Available online at www.sciencedirect.com

ScienceDirect

Journal of Molecular Graphics and Modelling 25 (2006) 313–323

Journal of Molecular Graphics and Modelling

www.elsevier.com/locate/JMGM

Prediction of estrogen receptor agonists and characterization of associated molecular descriptors by statistical learning methods

H. Li [a], C.Y. Ung [a,b], C.W. Yap [a], Y. Xue [c], Z.R. Li [c], Y.Z. Chen [a,d,*]

[a] *Bioinformatics and Drug Design Group, Department of Pharmacy, National University of Singapore, Blk SOC1, Level 7, 3 Science Drive 2, Singapore 117543, Singapore*

[b] *Department of Biochemistry, the Yong Loo Lin School of Medicine, National University of Singapore, Blk MD7, #02-03, 8 Medical Drive, Singapore 117597, Singapore*

[c] *College of Chemistry, Sichuan University, Chengdu 610064, PR China*

[d] *Shanghai Center for Bioinformation Technology,100 Qinzhou Road, Shanghai 200235, PR China*

Received 19 October 2005; received in revised form 21 December 2005; accepted 19 January 2006
Available online 23 February 2006

Abstract

Specific estrogen receptor (ER) agonists have been used for hormone replacement therapy, contraception, osteoporosis prevention, and prostate cancer treatment. Some ER agonists and partial-agonists induce cancer and endocrine function disruption. Methods for predicting ER agonists are useful for facilitating drug discovery and chemical safety evaluation. Structure–activity relationships and rule-based decision forest models have been derived for predicting ER binders at impressive accuracies of 87.1–97.6% for ER binders and 80.2–96.0% for ER non-binders. However, these are not designed for identifying ER agonists and they were developed from a subset of known ER binders. This work explored several statistical learning methods (support vector machines, k-nearest neighbor, probabilistic neural network and C4.5 decision tree) for predicting ER agonists from comprehensive set of known ER agonists and other compounds. The corresponding prediction systems were developed and tested by using 243 ER agonists and 463 ER non-agonists, respectively, which are significantly larger in number and structural diversity than those in previous studies. A feature selection method was used for selecting molecular descriptors responsible for distinguishing ER agonists from non-agonists, some of which are consistent with those used in other studies and the findings from X-ray crystallography data. The prediction accuracies of these methods are comparable to those of earlier studies despite the use of significantly more diverse range of compounds. SVM gives the best accuracy of 88.9% for ER agonists and 98.1% for non-agonists. Our study suggests that statistical learning methods such as SVM are potentially useful for facilitating the prediction of ER agonists and for characterizing the molecular descriptors associated with ER agonists.

Keywords: Estrogen receptor (ER); Estrogen receptor agonists; Statistical learning methods (SLMs); Support vector machine (SVM); Molecular descriptors; Classification

1. Introduction

Estrogen receptors (ERs) are members of nuclear receptor family play important roles in cell growth, development, and homeostasis processes in various tissues. There are two ER isoforms. ERα is primarily expressed in uterus, vagina, liver, and pituitary. ERβ is mainly expressed in ovary, prostate, epididymis, lung, hypothalamus, and bladder. Both ERα and ERβ share modest overall sequence identity (~47%) with highly conserved regions in the DNA-binding and ligand-binding domains and low homology at the N-terminal transactivation domain [1]. The C-terminal of the ligand-binding domain contains regions for ER dimerization and for recruiting transcriptional coactivators [2].

ER ligands are of three types: agonists, antagonists and selective ER modulators [3]. ER modulators can act as either agonists or antagonists depending on the cellular and promoter context as well as the ER isoforms [4]. In particular, ER agonists have been used as drugs for hormone replacement therapy, contraception, prevention of osteoporosis [5], and for the treatment of metastatic prostate cancer [6]. ER agonists also show neuroprotective actions both dependent and independent of ER activity [7] and are reported to have beneficial cardiovascular effects [8]. Apart from these beneficial

* Corresponding author. Tel.: +65 6874 6877; fax: +65 6779 1554.
E-mail address: yzchen@cz3.nus.edu.sg (Y.Z. Chen).

1093-3263/$ – see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2006.01.007

therapeutic applications, a number of environmental ER agonists and partial-agonists, produced as industrial compounds and pesticides, are known to disrupt human endocrine functions by mimicking endogenous estrogens [9]. Exposure to environmental ER agonists and partial agonists has been proposed to be a risk factor for the disruption of reproductive system development and tumorigenesis in humans [10].

As part of the effort for developing fast and low-cost tools for facilitating drug design and chemical safety evaluations, a few statistical learning methods have been used for computer prediction of ER binders [11–14]. These methods use specific structural and physicochemical properties of the known ER binders and non-binders to statistically derive structure–activity relationships (SAR) [11,13], quantitative structure–activity relationships (QSAR) [12,14], and the rule-based decision forest models [15] for predicting ER binding potential of a molecule. The prediction accuracies of these methods are in the range of 87.1–97.6% for ER binders and 80.2–96.0% for ER non-binders, which are at a useful level for facilitating the prediction of ER binders and non-binders.

