surface area, geometry, and connectivity are important for characterizing molecular recognition between PXR ligand-binding site and activators. The solved crystal structure of hPXR ligand-binding domain shows high mobility and flexibility in largely hydrophobic site that incorporates a few polar residues capable of forming hydrogen bonds with a binding ligand (Watkins et al., 2001, 2003a; Chrencik et al., 2005). Hydrogen bonds are important in determining the specificity of molecular recognition. Upon binding to PXR ligand-binding site, PXR activator is oriented in a specific orientation stabilized by hydrogen bonds and causes conformational change of PXR ligand binding domain to recruit the binding of coactivators. On the other hand, connectivity is important not only for discriminating between active from nonactive analogs but also for representing important molecular topological features involved in PXR activation. Moreover, electrotopological states, hydrophobicity, and quantum chemical descriptors describe polarity and charge of molecules that contribute in hydrogen bonding, polar, and salt-bridge interactions between PXR activators with the amino acid residues in the ligand-binding cavity of PXR.

PXR activators generally show higher content of halogen atoms, especially chlorine atoms, than nonactivators, as can be seen from higher mean values of halogen atom count (nhal) (1.16 versus 0.80), chlorine atom count (ncocl) (1.02 versus 0.27), and atom-type Estate sum for chlorine S(60) (6.33 versus 1.63). Moreover, PXR activators contain less nitrogen atoms (nnitro) than nonactivators (0.80 versus 1.79), and have lower values of several descriptors including the mean values of atom-type electrotopological state (Estate) sum for >NH, S(5) (0 versus 0.45); atom-type Estate sum for = N-, S(34) (0.31 versus 1.15); atom-type Estate sum for >N-, S(36) (0.15 versus 0.94); and atom-type Estate sum for −N≪, S(37) (0.05 versus 0.41). In addition, polar and salt bridges between PXR ligand binding domain residues and π-π stacking between aromatic rings of activators and ligand binding domain are also important for PXR activation. The descriptors for sums of solvent-accessible surface areas of positively charged atoms (Sapc, Sapcw, Svpc), negatively charged atoms (Sanc, Sancw), and ionization potential (IP) are associated with salt-bridge interactions. Those of atom-type Estate sum for CHn unsaturated, S(13), and atom-type Estate sum for :CH: aromatic, S(21), are relevant to π-π stacking.

Although PXR activators generally contain a fewer number of hydrogen bond donors (εa) and acceptors (εb) than those of nonactivators, nonetheless, hydrogen bonding plays some roles in activator binding to PXR. It was found that, on

average, PXR activators have higher number of HBAs (εb) than hydrogen bond donors (εa), which are consistent with the results from QSAR and pharmacophore studies (Ekins and Erickson, 2002; Jacobs, 2004; Schuster and Langer, 2005). A higher number of HBAs for PXR activators may result from the existence of the hydrogen bond donor-containing residues His327, His407, and Arg410 residues in the interior region of PXR ligand-binding site. These features are captured by the RFE-selected descriptors Svpcw and Svncw for the sum of weighted van der Waals surface areas of positive and negative atoms, respectively. The mean values for Svncw (36.25 versus 21.84) is larger than Svpcw for PXR activators showing complementary charge for activators to the PXR ligand-binding site may contribute to the entry of binding site and stable binding.

The computed mean values for the number of rotatable bonds (nrot) (4.45 versus 5.99), Kier molecular flexibility index (ϕ) (5.84 versus 6.24), polar molecular surface area (61.90 versus 69.53) of PXR activators are smaller than those of nonactivators, which is consistent with the view that PXR activators are generally smaller in size (Handschin and Meyer, 2005). The smaller size and number of rotatable bonds enable a better access to the ligand-binding site. Although how a ligand gains access to the PXR ligand-binding cavity remains unclear, it has been hypothesized that the flexible α2 (residues 192–205) that is unique to PXR may be

