Chemometrics and Intelligent Laboratory Systems 118 (2012) 271–279

Contents lists available at SciVerse ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

In silico identification of human pregnane X receptor activators from molecular descriptors by machine learning approaches

Hanbing Rao [a,*], Yanying Wang [a], Xianyin Zeng [a], Xianxiang Wang [a], Yong Liu [a], Jiajian Yin [a], Hua He [b], Feng Zhu [c], Zerong Li [d,*]

[a] College of Life and Science, Sichuan Agricultural University, Ya'an 625014, PR China
[b] Animal Genetics and Breeding Institute of Sichuan Agricultural University, Ya'An 625014, Sichuan, PR China
[c] Bioinformatics and Drug Design Group, Department of Pharmacy and Centre for Computational Science and Engineering, National University of Singapore, Blk. S16, Level 8, 3 Science Drive 2, Singapore, 117543, Singapore
[d] College of Chemistry, Sichuan University, Chengdu 610065, PR China

ARTICLE INFO

Article history:
Received 27 December 2011
Received in revised form 22 May 2012
Accepted 25 May 2012
Available online 6 June 2012

Keywords:
hPXR
Machine learning approaches
Applicability domain

ABSTRACT

In the current study, computational models for hPXR activators and hPXR non-activators were developed using support vector machine (SVM), *k*-nearest neighbor (*k*-NN), and artificial neural networks (ANN) algorithms. 73 molecular descriptors used for hPXR activator and hPXR non-activator prediction were selected from a pool of 548 descriptors by using a multi-step hybrid feature selection method combining Fischer's score and Monte Carlo simulated annealing method. The y-scrambling method was used to test if there is a chance correlation in the developed SVM model. In the meantime, five-fold cross validation of these machine learning methods results in the prediction accuracies of 87.2–92.5% for hPXR activators and 73.8–87.8% for hPXR non-activators, and the prediction accuracies for external test set are 93.8–95.8% for hPXR activators and 86.7–92.8% for hPXR non-activators. Our study suggested that the tested machine learning methods are potentially useful for hPXR activators identification.

1. Introduction

The human pregnane X receptor (hPXR) is a transcriptional regulator of a large number of genes involved in xenobiotic metabolism and excretion. The hPXR activators include a wide range of prescriptions and herbal drugs such as paclitaxel, troglitazone, rifampicin, ritonavir, and clotrimazole, which can be involved in clinically relevant drug–drug interactions [1]. In addition to xenobiotics, PXR is also activated by pregnanes, androstanes, bile acids, hormones, dietary vitamins, and a wide array of endogenous molecules recently reviewed [2].

The PXR ligand-binding domain (LBD) consists of 12 α-helices that fold to form a hydrophobic pocket and a short region of β-strands. The pocket is lined with 28 amino acid residues, 10 hydrophobic, four polar, and four charged [3–7]. The potential for molecules to bind to numerous locations in the LBD complicates the reliable prediction of PXR activators (Y) and non-activators (N) by using structure-based drug design methods alone. Computational models, such as ligand based pharmacophores [8–11], quantitative structure–activity relationships (QSAR) [12–14], machine learning methods [14,15], and homology modeling with molecular dynamics [16] (for identifying protein–corepressor interactions), were frequently used to predict PXR ligand binding [2]. These methods focused on diverse structural types of agonists and in one case used structural analogs [2], which may have assessed specific binding locations within the LBD like steroidal compounds. A likely consensus emerged that different chemical types of PXR agonists (such as imidazole, steroidal, and other chemicals with structural diversity) tend to fit to multiple hydrophobic features and at least one hydrogen bond acceptor (in some cases an additional hydrogen bond donor feature) [2]. PXR agonist QSAR or pharmacophore models are highly dependent on the promiscuous nature of the molecules in the training dataset, and the prediction results of models generated by different training datasets may overlap with each other [2]. Moreover, the published QSAR models rarely utilize a large external test set to validate their predictive nature or assess their applicability domain [17–19]. In other words, it is unclear how structural similarity between training and testing data would affect the prediction accuracy, especially for structurally promiscuous protein as PXR. Currently, many EC50 data of PXR agonists were reported as greater or less than a certain value without illustrating their exact activity value, which makes it difficult to construct quantitative QSAR models for hPXR. In the meantime, machine learning methods (MLMs), such as support vector machine (SVM), *k*-nearest neighbors (*k*-NN), and probabilistic neural network (PNN), were applied to identify PXR activators [14]. Binary classification data of 98 hPXR activators and 79 non-activators were used [14] to construct MLM models, and the prediction results ranged from 80.8% to 85.0% of hPXR activators

* Corresponding authors.
E-mail addresses: rhbsau@gmail.com (H. Rao), lizerong@scu.edu.cn (Z. Li).

0169-7439/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2012.05.012

and 67.7% to 73.6% of hPXR non-activators (in the training set). For 15 known hPXR activators in external test set, the prediction accuracies are between 53.3% and 66.7% across all three machine learning methods, with SVM performing the best [14]. Khandelwal et al. [15] have compared recursive partitioning (RP), random forest (RF), and SVM machine learning methods for building hPXR models derived with VolSurf three-dimensional (3D) descriptors. The predictive ability of the classification and docking models was further evaluated by using a novel large external test set containing 145 hPXR activators and non-activators, which aims at prioritizing molecules for in vitro testing.

The purpose of this work is to develop a new hPXR activator prediction model based on a more diverse training dataset using various machine learning methods.

One important step in developing machine learning hPXR activator prediction model is to compute and select appropriate molecular descriptors. A single or a standard set of descriptor according to experience may reflect adducting features to some extent, but cannot guarantee a full capture of the whole properties. In other words, there is no pre-knowledge on descriptors that are most relevant to hPXR activator prediction, so a priori feature selection is not feasible. In this work, we calculated molecular descriptors as many as possible and select the appropriate ones by using feature selection algorithms (FSAs).

There are two major classes of FSAs: classifier independent and classifier dependent. A classifier independent approach is a filter method [20–22] as outlined in Fig. 1a, which is computationally efficient. The filter method attempts to identify relevant features by selecting a feature subset using a preprocessing step independent of the learning algorithm, which is less useful for redundant features and data with strongly correlated features. Classifier dependent FSA is also called wrapper approach [23–26] as shown in Fig. 1b, which uses a specific learning algorithm, such as decision trees and support vector machines, to evaluate the feature subset based on their contribution to the performance of the learner. The wrapper approach has the advantage of selecting features suitable to the specific learner, and hence generally results in higher learning performance than filter method. In the wrapper approach, the selections of subset of features are imbedded in the classifier, such as recursive feature elimination (RFE) [27], genetic algorithm (GA) [28] and simulated annealing method (SA) [29]. Compared with the filter method, the wrapper approach is much more computationally expensive, but is able to produce better results. A detailed introduction to the wrapper approach can be found in ref. [30].