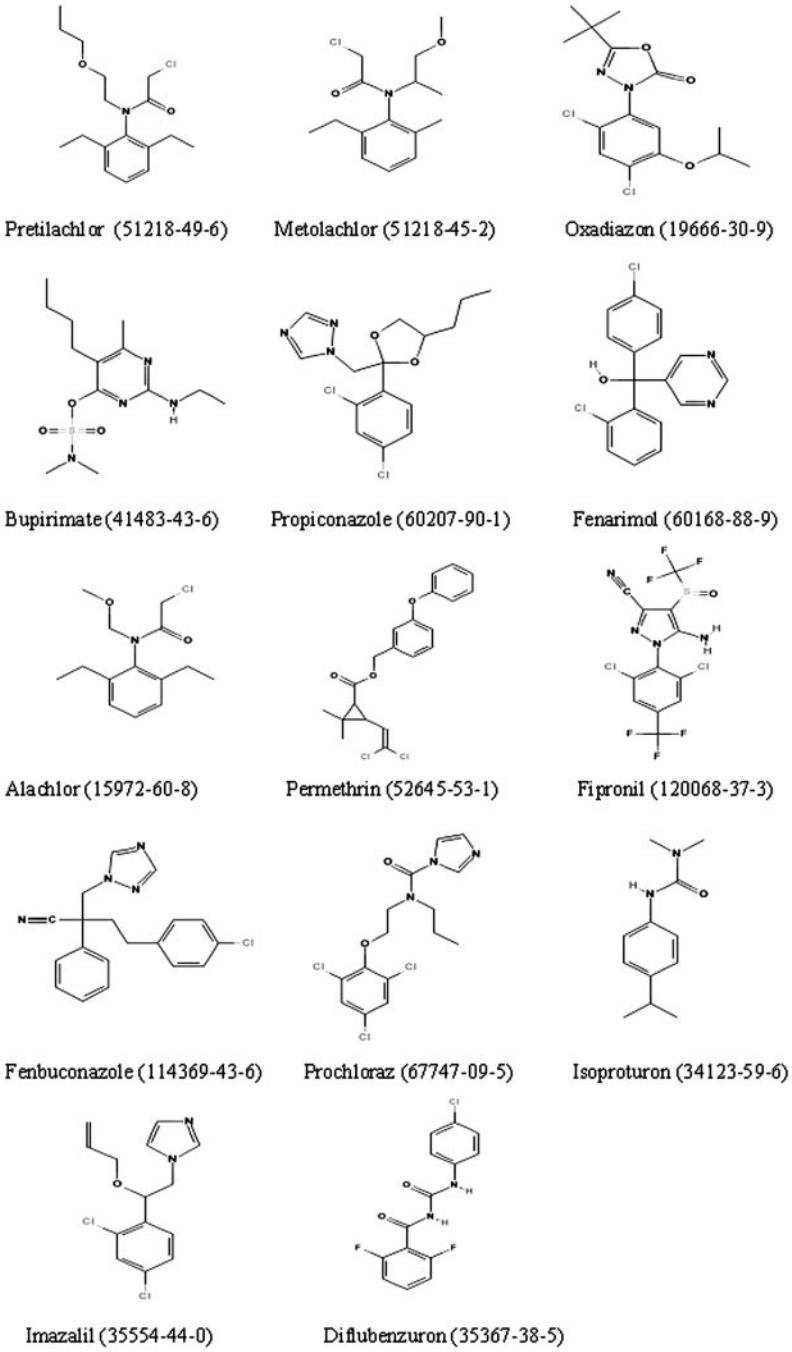

Fig. 6. Structure of 14 novel PXR activators from a recent publications (Lemaire et al., 2006). The CAS number for each compound is also given.

critical component for ligand entry and exit site (Watkins et al., 2003a). The flexible region may operate like a trapping-door allowing ligands to enter the central of the ligand-binding site. In addition, Leu209 located near the C terminus of α2 shifted in position by up to 7.7 E when bound by different ligands (Watkins et al., 2003b). Binding by coactivators further stabilizes the bound orientations of ligands. Taken together, the large and flexible ligand-binding pocket of PXR explains the promiscuous nature of PXR to bind to a variety of endogenous and xenobiotic compounds.