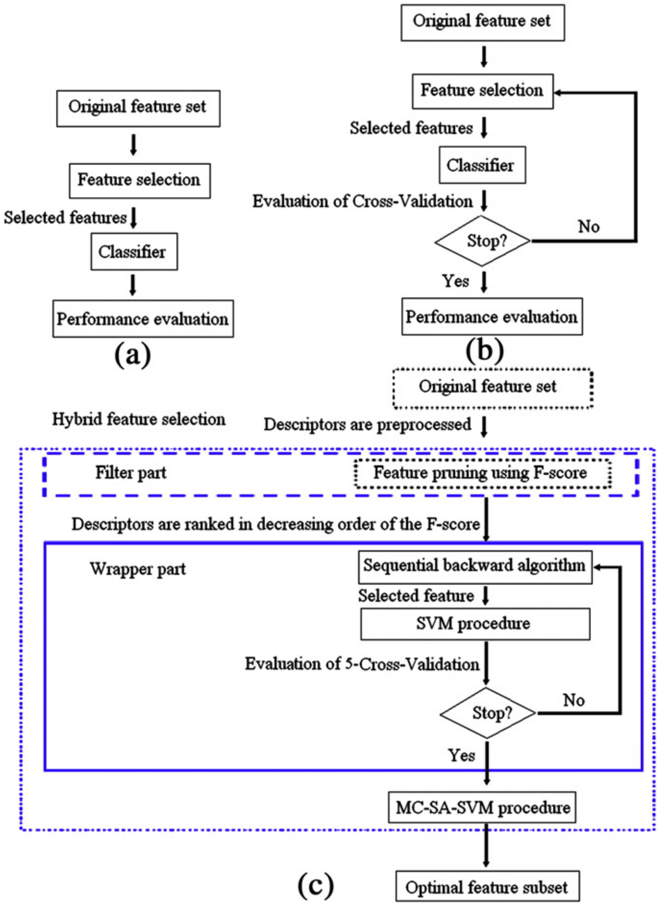

Fig. 1. Comparison of feature selection method (a) filter method, (b) wrapper method and (c) hybrid method.

In this study, to overcome the computational cost of wrapper approach and the low accuracy of filter method, a multi-step hybrid FSA combining F-score and Monte Carlo simulated annealing (F-MC-SA), as shown in Fig. 1c, was used to select most relevant descriptors for hPXR activator prediction. Different from other ranking algorithms like information based method, the F-score filter approach is capable of calculating continuous features, without discretizing them. Moreover, Monte Carlo simulated annealing, a wrapper approach, is very efficient for searching global minimum. In the meantime, several hPXR activator prediction machine learning models were developed in combination with our hybrid feature selection method. The performance of the developed models was further evaluated by different approaches: y-scrambling, five cross-validations and an external test dataset.

2. Materials and methods

2.1. Datasets

A diverse set of 362 hPXR and hPXR non-activator compounds (Supporting Information Table S1) is collected from literatures [14,15,31,32]. The term half maximal effective concentration (EC_{50}) refers to the concentration of a drug, antibody or toxicant which induces a response halfway between the baseline and maximum after some specified exposure time. It is commonly used as a measure of drug's potency. In accordance with the work of Khandelwal [15], $EC_{50} = 100$ μM is used as the threshold of classifying the compounds as hPXR activators or hPXR non-activators: compounds with $EC_{50} < 100$ μM are considered as hPXR activators and compounds with $EC_{50} \geq 100$ μM are considered as the hPXR non-activators. Overall, a total of 222 hPXR activators and 140 hPXR non-activators were used to develop and test machine learning models.

The SMILES string for each molecule was downloaded from PubChem (http://pubchem.ncbi.nlm.nih.gov/) and ChemSpider (http://www.chemspider.com/), or sketched using ChemDraw. Then, each molecule is drawn and optimized by using the MM+ force field implemented in HyperChem7.0 [33] software.

2.2. Measurement of structural diversity of compounds

The diversity of a dataset can be assessed by using diversity index (*DI*), which is average value of the dissimilarity between all pairs of molecules in the dataset [34].

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

where N is the number of compounds in the data set, and $diss(i,j)$ is a measure of the dissimilarity between objects i and j. Dissimilarity is a complementary measure of similarity, so that if a measure of similarity between two objects is defined between 0 and 1, its dissimilarity is simply defined as (1-similarity), accordingly, we have:

$$DI = 1 - \frac{\sum_{i=1}^{N}\sum_{j=1,i\neq j}^{N} sim(i,j)}{N(N-1)} \tag{2}$$

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.

One of the most widely used measure of similarity is the Tanimoto coefficient for continuous variables [35,36], defined as follows:

$$s(i,j) = \frac{\sum_{d=1}^{l} x_{di}x_{dj}}{\sum_{d=1}^{l}(x_{di})^2 + \sum_{d=1}^{l}\left(x_{dj}\right)^2 - \sum_{d=1}^{l} x_{di}x_{dj}} \tag{3}$$

where l is the number of descriptors computed for the molecules in the data set, i and j represent attributes, x_{di} and x_{dj} are the values of dth attribute in objects i and j, respectively. The Tanimoto coefficient ranges from -0.333 to $+1.0$, (continuous). $s(i,j)=1$ when $x_{di}=x_{dj}(d=1,2,\ldots,n)$ and $s(i,j) = -\frac{1}{3}$ when $x_{di}=-x_{dj}(d=1,2,\ldots,n)$. Therefore, in this study, the Tanimoto coefficient is normalized to between 0 and 1 before substituted to Eq. (2) by the following procedure:

$$sim(i,j) = \frac{s(i,j)+0.333}{1.333} \tag{4}$$

where $s(i, j)$ is the Tanimoto coefficient and $sim(i, j)$ is the normalized Tanimoto coefficient. Now, DI is within the range from 0 to 1, and the structural diversity of a dataset increases with increasing DI value. When $DI=1$, compounds in the dataset have a zero-valued similarity, that is, the dataset is sufficiently diverse for the given molecular descriptors. When $DI=0$, all compounds have identical molecular descriptors. Obviously, the closer the DI value is to 1, the more diverse the dataset investigated is. In this study, the DI of 362 hPXR agonists equals to 0.765, which indicates that the investigated compounds are structurally diverse.

2.3. Construction of training and testing sets

In total, 362 hPXR activators and hPXR non-activators were divided into training set (279 compounds) and external test set (83 compounds) using Kennard–Store (KS) method. It is found that the KS method outperforms the other methods [37], because the samples chosen by the KS method can span the largest chemical space, so the prediction for most of the compounds in the test set will be interpolation and fall into the applicability domain of the chemical space covered by the training set and the model can have best prediction ability for unknown compounds. The training set is used in the model building and model optimization by five-fold cross-validation method: the 172 hPXR activators and 107 hPXR non-activators in the original training set were randomly divided into 5 subsets of approximately equal size. Four of the subsets were combined and used as the training set, and the remaining subset was used as the testing set. This process was repeated 5 times such that every subset is used as the testing set once. An additional set of 83 compounds, including 48 hPXR activators and 35 hPXR non-activators, was used as the external test set for validating the prediction systems.

2.4. Molecular descriptor calculation

Geometry optimization of each molecule was performed using the MM+ force field of HyperChem7 [33] before computing the molecular descriptors. Molecular descriptors have been routinely used for quantitative description of the structural and physicochemical properties of molecules in the development of various QSAR models [38–41]. We used 548 1D and 2D descriptors (see Supplementary Table S2) by the web based software Model [42], which include 72 fingerprint descriptors [38], 30 constitutional descriptors [38], 92 molecular connectivity and molecular shape descriptors [38,43], 108 electro-topological state descriptors [31,44], 60 BCUT molecular descriptors [31,45] and 186 autocorrelation descriptors [31,46].

2.5. Feature selection method

Obviously, not all of the molecular descriptors are relevant to discriminate hPXR activators from hPXR non-activators. Elimination of the redundant descriptors can improve the prediction accuracy, and facilitate the interpretation of the model by focusing on the most relevant descriptors. In this study, a hybrid feature selection method is used to find the optimal subset of features with the following procedures:

1) Preprocessing: Firstly, any descriptor that has an identical value for more than 90% of the samples is removed. Secondly, any descriptor with the relative standard deviation of less than 0.05 is removed. Finally, one of any two descriptors with the absolute value of Pearson correlation coefficient above 0.9 is removed.
2) F-score ranking and backward selecting: The descriptors left after preprocessing are ranked in decreasing order of F-score. F-score is a simple technique which measures the discrimination of two sets of real numbers. Given training vectors $\boldsymbol{x}_k$, $k=1,\ldots$, n, if the number of positive and negative samples are n_+ and n_-, respectively, then, the F-score of the ith feature is defined as [47]:

$$F(i) = \frac{\left(\overline{x_i}^{(+)} - \overline{x_i}\right)^2 + \left(\overline{x_i}^{(-)} - \overline{x_i}\right)^2}{\frac{1}{n_+-1}\sum_{k=1}^{n_+}\left(x_{k,i}^{(+)} - \overline{x_i}^{(+)}\right)^2 + \frac{1}{n_--1}\sum_{k=1}^{n_-}\left(x_{k,i}^{(-)} - \overline{x_i}^{(-)}\right)^2} \tag{5}$$

where x_i, $\overline{x_i}^{(+)}$, and $\overline{x_i}^{(-)}$ are the average of the ith feature of the whole, positive and negative data sets, respectively; $x_{k,i}^{(+)}$ denotes the ith feature of the kth positive sample, and $x_{k,i}^{(-)}$ denotes the ith feature of the kth negative sample. The larger the F-score is, the more likely this feature is more discriminative. In this work, features are ranked in decreasing order of its importance according to F-score and the number of relevant descriptors is chosen through a sequential backward selection algorithm: starting with all descriptors in the descriptor set, each time one feature with the smallest F-score is removed from the candidate set if the performance of the subset of features is improved.
3) Simulated annealing selection: molecular descriptors relevant to the classification are further reduced by Monte Carlo simulated annealing procedure. Simulated annealing is the simulation of a physical process, ‘annealing’, which involves heating the system to a high temperature and then gradually cooling it down to a preset temperature (e.g., room temperature). During this process, the possible configurations of the samples obey the Boltzmann distribution and hence the low energy states are the most populated at equilibrium. The implementation of Monte Carlo simulated annealing combined with SVM reported here is similar to that described in ref. [48] and can be summarized as follows:

① Giving an initial σ value for the Gaussian kernel function.
② Setting the initial simulation temperature T.
③ Generating a trial solution to the underlying optimization problem; i.e., a MC-SA-SVM model is built based on a random selection of descriptors.
④ Calculating the value of the fitness function, which characterizes the quality of the trial solution to the underlying problem, i.e., the performance of the trial subset.
⑤ Perturbing the trial solution to obtain a new solution and build a new MC-SA-SVM model for the new trial solution.

⑥ Calculating the value of the fitness function Q_{new} for the new trial solution.

⑦ Applying the optimization criteria: If $Q_{old} < Q_{new}$, the new solution is accepted and used to replace the old trial solution; if $Q_{old} > Q_{new}$, the new solution is accepted only if the Metropolis criterion is satisfied; i.e.

$$rnd < \exp(-(Q_{old} - Q_{new}))\Big/T \tag{6}$$

where *rnd* is a random number uniformly distributed between 0 and 1, *T* is a parameter analogous to the temperature in the Boltzmann distribution.

⑧ Lowering the simulation temperature *T* to a predetermined value and return to step ③ until the termination condition is satisfied.

⑨ Systematically adjusting the σ value and going back to step ② until the maximal fitness function is achieved.

After these steps, an optimal subset of molecular descriptors and σ value will be obtained and the final MC-SA-SVM model will give the minimum generalization error.

2.6. SVM method

SVM is based on the structural risk minimization principle of the statistical learning theory [49,50], which consistently shows outstanding classification performance, is less penalized by sample redundancy, and has lower risk for overfitting [51,52]. In linearly separable case, SVM constructs a hyperplane to separate active and non-active classes of compounds with a maximum margin for a given training set (x_i, y_i) $(i = 1, 2, \ldots, l)$, $y_i \in \{-1, +1\}$, *l* is the number of samples in the training set. A compound is represented by a vector x_i composed of its molecular descriptors. The hyperplane is constructed by finding another vector **w** and a parameter *b* that minimizes $\|\mathbf{w}\|^2\big/2$ which satisfies the following conditions:

$$\mathbf{w}\cdot x_i + b \geq +1 \quad \text{for (positive class)} \tag{7}$$

$$\mathbf{w}\cdot x_i + b \leq -1 \quad \text{for (negative class)} \tag{8}$$

where y_i is the class index of compound *i*, **w** is a vector normal to the hyperplane, $|b|\big/\|\mathbf{w}\|$ is the perpendicular distance from the hyperplane to the origin and $\|\mathbf{w}\|^2$ is the Euclidean norm of **w**. Based on **w** and *b*, a given vector $\boldsymbol{x}$ can be classified by:

$$f(\boldsymbol{x}) = \operatorname{sgn}(\mathbf{w}\cdot\boldsymbol{x} + b). \tag{9}$$

A positive or negative $f(\boldsymbol{x})$ value indicates that the vector $\boldsymbol{x}$ belongs to the active or non-active class, respectively.

In nonlinearly separable cases, which frequently occur in classifying compounds of diverse structures [53–60], SVM maps the input vectors into a higher dimensional feature space implicitly using a kernel function $K(\boldsymbol{x}_i, \boldsymbol{x})$. We used the Gaussian radial basis function kernel, which has been extensively used and has consistently shown better performance than other kernel functions [61–63].

$$K(\boldsymbol{x}_i, \boldsymbol{x}) = \exp\left(-\|\boldsymbol{x} - \boldsymbol{x}_i\|^2 \Big/ 2\sigma^2\right) \tag{10}$$

where $\sigma > 0$ is a parameter defining the kernel width. Linear SVM can then applied to this feature space based on the following decision function:

$$f(\boldsymbol{x}) = \operatorname{sgn}\left[\sum_{i}^{l} \alpha_i y_i K(\boldsymbol{x}_i, \boldsymbol{x}) + b\right] \tag{11}$$

where the coefficients α_i and *b* are determined by maximizing the following decision Langrangian expression:

$$L(\mathbf{w}, b, \alpha) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{l} \alpha_i [y_i(\mathbf{w}\cdot\boldsymbol{x}_i + b) - 1] \tag{12}$$

under the Karush–Kuhn–Tucker (KKT) conditions [64], the derivatives of L at the saddle point with respect to the primal variables must vanish

$$\frac{\partial}{\partial b} L(\mathbf{w}, b, \alpha) = 0, \quad \frac{\partial}{\partial \mathbf{w}} L(\mathbf{w}, b, \alpha) = 0. \tag{13}$$

Which lead to $\sum_{i=1}^{l} \alpha_i y_i = 0$ and $\mathbf{w} = \sum_{i=1}^{l} \alpha_i y_i \boldsymbol{x}_i$

By substituting Eq. (14) into Eq. (12) the primal variables can be eliminated and the equations are converted into the Wolfe dual optimization problem [65] for finding multipliers α_i to maximize

$$Q(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j \left(\boldsymbol{x}_i \cdot \boldsymbol{x}_j\right) \tag{14}$$

subject to $\alpha_i \geq 0$, $i = 1, \ldots, l$ and $\sum_{i=1}^{l} \alpha_i y_i = 0$.