Although some aspects of activator binding to PXR can be exhibited by analyzing the selected descriptors, these descriptors are quantitative representations of structural and physicochemical features. Therefore, analysis of these descriptors without consideration of the receptor site structure is insufficient for providing a molecular level picture about the connection between a descriptor and the predicted activity. In the protein 3D structure database PDB (http://www.rcsb.org/pdb/Welcome.do), there are four entries of ligand-bound PXR structures. Analysis of some of these structures provides useful information about the atomic-level interactions represented by some of our selected descriptors. Figures 7 and 8 show the binding site structure of PXR bound by activator SR12813 (Watkins et al., 2001, 2003a; Chrencik et al., 2005) and hyperforin (Watkins et al., 2003b), respectively. Both activators form hydrophobic contacts with two hydrophobic residues, and they form hydrogen bonds with two polar residues. These provide clear molecular picture about the connection between our selected descriptors, hydrophobic and hydrogen bond descriptors, and activator-binding to PXR.

k-NN is based on a nearest neighbor algorithm that works best when activators and nonactivators tend to cluster in different regions or pockets of chemical space. SVM and PNN are based on nonlinear algorithms that are generally effective for all cases of distributions. SVM has fewer parameters than PNN, which makes it easier for deriving an optimal prediction system. MLMs are subjected to some degree of error due to such factors as data-set quality and the inherent limitation in predicting biological activities solely based on structure-derived molecular descriptors.

From the chemistry point of view, one can state that the molecular structure of a compound is the key in understanding its physicochemical properties and ultimately its biolog-

TABLE 3
Performance of the PXR and hPXR activator prediction systems for predicting the 15 recently published hPXR activators

SN	Compound Name	CAS Number	Relative Activity	Activator Prediction System					
				PXR			hPXR		
				SVM	PNN	k-NN	SVM	PNN	k-NN
1	Pretilachlor	51218-49-6	129.5	A	A	A	N	A	N
2	Toxaphene	8001-35-2	114.2	A	A	A	A	A	A
3	Metolachlor	51218-45-2	107.2	A	A	A	N	A	A
4	Oxadiazon	19666-30-9	94.2	A	A	A	A	A	N
5	Bupirimate	41483-43-6	93.5	A	A	A	A	A	A
6	Fenarimol	60168-88-9	89.6	A	A	N	A	N	A
7	Permethrin	52645-53-1	88.4	A	A	A	A	N	A
8	Propiconazole	60207-90-1	85.1	A	N	N	A	A	A
9	Alachlor	15972-60-8	71.3	A	A	A	N	A	A
10	Fipronil	120068-37-3	58.7	A	A	A	A	A	N
11	Fenbuconazole	114369-43-6	56.1	N	N	N	N	N	N
12	Prochloraz	67747-09-5	50.5	A	A	A	A	A	N
13	Isoproturon	34123-59-6	50.1	A	A	A	A	N	A
14	Imazalil	35554-44-0	46.5	N	N	A	N	N	N
15	Diflubenzuron	35367-38-5	33.0	A	N	N	A	A	N

A, predicted PXR activator; N, predicted PXR nonactivators.

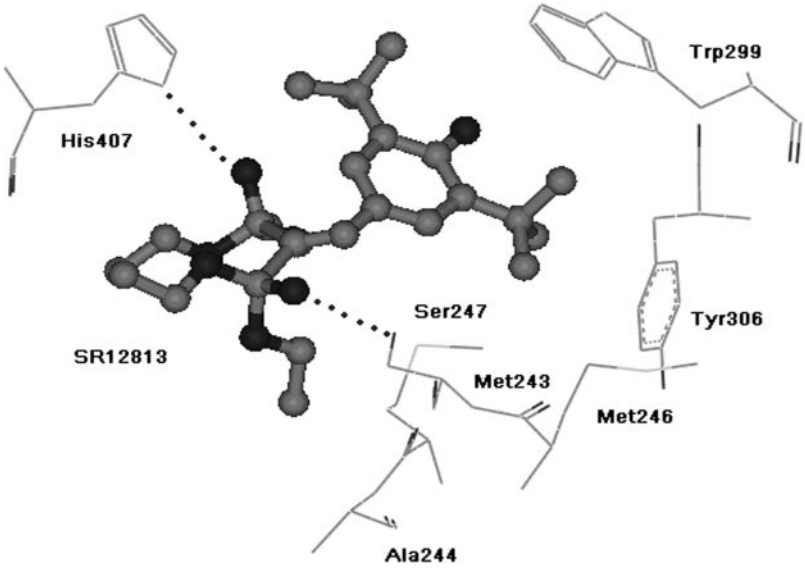

Fig. 7. Binding of PXR activator SR12813 (in ball-and-stick) at PXR (in wire frame) ligand-binding site. The activator forms hydrogen bonds with Ser247 and His407 and hydrophobic contact with Met243 and Met246.