From the KKT complementary conditions for all support vectors $\boldsymbol{x}_i$ with $\alpha_i > 0$:

$$\alpha_i \left[y_i \left(\sum_{j}^{l} \alpha_j y_i K\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) + \mathrm{b} \right) - 1 \right] = 0, \quad i = 1, \ldots, l$$

one finds:

$$\sum_{j=1}^{l} \alpha_j y_i K\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) + b = y_i. \tag{15}$$

The threshold *b* can be computed from Eq. (15) for any support vector. This is the so called hard margin SVM classifier. The exponent σ value of the Gaussian kernel in Eq. (10) is optimized by maximizing the generalization ability for the five-fold cross validation.

2.7. k-NN method

In *k*-NN, the Euclidean distance between an unclassified vector $\boldsymbol{x}$ and each individual vector $\boldsymbol{x}_i$ in the training set is measured [65,66]. A total of *k* number of vectors nearest to the unclassified vector $\boldsymbol{x}$ are used to determine the class of that unclassified vector. The class of the majority of *k* nearest neighbors is chosen as the predicted class of the unclassified vector $\boldsymbol{x}$. In this work, the *k*-NN prediction accuracies are estimated through five-fold cross-validation with the same dataset and molecular descriptors selected in the SVM classification model.

2.8. ANN method

ANN is a mathematical tool that can be used to regression and classification, which was originally inspired by the neuron structure in the brain. It consists of a series of nodes (the analogy of neurons) which have multiple connections with other nodes. Our neural network adopts a three-layer architecture which has an input layer consisting of inputs from the molecular descriptors and a bias, a hidden layer containing a number of hidden neurons, and an output layer that outputs the class of a sample [67]. The error back-propagation method using the gradient descent with momentum is used to train the ANN model [68–70]. In this work, the optimal number of neurons in the hidden layer is chosen by maximizing the generalization ability for the five-fold cross-validation. In this work, the ANN prediction accuracies are estimated through five-fold cross-validation with the same dataset and molecular descriptors selected in the SVM classification model.

3. Evaluation of prediction performance

As in the case of all discriminative methods [71,72], the performance of machine learning methods can be evaluated by the quantity of true positive (*TP*: the number of true hPXR activators), true negative (*TN*: the number of true hPXR non-activators), false positive (*FP*: the number of falsely classified hPXR activators), and false negative (*FN*: the number of falsely classified hPXR non-activators). Sensitivity $SE = {}^{TP}\!/_{(TP+FN)}$ and specificity $SP = {}^{TN}\!/_{(TN+FP)}$ are the prediction accuracy for hPXR activators and non-activators, respectively. The overall prediction accuracy (*Q*) and Matthews' correlation coefficient (*C*) [73] are used to measure the overall prediction performance:

$$Q = \frac{TP+TN}{TP+TN+FP+FN} \tag{16}$$

$$C = \frac{TP \times TN - FN \times FP}{\sqrt{(TP+FN)(TP+FP)(TN+FN)(TN+FP)}}. \tag{17}$$

4. Results and discussion

4.1. Feature selection for SVM model

Table 1 gives the SVM performance evaluated by five-fold cross-validation in each step of the feature selection processes. The number of descriptors is promptly decreased from the initial 548 to 159 by the first preprocessing step. By using these 159 descriptors, SVM gives an average prediction accuracy of 89.0% for the hPXR activators, 73.9% for the hPXR non-activators and 83.2% for the all samples. The second ranking step then further reduce the number of descriptors to 145, and the average prediction accuracies of the corresponding SVM model for the hPXR activators, the hPXR non-activators and all samples are 90.7%, 80.4% and 86.7%, respectively. The results indicate that the filter step slightly improves the average predictive accuracies. To further reduce the number of descriptors, MC-SA is applied to reduce it to 73. The selected 73 descriptors are listed in Table 2. This final model is named MC-SA-SVM. The average prediction accuracies for the MC-SA-SVM model are 92.5%, 87.8% and 90.7% for the hPXR activators, the hPXR non-activators and the all samples, respectively. These results indicate that the use of MC-SA feature selection method can further improve the predictive accuracies, and suggest that MC-SA is useful for removing redundant descriptors and helpful for the improvement of computational efficiency of statistical system.

4.2. Model validation through y-scrambling

y-Scrambling was applied to exclude the possibility of chance correlation, i.e., fortuitous correlation without any predictive ability. The classification labels (true, negative) of the 318 compounds in the training set were reordered in a random manner. Afterward, attempts were made to build SVM model with the scrambled activity data. A total of 30 randomization runs were performed. The results of the y-scrambling test are given in (Supporting Information Table S3) of the supporting information. The average accuracies for the hPXR activators, hPXR non-activators and overall samples are 20.3–44.5%, 55.1–63.4% and 51.4–62.6%, respectively. In all cases, the obtained random models have much lower prediction accuracies than the model based on the real data, indicating no obvious chance correlation in the SVM model.

4.3. Application of the selected descriptors to other machine learning approaches

To test whether the selected descriptors are truly relevant to the discrimination between hPXR compounds and hPXR non-activators, the 73 selected descriptors were used to develop ANN and *k*-NN prediction models. The prediction accuracies of these methods and the SVM method are given in Table 3. The prediction accuracies for hPXR compounds, hPXR non-activators, total agents and MCC are between 87.2 and 92.5%, 73.8 and 87.8%, 82.1 and 90.7%, and 0.618 and 0.805, respectively. SVM and ANN were found to outperform *k*-NN. Our study suggests that the descriptors selected by our multi-step

Table 1
Performance of SVM in each step of the feature selection estimated by five-fold cross-validation.

Step[a] (number of descriptors)	Optimal σ[b]	Cross-validation	Prediction for hPXR activators			Prediction for hPXR non-activators			% *Q*	*C*
			TP	FN	% *SE*	TN	FP	% *SP*		
Step 1 (159)	8.0	1	32	2	94.1	15	6	71.4	85.5	0.689
		2	29	5	85.3	17	4	81.0	83.6	0.657
		3	31	4	88.6	16	5	76.2	83.9	0.654
		4	34	0	100.0	14	8	63.6	85.7	0.718
		5	27	8	77.1	17	5	77.3	77.2	0.534
		Average			**89.0**			**73.9**	**83.2**	**0.650**
		SD[c]			8.7			6.7	3.5	0.070
Step 2 (145)	7.0	1	29	5	85.3	15	6	71.4	80.0	0.573
		2	30	4	88.2	18	3	85.7	87.3	0.733
		3	33	2	94.3	18	3	85.7	91.1	0.808
		4	34	0	100.0	17	5	77.3	91.1	0.821
		5	30	5	85.7	18	4	81.8	84.2	0.670
		Average			**90.7**			**80.4**	**86.7**	**0.721**
		SD[c]			6.3			6.1	4.7	0.103
Step 3 (73)	6.0	1	32	2	94.1	17	4	80.9	89.1	0.767
		2	31	3	91.2	19	2	90.5	90.9	0.810
		3	32	3	91.4	19	2	90.5	91.1	0.812
		4	34	0	100.0	20	2	90.9	96.4	0.927
		5	30	5	85.7	19	3	86.4	86.0	0.711
		Average			**92.5**			**87.8**	**90.7**	**0.805**
		SD[c]			5.2			4.2	3.8	0.079

[a] Step 1: Preprocessing; Step 2: Filter step through Fisher-score ranking and backward selection; and Step3: Monte Carlo simulated annealing.
[b] σ: exponent of the Gaussian kernel of SVM.
[c] SD: standard deviation.

Table 2
MC-SA selected 73 molecular descriptors.

Descriptor class	Descriptions	N[a]
Simple molecular properties	Number of H-bond donor, number of 6-member non-aromatic rings, number of atoms, number of O atoms, number of S atoms, average molecular weight (AMW), molecular weight (MW), and number of rings	8
BCUT descriptors	The fifth highest eigenvalue of BCUT descriptors weighted by atomic polarizability, the third lowest eigenvalue of BCUT descriptors weighted by atomic E-state, the fourth lowest eigenvalue of BCUT descriptors weighted by atomic E-state, the third lowest eigenvalue of BCUT descriptors weighted by atomic mass, the third lowest eigenvalue of BCUT descriptors weighted by atomic polarizability, the first highest eigenvalue of BCUT descriptors weighted by atomic polarizability, the fourth lowest eigenvalue of BCUT descriptors weighted by atomic polarizability, the fifth highest eigenvalue of BCUT descriptors weighted by atomic E-state, the second highest eigenvalue of BCUT descriptors weighted by atomic electronegativity, the second highest eigenvalue of BCUT descriptors weighted by atomic polarizability, the third highest eigenvalue of BCUT descriptors weighted by atomic E-state, the third highest eigenvalue of BCUT descriptors weighted by atomic polarizability, the first lowest eigenvalue of BCUT descriptors weighted by atomic mass	13
Electro-topological state	Sum of estate of atom type ssO, sum of estate of atom type dsCH, sum of estate of atom type aasC, sum of H estate of atom type HdsCH, sum of estate of atom type aaaC, sum of estate of atom type dO, sum of estate of atom type ssNH, sum of H estate of atom type HaaCH, sum of estate of atom type aaCH, sum of estate of all C atoms, sum of estate of atom type ssCH2, sum of estate of all halogen atoms	12
Autocorrelation descriptors (2D)	Moreau–Broto autocorrelation of lag9 weighted by atomic E-state indices, Moreau–Broto autocorrelation of lag0 weighted by atomic E-state indices, Moreau–Broto autocorrelation of lag3 weighted by atomic E-state indices, Moreau–Broto autocorrelation of lag7 weighted by atomic E-state indices, Moreau–Broto autocorrelation of lag6 weighted by atomic E-state indices, Moreau–Broto autocorrelation of lag2 weighted by atomic E-state indices, Moran autocorrelation of lag2 weighted by atomic electronegativity, Geary autocorrelation of lag6 weighted by atomic E-state indices, Geary autocorrelation of lag3 weighted by weighted by atomic mass, Geary autocorrelation of lag7 weighted by atomic VDW radius, Geary autocorrelation of lag3 weighted by atomic electronegativity, Geary autocorrelation of lag3 weighted by atomic polarizability, Geary autocorrelation of lag2 weighted by atomic polarizability, Geary autocorrelation of lag2 weighted by atomic electronegativity, Geary autocorrelation of lag2 weighted by atomic VDW, Geary autocorrelation of lag1 weighted by atomic E-state indices, Geary autocorrelation of lag3 weighted by atomic E-state indices, Geary autocorrelation of lag1 weighted by atomic VDW volume, Geary autocorrelation of lag9 weighted by atomic E-state indices, Geary autocorrelation of lag1 weighted by atomic mass, Geary autocorrelation of lag2 weighted by atomic mass	21
Molecular connectivity and shape	Mean topological charge index J2, optimized 1st connectivity index, simple topological index by Narumi, topological charge index G4, molecular path count of length 5, gravitational topological index,1th order delta chi index, dispersion	8
Fingerprint descriptors	Fingerprint for phenol (Ph-OH), fingerprint for 6-member aromatic rings, fingerprint for heterocyclic rings, fingerprint for 6-member non-aromatic rings, fingerprint for containing rings connected by 3 non-ring edges, fingerprint for containing rings connected by 2 non-ring edges, fingerprint for secondary ammonium, fingerprint for tertiary ammonium, fingerprint for fused rings with 3 rings, fingerprint for second alcohol, fingerprint for 5-member non-aromatic rings.	11

[a] The number of molecular descriptors.

hybrid feature selection method in developing SVM hPXR compound prediction model are useful for developing other machine learning models for predicting hPXR compounds. Therefore, these selected descriptors are likely relevant to the classification of hPXR compounds from hPXR non-activators. Moreover, all the developed machine learning models show no apparent over-fitting phenomenon, which frequently occur in the application of wrapper methods (http://www.scss.tcd.ie/publications/tech-reports/reports.05/TCD-CS-2005-17.pdf).

4.4. Model validation through external test set

According to Gobraikh and Tropsha [74], the only way to establish a reliable model is by means of external validation. In the external validation method, the data in the external testing set should not take part in the training, and hence it can measure the prediction ability and check the chance correlation. In this work, all of the models are also validated using external test sets.

Before giving a prediction for a compound in the external testing set, the applicability domain of the models should be considered in advance. The applicability domain of QSAR is defined by the physicochemical, structural, or biological space knowledge on which the training data have been developed, and for which it is applicable to make predictions for new compounds. Ideally the model should only be used to make predictions within that domain by interpolation not extrapolation [75]. One of the approaches of defining the applicability domain is to estimate the training set coverage in the model's descriptor space. In mathematical terms, it means estimation of interpolation regions in the multivariate space of training set, because the

Table 3
Performance of SVM and other machine learning methods using the selected 73 descriptors.

Method	Parameter	Cross-validation	Prediction for hPXR activators			Prediction for hPXR non-activators			% *Q*	*C*
			TP	*FN*	*% SE*	*TN*	*FP*	*% SP*		
k-NN	$k=1$	1	30	4	88.2	16	5	76.2	83.6	0.651
		2	30	4	88.2	15	6	71.4	81.8	0.610
		3	31	4	88.6	15	6	71.4	82.1	0.614
		4	29	5	85.3	19	3	86.4	85.7	0.707
		5	30	5	85.7	14	8	63.6	77.2	0.510
		Average			**87.2**			**73.8**	**82.1**	**0.618**
ANN	$n=37$	1	32	2	94.1	18	3	85.7	90.9	0.806
		2	29	5	85.3	17	4	81.0	83.6	0.657
		3	33	2	94.3	17	4	81.0	89.3	0.769
		4	32	2	94.1	20	2	90.9	92.9	0.850
		5	28	7	80.0	15	7	68.2	75.4	0.482
		Average			**89.6**			**81.4**	**86.4**	**0.713**
SVM	$\sigma=6.0$	1	32	2	94.1	17	4	80.9	89.1	0.767
		2	31	3	91.2	19	2	90.5	90.9	0.810
		3	32	3	91.4	19	2	90.5	91.1	0.812
		4	34	0	100.0	20	2	90.9	96.4	0.927
		5	30	5	85.7	19	3	86.4	86.0	0.711
		Average			**92.5**			**87.8**	**90.7**	**0.805**

Table 4
Comparison of prediction accuracies of different machine learning approaches by external testing set with the selected molecular descriptors.