ical activity and physiological effect (Johnson and Maggiora, 1990). Although hydrophobic interactions and hydrogen bondings are known to play important roles in molecular recognition from ligand-protein, protein-protein, up to macromolecular assemblies, there are many ways to describe these interactions from chemistry point of views as can be expressed by various molecular descriptors. However, which descriptions are more relevant to a given activity has to be further characterized by various means such as using feature selection methods in the machine learning methods.

Current representations of molecular physicochemical properties by molecular descriptors are still far from complete. Further refinement to develop a more sophisticated set of molecular descriptors is definitely an important task. Moreover, it is essential to include more PXR activators and nonactivators from future experimental works. Currently, we used a set of 199 molecular descriptors. However, when the data set grows in future, we believe more completed set of molecular descriptors is required. Furthermore, the biological activity of a compound is an induced response that is influenced by numerous factors dictated by many levels of biological complexity. The relationship between structure and activity is thus more implicit and thereby requires a more thorough investigation and rigorous validation (Tong et al., 2004). Hence, the choice for better descriptors is still under investigation.

Conclusion

Identification of novel PXR activators from structurally diverse compounds is important for the discovery of drugs with desired metabolic and toxicological profiles. This study shows that MLMs and especially SVM are useful for in silico prediction of the activators of highly promiscuous proteins, such as PXR and for characterizing the molecular features of PXR activation. By incorporating feature-selection methods such as RFE into MLMs, molecular descriptors relevant to PXR activators can be identified. Most of these selected molecular descriptors are consistent to those used in previous pharmacophore and QSAR studies and with the findings from X-ray crystallography studies. Further works on the improvement and refinement of feature selection methods and molecular descriptors are needed to improve the capability of MLMs for accurately identifying PXR activators and the related molecular characteristics.

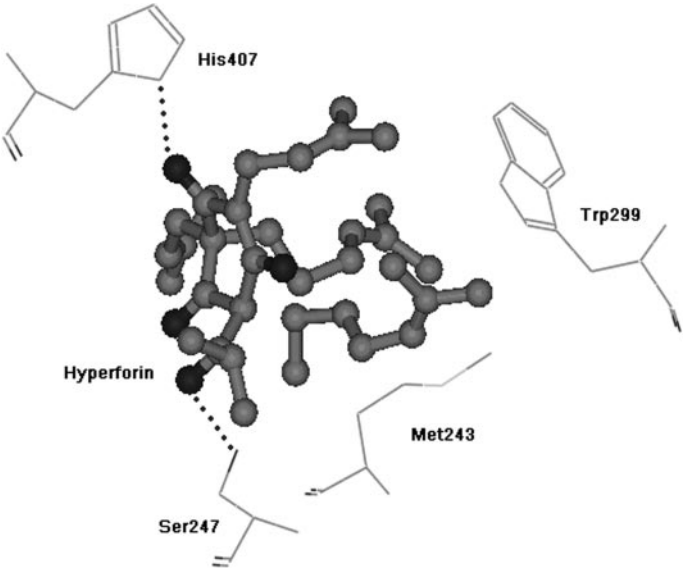

Fig. 8. Binding of PXR activator hyperforin (in ball-and-stick) at PXR (in wire frame) ligand-binding site. The activator forms hydrogen bonds with Ser247 and His407 and hydrophobic contact with Met243 and TRP299.

References

Baldi P, Brunak S, Chauvin Y, Andersen CA, and Nielsen H (2000) Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics* **16:**412–424.

Burges CJC (1998) A tutorial on support vector machines for pattern recognition. *Data Min Knowl Disc* **2:**127–167.