Method	Parameter	External testing set						% Q	C
		TP	*FN*	*% SE*	*TN*	*FP*	*% SP*		
k-NN	$k=1$	42	6	87.5	30	5	85.7	86.7	0.730
ANN	$n=37$	43	5	89.6	32	3	91.4	90.4	0.805
SVM	$\sigma=6.0$	45	3	93.8	32	3	91.4	92.8	0.852

interpolated prediction results are more reliable than the extrapolated prediction results. This approach is especially suitable to those models based on statistical mining techniques. There are four major approaches (range based, distance based, geometrical and probability density distribution based) to estimate interpolation regions in a multivariate space [76].

In this study, to assess the application domain, i.e., the chemical space covered by the training set of the above hold-out model, the distance based approach is adopted. In this approach, to check if the prediction for a new compound is in the applicability domain, the distance between the compound and the center of the training set is calculated and if this distance exceeds a threshold, which is the largest distance of a training set data point to the center of the training data set, this compound is labeled out of the domain [77].

Since Euclidean distance has been adopted to choose the training set points in order to make these points span the largest chemical space, it is also employed as the distance of assessing applicability domain. Before the selected descriptors of the training set data are employed, they should be scaled, centered and rotated to principal components data. But the compound mirtazapine in the external testing set is out of the application domain.

Table 4 gives the prediction accuracies of the external testing set. As shown in Table 4, the prediction accuracies for hPXR activators, hPXR non-activators and overall samples range between 87.5 and 93.8%, 85.7 and 91.4%, and 86.7 and 92.8%, respectively. The *C* value ranges between 0.730 and 0.852. These results indicate that the classification model is reliable.

4.5. Comparison with literature studies

Table 5 gives an overview on recent results on in silico models of hPXR and compares them with our binary QSAR models. Considering the size of the dataset and the structural diversity of compounds, our results seem acceptable. The *C* value ranges between $+1$ and -1, where a value of $+1$ indicates perfect prediction, -1 represents an inverse prediction, and 0 indicates that the prediction is equivalent to a completely random prediction. The previous best SVM reported [15] had an external test set accuracy with a *C* value of 0.332 as compared with 0.852 in this study. However, it should be noted that direct comparisons of the results by different works may not be very appropriate because of the use of different sets of samples, molecular descriptors, classification methods and parameters, and methods for validation. Nonetheless, a tentative comparison may provide some crude estimate regarding the approximate level of accuracy of the hPXR predictions. From Table 5, we can come to the conclusion that the results of our SVM model are superior to that of the others. Especially, our external testing set gives the best result than the literature results. Therefore, our SVM model is useful for predicting hPXR.

5. Conclusions

Identification of novel hPXR activators from structurally diverse compounds is important for the discovery of drugs with desired metabolic and toxicological profiles. In this study, a hybrid feature selection method, which is a combination of a preprocessing step, a filtering step through ranking of the Fisher scores and a wrapper approach step by Monte Carlo simulated annealing, is developed to select relevant descriptors from a large pool of molecular descriptors for the prediction of the hPXR activation. The optimal subset of descriptors and the optimal model parameter for SVM are obtained based on a five-fold cross-validation. It is shown that the SVM model using the selected descriptors has very good prediction ability, showing that the hybrid feature selection method is an efficient method. The study reveals that the five-fold cross-validation method may be used to optimize the model parameters and select the relevant descriptors to overcome the over-fitting problem and the external test validation method by designing a representative training set may be used to build the final classification model that has improved prediction ability. We can draw a conclusion that the SVM method can be used for high throughput virtual screening to assess hPXR activation.

Acknowledgments

This work was supported by a grant from the Two-Way Support Programs of Sichuan Agricultural University (Project No. 00770117) and by a grant from the National Natural Science Foundation of China (Project No. 20973118).

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.chemolab.2012.05.012.

References

[1] S. Harmsen, I. Meijerman, J.H. Beijnen, J.H. Schellens, The role of nuclear receptors in pharmacokinetic drug–drug interactions in oncology, Cancer Treatment Reviews 33 (2007) 369–380.
[2] S. Ekins, C. Chang, S. Mani, M.D. Krasowski, E.J. Reschly, M. Iyer, V. Kholodovych, N. Ai, W.J. Welsh, M. Sinz, P.W. Swaan, R. Patel, K. Bachmann, Human pregnane X receptor antagonists and agonists define molecular requirements for different binding sites, Molecular Pharmacology 72 (2007) 592–603.
[3] R.E. Watkins, P.R. Davis-Searles, M.H. Lambert, M.R. Redinbo, Coactivator binding promotes the specific interaction between ligand and the pregnane X receptor, Journal of Molecular Biology 331 (2003) 815–828.
[4] R.E. Watkins, J.M. Maglich, L.B. Moore, G.B. Wisely, S.M. Noble, P.R. Davis-Searles, M.H. Lambert, S.A. Kliewer, M.R. Redinbo, 2.1 Å crystal structure of human PXR in complex with the St. John's wort compound hyperforin, Biochemistry 42 (2003) 1430–1438.
[5] R.E. Watkins, S.M. Noble, M.R. Redinbo, Structural insights into the promiscuity and function of the human pregnane X receptor, Current Opinion in Drug Discovery & Development 5 (2002) 150–158.

Table 5
Comparison of prediction accuracies with literature studies.

Study	Method	Training set					External testing set				
		No.	*%SE*	*%SP*	*%Q*	C	No.	*%SE*	*%SP*	*%Q*	C
Chen et al. [14]	SVM[a]	177	84.4	73.6	79.6	0.598	15			53.3–66.7	
Khandelwal et al. [15]	SVM[a]	177	98.98	88.61	94.35	0.888	145	68.29	65.08	66.9	0.332
This study	SVM[b]	279	92.5	87.8	90.7	0.805	83	93.8	91.4	92.8	0.852

[a] 10-fold cross validation.
[b] 5-fold cross validation.

[6] R. Watkins, G.B. Wisely, L.B. Moore, J.L. Collins, M.H. Lambert, S.P. Williams, T.M. Willson, S.A. Kliewer, M.R. Redinbo, The human nuclear xenobiotic receptor PSR: structural determinants of directed promiscuity, Science 292 (2001) 2329–2333.

[7] Y. Xue, L.B. Moore, J. Orans, L. Peng, S. Bencharit, S.A. Kliewer, M.R. Redinbo, Crystal structure of the pregnane X receptor–estradiol complex provides insights into endobiotic recognition, Molecular Endocrinology 21 (2007) 1028–1038.

[8] S. Ekins, J.A. Erickson, A pharmacophore for human pregnane X receptor ligands, Drug Metabolism and Disposition 30 (2002) 96–99.