Chrencik JE, Orans J, Moore LB, Xue Y, Peng L, Collins JL, Wisely GB, Lambert MH, Kliewer SA, and Redinbo MR (2005) Structural disorder in the complex of human pregnane X receptor and the macrolide antibiotic rifampicin. *Mol Endocrinol* **19:**1125–1134.

Doniger S, Hofman T, and Yeh J (2002) Predicting CNS Permeability of drug molecules: comparison of neural network and support vector machine algorithms. *J Comput Biol* **9:**849–864.

Ekins S (2004) Predicting undesirable drug interactions with promiscuous proteins in silico. *Drug Discov Today* **9:**276–285.

Ekins S, Bravi G, Binkley S, Gillespie JS, Ring BJ, Wikel JH, and Wrighton SA (2000) Three- and four-dimensional-quantitative structure activity relationship (3D/4D-QSAR) analyses of CYP2C9 inhibitors. *Drug Metab Dispos* **28:**994–1002.

Ekins S and Erickson JA (2002) A pharmacophore for human pregnane X receptor ligands. *Drug Metab Dispos* **30:**96–99.

Fang H, Tong W, Shi LM, Blair R, Perkins R, Branham W, Hass BS, Xie Q, Dial SL, Moland CL, et al. (2001) Structure-activity relationships for a large diverse set of natural, synthetic, and environmental estrogens. *Chem Res Toxicol* **14:**280–294.

Guyon I, Weston J, Barnhill S, and Vapnik V (2002) Gene selection for cancer classification using support vector machines. *Mach Learn* **46:**389–422.

Handschin C and Meyer UA (2005) Regulatory network of lipid-sensing nuclear receptors: roles for CAR, PXR, LXR, and FXR. *Arch Biochem Biophys* **433:**387–396.

Jacobs MN (2004) In silico tools to aid risk assessment of endocrine disrupting chemicals. *Toxicology* **205:**43–53.

Johnson M and Maggiora GM (1990) *Concepts and Applications of Molecular Similarity.* Wiley, New York.

Johnson RA and Wichern DW (1982) *Applied Multivariate Statistical Analysis.* Prentice Hall, Englewood Cliffs, NJ.

Jones SA, Moore LB, Shenk JL, Wisely GB, Hamilton GA, McKee DD, Tomkinson NC, LeCluyse EL, Lambert MH, Willson TM, et al. (2000) The pregnane X receptor: a promiscuous xenobiotic receptor that has diverged during evolution. *Mol Endocrinol* **14:**27–39.

Kliewer SA, Goodwin B, and Willson TM (2002) The nuclear pregnane X receptor: a key regulator of xenobiotic metabolism. *Endocr Rev* **23:**687–702.

Kohavi R and John GH (1997) Wrappers for feature subset selection. *Artif Intell Med* **97:**273–324.

Krasowski MD, Yasuda K, Hagey LR, and Schuetz EG (2005) Evolutionary selection across the nuclear hormone receptor superfamily with a focus on the NR1I subfamily (vitamin D, pregnane X, and constitutive androstane receptors). *Nucl Recept* **3:**2.

Lehmann JM, McKee DD, Watson MA, Willson TM, Moore JT, and Kliewer SA (1998) The human orphan nuclear receptor PXR is activated by compounds that regulate CYP3A4 gene expression and cause drug interactions. *J Clin Investig* **102:**1016–1023.

Lemaire G, Mnif W, Pascussi JM, Pillon A, Rabenoelina F, Fenet H, Gomez E, Casellas C, Nicolas JC, Cavailles V, et al. (2006) Identification of new human PXR ligands among pesticides using a stable reporter cell system. *Toxicol Sci* **91:**501–509.

Li H, Ung CY, Yap CW, Xue Y, Li ZR, Cao ZW, and Chen YZ (2005a) Prediction of genotoxicity of chemical compounds by statistical learning methods. *Chem Res Toxicol* **18:**1071–1080.

Li H, Ung CY, Yap CW, Xue Y, Li ZR, and Chen YZ (2006) Prediction of estrogen receptor agonists and characterization of associated molecular descriptors by statistical learning methods. *J Mol Graph Model* **25:**313–323.