[9] K. Bachmann, H. Patel, Z. Batayneh, J. Slama, D. White, J. Posey, S. Ekins, D. Gold, L. PXR and the regulation of apoA1 and HDL-cholesterol in rodents, Pharmacological Research 50 (2004) 237–246.

[10] D. Schuster, C. Laggner, T.M. Steindl, A. Palusczak, R.W. Hartmann, T.J. Langer, Pharmacophore modeling and in silico screening for new P450 19 (aromatase) inhibitors, Journal of Chemical Information and Modeling 46 (2006) 1301–1311.

[11] S. Ekins, L. Mirny, E. Schuetz, A ligand-based approach to understanding selectivity of nuclear receptors PXR, CAR, FXR, LXRα and LXRβ, Pharm, PXR, CAR, FXR, LXRα and LXRβ, Pharmaceutical Research 19 (2002) 1788–1800.

[12] M.N. Jacobs, In silico tools to aid risk assessment of endocrine disrupting chemicals, Toxicology 205 (2004) 43–53.

[13] S. Ekins, S. Andreyev, A. Ryabov, E. Kirillov, E.A. Rakhmatulin, S. Sorokina, A. Bugrim, T. Nikolskaya, A combined approach to drug metabolism and toxicity assessment, Drug Metabolism and Disposition 34 (2006) 495–503.

[14] C.Y. Ung, H. Li, C.W. Yap, Y.Z. Chen, In silico prediction of pregnane X receptor activators by machine learning approaches, Molecular Pharmacology 71 (2007) 158–168.

[15] A. Khandelwal, M.D. Krasowski, E.J. Reschly, M.W. Sinz, P.W. Swaan, S. Ekins, Machine learning methods and docking for predicting human pregnane X receptor activation, Chemical Research in Toxicology 21 (2008) 1457–1467.

[16] C.Y. Wang, C.W. Li, J.D. Chen, W.J. Welsh, Interactions in the assembly of the pregnane X receptor/corepressor complex, Molecular Pharmacology 69 (2006) 1513–1517.

[17] I.V. Tetko, P. Bruneau, H.W. Mewes, D.C. Rohrer, G.I. Poda, Can we estimate the accuracy of ADME-Tox predictions? Drug Discovery Today 11 (2006) 700–707.

[18] S. Dimitrov, G. Dimitrova, T. Pavlov, N. Dimitrova, G. Patlewicz, J. Niemela, O. Mekenyan, A stepwise approach for defining the applicability domain of SAR and QSAR models, Journal of Chemical Information and Modeling 45 (2005) 839–849.

[19] R.P. Sheridan, B.P. Feuston, V.N. Maiorov, S.K. Kearsley, Similarity to molecules in the training set is a good discriminator for prediction accuracy in QSAR, Journal of Chemical Information and Computer Sciences 44 (2004) 1912–1928.

[20] M. Dash, K. Choi, P. Scheuermann, H. Liu, Proceedings of the Second International Conference on Data Mining, 2000, pp. 15–122.

[21] M.A. Hall, Correlation-based feature selection for discrete and numeric class machine learning, Proceedings of the 17th International Conference on Machine Learning, 2000, pp. 359–366.

[22] H. Liu, R. Setiono, A probabilistic approach to feature selection—a filter solution, Proceedings of the 13th International Conference on Machine Learning, In 13th International Conference on Machine Learning (ICML'96), 1996, pp. 319–327.

[23] R. Caruana, D. Freitag, Greedy attribute selection, Proceedings of the 11th International Conference on Machine Learning, 1994, pp. 28–36.

[24] J.G. Dy, C.E. Brodley, Feature subset selection and order identification for unsupervised learning, Proceedings of the 17th International Conference on Machine Learning, 2000, pp. 247–254.

[25] Y. Kim, W. Street, F. Menczer, Feature selection for unsupervised learning via evolutionary search, Proceedings of the Sixth ACM SIGKDD International Conference Knowledge Discovery and Data Mining, 2000, pp. 365–369.

[26] Y. Leung, Y. Hung, A multiple-filter–multiple-wrapper approach to gene selection and microarray data classification, IEEE/ACM Transactions on Computational Biology and Bioinformatics 7 (2010) 108–117.

[27] H. Li, C.Y. Ung, C.W. Yap, Y. Xue, Z.R. Li, Z.W. Cao, Y.Z. Chen, Prediction of genotoxicity of chemical compounds by statistical learning, Chemical Research in Toxicology 18 (2005) 1071–1080.

[28] S.W. Chen, Z.R. Li, X.Y. Li, Prediction of antifungal activity by support vector machine approach, Journal of Molecular Structure (THEOCHEM) 731 (2005) 73–81.

[29] R.M .Balabin, S.V. Smirnov, Variable selection in near-infrared spectroscopy: benchmarking of feature selection methods on biodiesel data, 692 (1–2) (2011) 63–72.

[30] R. Kohavi, G.H. John, Wrappers for feature subset selection, Artificial Intelligence 97 (1997) 273–324.

[31] H. Kojima, F. Sata, S. Takeuchi, T. Sueyoshi, T. Nagai, Comparative study of human and mouse pregnane X receptor agonistic activity in 200 pesticides using in vitro reporter gene assays, Toxicology 280 (2011) 77–87.

[32] C. Benod, G. Subra, V. Nahoum, A. Mallavialle, J.F. Guichou, J. Milhau, S. Roblés, W. Bourguet, J.M. Pascussi, P. Balaguer, A. Chavanieu, N-1H-Benzimidazol-5-ylbenzenesulfonamide derivatives as potent hPXR agonists. Bioorganic & Medicinal Chemistry 7 (2008) 3537–3549.

[33] Hyperchem Release 7.0(Beta 1.04)software available at http://www.hyper.com.

[34] J.J. Perez, Managing molecular diversity, Chemical Society Reviews 34 (2005) 143–152.

[35] P. Willett, J.M. Barnard, G.M. Downs, Chemical similarity searching, Journal of Chemical Information and Computer Sciences 38 (1998) 983–996.

[36] P. Willett, V.A. Winterman, Comparison of some measures for the determination of intermolecular structural similarity, Quantitative Structure–Activity Relationships 5 (1986) 18–25.

[37] W. Wu, B. Walczak, D.L. Massart, S. Heuerding, F. Erni, I.R. Last, K.A. Pebble, Artificial neural networks in classification of NIR spectral data: design of the training set, Chemometrics and Intelligent Laborary Systems 33 (1996) 35–46.

[38] R. Todeschini, V. Consonni, Handbook of Molecular Descriptors, Wiley-VCH, Weinheim, 2000, pp. 50–70.

[39] A.R. Katritzky, E.V. Gordeeva, Electronic, geometrical, and combined molecular descriptors in QSAR/QSPR research, Journal of Chemical Information and Computer Sciences 33 (1993) 835–857.

[40] L.B. Kier, L.H. Hall, Molecular Structure Description: The Electrotopological State, Academic Press, San Diego, 1999.

[41] M. Karelson, V.S. Lobanov, A.R. Katritzky, Quantum-chemical descriptors in QSAR/QSPR studies, Chemical Reviews 96 (1996) 1027–1043.

[42] Z.R. Li, L.Y. Han, Y. Xue, C.W. Yap, H. Li, L. Jiang, Y.Z. Chen, Model-molecular descriptor lab: a web-based sever for computing structural and physicochemical feature of compounds, Biotechnology and Bioengineering 97 (2007) 389–396.