Li H, Yap CW, Ung CY, Xue Y, Cao ZW, and Chen YZ (2005b) Effect of selection of molecular descriptors on the prediction of blood-brain barrier penetrating and nonpenetrating agents by statistical learning methods. *J Chem Inf Model* **45:**1376–1384.

Matthews BW (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta* **405:**442–451.

Moore LB, Maglich JM, McKee DD, Wisely B, Willson TM, Kliewer SA, Lambert MH, and Moore JT (2002) Pregnane X receptor (PXR), constitutive androstane receptor (CAR), and benzoate X receptor (BXR) define three pharmacologically distinct classes of nuclear receptors. *Mol Endocrinol* **16:**977–986.

Perez JJ (2005) Managing molecular diversity. *Chem Soc Rev* **34:**143–152.

Schuster D and Langer T (2005) The identification of ligand features essential for PXR activation by pharmacophore modeling. *J Chem Inf Model* **45:**431–439.

Specht DF (1990) Probabilistic neural networks. *Neural Netw* **3:**109–118.

Tong W, Fang H, Hong H, Xie Q, Perkins R, and Sheehan D (2004) Receptor-mediated toxicity: QSARs for oestrogen receptor binding and priority setting of potential oestrogenic endocrine disruptors, in *Predicting Chem Toxicity and Fate* (Cronin MTD and Livingstone D eds) pp 285–314, CRC Press, Boca Raton, FL.

Vapnik VN (1995) *The Nature of Statistical Learning Theory.* Springer, New York.

Watkins RE, Davis-Searles PR, Lambert MH, and Redinbo MR (2003a) Coactivator

binding promotes the specific interaction between ligand and the pregnane X receptor. *J Mol Biol* **331:**815–828.
Watkins RE, Maglich JM, Moore LB, Wisely GB, Noble SM, Davis-Searles PR, Lambert MH, Kliewer SA, and Redinbo MR (2003b) 2.1 A crystal structure of human PXR in complex with the St. John's wort compound hyperforin. *Biochemistry* **42:**1430–1438.
Watkins RE, Wisely GB, Moore LB, Collins JL, Lambert MH, Williams SP, Willson TM, Kliewer SA, and Redinbo MR (2001) The human nuclear xenobiotic receptor PXR: structural determinants of directed promiscuity. *Science (Wash DC)* **292:** 2329–2333.
Willett P, Barnard JM, and Downs GM (1998) Chem similarity searching. *J Chem Inf Comput Sci* **38:**983–996.
Xie W, Uppal H, Saini SP, Mu Y, Little JM, Radominska-Pandya A, and Zemaitis MA (2004) Orphan nuclear receptor-mediated xenobiotic regulation in drug metabolism. *Drug Discov Today* **9:**442–449.
Xue Y, Li ZR, Yap CW, Sun LZ, Chen X, and Chen YZ (2004a) Effect of molecular descriptor feature selection in support vector machine classification of pharmacokinetic and toxicological properties of chemical agents. *J Chem Inf Comput Sci* **44:**1630–1638.
Xue Y, Yap CW, Sun LZ, Cao ZW, Wang JF, and Chen YZ (2004b) Prediction of p-glycoprotein substrates by support vector machine approach. *J Chem Inf Comput Sci* **44:**1497–1505.
Yap CW and Chen YZ (2004) Quantitative structure-pharmacokinetic relationships for drug distribution properties by using general regression neural network. *J Pharm Sci* **94:**153–168.
Yap CW and Chen YZ (2005) Prediction of cytochrome P450 3A4, 2D6, and 2C9 inhibitors and substrates by using support vector machines. *J Chem Inf Model* **45:**982–992.
Zernov VV, Balakin KV, Ivaschenko AA, Savchuk NP, and Pletnev IV (2003) Drug discovery using support vector machines. The case studies of drug-likeness, agrochemical-likeness, and enzyme inhibition predictions. *J Chem Inf Comput Sci* **43:**2048–2056.

Address correspondence to: Dr. Y. Z. Chen, Bioinformatics and Drug Design Group, Department of Pharmacy and Department of Computational Science, National University of Singapore, Blk S16, Level 8, 3 Science Drive 2, Singapore 117543. E-mail: phacyz@nus.edu.sg