[43] H.P. Schultz, Topological organic chemistry. 1. Graph theory and topological indices of alkanes, Journal of Chemical Information and Computer Sciences 29 (1989) 227–228.

[44] L.H. Hall, L.B. Kier, Electrotopological state indices for atom types: a novel combination of electronic, topological and valence state information, Journal of Chemical Information and Computer Sciences 35 (1995) 1039–1045.

[45] R.S. Pearlman, K.M. Smith, Novel software tools for chemical diversity, Perspectives in Drug Discovery and Design 9–11 (1998) 339–353.

[46] J. Caballero, F.D.F.M. Gonza'lez-Nilo, Structural requirements of pyrido[2,3-d] pyrimidin-7-one as CDK4/D inhibitors: 2D autocorrelation, CoMFA and CoMSIA analyses, Bioorganic & Medicinal Chemistry 16 (2008) 6103–6115.

[47] Y.W. Chen, C.J. Lin, Combining SVMs with various feature selection strategies, 2005. Available at http://www.csie.ntu.edu.tw/~cjlin/papers/features.pdf.

[48] S. Ajmani, K. Jadhav, S.A. Kulkarni, Three-dimensional QSAR using the k-nearest neighbor method and its interpretation, Journal of Chemical Information and Modeling 46 (2006) 24–31.

[49] V.N. Vapnik, The Nature of Statistical Learning Theory, Springer, New York, 1995.

[50] C.J.C. Burges, A tutorial on support vector machines for pattern recognition, Data Mining and Knowledge Discovery 2 (1998) 127–167.

[51] N. Pochet, F. De Smet, J.A. Suykens, B.L. De Moor, Systematic benchmarking of microarray data classification: assessing the role of nonlinearity and dimensionality reduction, Bioinformatics 20 (2004) 3185–3195.

[52] F. Li, Y. Yang, Analysis of recursive gene selection approaches from microarray data, Bioinformatics 21 (2005) 3741–3747.

[53] R.N. Jorissen, M.K. Gilson, Virtual screening of molecular databases using a support vector machine, Journal of Chemical Information and Modeling 45 (2005) 549–561.

[54] Z. Lepp, T. Kinoshita, H. Chuman, Screening for new antidepressant leads of multiple activities by support vector machines, Journal of Chemical Information and Modeling 46 (2006) 158–167.

[55] M. Glick, J.L. Jenkins, J.H. Nettles, H. Hitchings, J.W. Davies, Enrichment of high-throughput screening data with increasing levels of noise using support vector machines, recursive partitioning, and Laplacian-modified naive Bayesian classifiers, Journal of Chemical Information and Modeling 46 (2006) 193–200.

[56] J. Hert, P. Willett, D.J. Wilton, P. Acklin, K. Azzaoui, E. Jacoby, A. Schuffenhauer, New methods for ligand-based virtual screening: use of data fusion and machine learning to enhance the effectiveness of similarity searching, Journal of Chemical Information and Modeling 46 (2006) 462–470.

[57] C.W. Yap, Y.Z. Chen, Quantitative structure–pharmacokinetic relationships for drug distribution properties by using general regression neural network, Journal of Pharmaceutical Sciences 94 (2005) 153–168.

[58] J. Cui, L.Y. Han, H.H. Lin, H.L. Zhang, Z.Q. Tang, C.J. Zheng, Z.W. Cao, Y.Z. Chen, Prediction of MHC-binding peptides of flexible lengths from sequence-derived structural and physicochemical properties, Molecular Immunology 44 (2007) 866–877.

[59] C.W. Yap, Y.Z. Chen, Prediction of cytochrome P450 3A4, 2D6, and 2C9 inhibitors and substrates by using support vector machines, Journal of Chemical Information and Modeling 45 (2005) 982–992.

[60] M. Grover, B. Singh, M. Bakshi, S. Singh, Quantitative structure–property relationships in pharmaceutical research — part 2, Pharmaceutical Science & Technology Today 3 (2000) 50–57.

[61] M.W.B. Trotter, B.F. Buxton, S.B. Holden, Support vector machines in combinatorial chemistry, Measurement and Control 34 (2001) 235–239.

[62] R. Burbidge, M. Trotter, B. Buxton, S. Holden, Drug design by machine learning: support vector machines for pharmaceutical data analysis, Computers & Chemistry 26 (2001) 5–14.

[63] R. Czerminski, A. Yasri, D. Hartsough, Use of support vector machine in pattern classification: application to QSAR studies, Quantitative Structure–Activity Relationships 20 (2001) 227–240.

[64] P. Bertsekas, Nonlinear Programming, Athena Scientific, Belmont, MA, 1995.

[65] E. Fix, J.L. Hodges, Discriminatory Analysis: Non-Parametric Discrimination: Consistency Properties, USAF School of Aviation Medicine, Randolph Field, Texas, 1951.

[66] R.A. Johnson, D.W. Wichern, Applied Multivariate Statistical Analysis, Prentice Hall, Englewood Cliffs, NJ, 1982.

[67] I.V. Tetko, V. Yu, N.P. Tanchuk, N.P. Chentsova, S.V. Antonenko, G.I. Poda, V.P. Kukhar, A.I. Luik, HIV-1 reverse transcriptase inhibitor design using artificial neural networks, Journal of Medicinal Chemistry 37 (1994) 2520–2526.

[68] D.E. Rumelhart, G.E. Hinton, R.J. Williams, in: D.E. Rumelhart, J.L. McClelland (Eds.), Parallel Distributed Processing: Explorations in the Microstructure of Cognition, MIT Press, Cambridge, MA, 1986.

[69] J. Zupan, J. Gasteiger, Neural Networks for Chemistry and Drug Design: An Introduction, 2nd ed. VCH, Weinheim, 1999.

[70] I.V. Tetko, A.I. Luik, G.I. Poda, Application of neural networks in structure–activity relationships of a small number of molecules, Journal of Medicinal Chemistry 36 (1993) 811–814.

[71] P. Baldi, S. Brunak, Y. Chauvin, C.A. Andersen, H. Nielsen, Assessing the accuracy of prediction algorithms for classification: an overview, Bioinformatics 16 (2000) 412–424.

[72] J.E. Roulston, Screening with tumor markers, Molecular Pharmacology 20 (2002) 153–162.

[73] B.W. Matthews, Comparison of the predicted and observed second of T4 phage lysozyme, Biochimica et Biophysica Acta 405 (1975) 442–451.

[74] A. Golbraikh, A. Tropsha, Beware of q2! Journal of Molecular Graphics & Modelling 20 (2002) 269–276.

[75] N. Nikolova-Jeliazkova, J. Jaworska, An approach to determining applicability domains for QSAR group contribution models: an analysis of SRC KOWWIN, Alternatives to Laboratory Animals 33 (2005) 461–470.

[76] http://ecb.jrc.ec.europa.eu/documents/QSAR/INFORMATION_SOURCES/applicability_domain_overview.pdf.

[77] J. Jaworska, N. Nikolova-Jeliazkova, T. Aldenberg, QSAR applicability domain estimation by projection of the training set in descriptor space: a review, Alternatives To Laboratory Animals 33 (2005) 445–459.