However, prediction of ER binders does not automatically enable the identification of ER agonists. Moreover, the current ER binder prediction models have primarily been developed by using compounds that are significantly less in number and narrower in structural diversity than the currently known ER binders and non-binders. For instance, the largest number of ER binders used in previous study is 130 with a structure diversity index (DI) value of 0.645, which is compared to the 243 known ER agonists with a DI value of 0.598. Therefore, it is desirable to develop methods for predicting ER agonists from a more diverse set of ER agonists and ER non-agonists.

Several statistical learning methods (SLMs) have been explored for the prediction of various pharmacodynamic, pharmacokinetic and toxicological classes of chemical agents including drug-like molecules [16,17], *p*-glycoprotein substrates [18], CNS drugs [19], genotoxic agents [20,21], torsade-causing drugs [22], and agents of other specific pharmacokinetic properties [23]. The most widely used SLMs are support vector machines (SVM) [24,25], k nearest neighbor (k-NN) [26], probabilistic neural network (PNN) [27] and C4.5 decision tree (C4.5 DT) [28]. These methods have been shown to be particularly useful for predicting compounds of diverse structures, and some of them consistently show better prediction performance than those of other statistical learning methods. Moreover, they can be used to determine molecular descriptors.

This work evaluated the capability of the most widely used SLMs for predicting ER agonists by using a significantly larger number and more diverse range of ER agonists and non-agonists than in previous studies. A comprehensive literature search was conducted to collect diverse set of literature-reported ER agonists and non-agonists. A feature selection method, recursive feature elimination (RFE), which has been used for extracting molecular descriptors relevant to specific types of pharmaceutical agents [29–31], was used for selecting molecular descriptors relevant to the prediction of ER agonists. Two evaluation methods were used to objectively assess the performance of these methods. One is five-fold cross validation and the other one is validation by the use of an independent validation set of known ER agonists and non-agonists.

This study is focused on the classification of ER agonist versus non-agonist regardless of ER isoforms. ER agonists refer to compounds that show estrogenic activity to at least one of the ER isoforms regardless of whether it has antagonist activity to another isoform. ER agonists can be found from several sources including endogenous estrogens such as estradiol from human body, phytoestrogens such as genistein [32] from plants, mycoestrogens such as α-zearalenol [32] from fungi, xenoestrogens such as chlorothalonil and *o*,*p*′-DDT [33] from pesticides and environmental pollutants, and drug leads or candidates such as diphenolic azoles. ER non-agonists include all ER non-binders and ER antagonists.

X-ray crystallography studies have shown that agonist and pure antagonist has different binding modes in the ligand-binding domain [34]. Antagonists can be divided into ‘‘active’’ type that induces conformational change of the ligand-binding domain and ‘‘passive’’ type that lacks some bulky side chain [35]. These distinguishing features are determined by the structural and physicochemical properties of the compounds, which can be exploited by SLMs for separating ER agonists from ER non-agonists. Moreover, molecular descriptors associated with these features, which are responsible for separating ER agonists from non-agonists, can be extracted by means of feature selection methods [36,37]. Some of these molecular descriptors have been found and extensively used for deriving SAR [11,13], QSAR [12,14] and the rule-based decision forest models [15]. It is likely that not all of these molecular descriptors have been found in previous studies due to the limited coverage of compounds. Therefore, in addition to the improvement of the performance of SLMs, our feature selection method can be used for finding additional molecular descriptors to complement earlier studies.

2. Methods

2.1. Collection of ER agonists and ER non-agonists

Data for ER agonists and non-agonists were collected from several sources including National Center for Toxicological Research Estrogen Receptor Binding Database (NCTRER) (http://www.epa.gov/nheerl/dsstox/sdf_nctrer.html), Endocrine Disruptor Knowledge Base (EDKB) (http://edkb.fda.gov/databasedoor.html), and other publications [11,38,39]. Many of the compounds in the NCTR ER database and EDKB are given as ER binders without specific description whether or not they are ER agonists. Thus, additional literature search is conducted to confirm their ER agonistic status. Those ER binders reported to be as ER antagonists and non-binders are classified as ER non-agonists. A total of 243 ER agonists and 463 ER non-agonists are collected and used in this study. A complete list of these compounds is given in Table S1 in the supplementary material.

The 2D structure of each of the compounds were generated by using ChemDraw [40] and DS ViewerPro 5.0 [41], and were subsequently converted into 3D structure by using CONCORD [42] followed by optimization using the semi-empirical AM1 method [43]. All the generated geometries had been fully optimized without symmetry restrictions. The 3D structure of each compound was manually inspected to ensure that the chirality of each chiral agent is properly generated.

There are 130 compounds with chiral centers in our collected dataset. The active enantiomer of each of these compounds was selected on the basis of literature reports [11,44]. In some cases, an enantiomer is a full agonist of both ER isoforms, while another enantiomer is a full agonist of one isoform and weaker or non-agonist of the other isoform. For instance, S-indenestrol is a full ERα/β agonist but R-indenestrol only shows full agonism to ERβ [44]. In this study, the transactivation criteria of agonists were used rather than their binding activities. For those compounds where the transactivation activities have been determined by using racemate mixtures with no active enantiomers reported, the default enantiomer structure in the chemical database such as PubChem [45] and ChemFinder [46] was straightforwardly used.

2.2. Structural diversity

Structural diversity of these 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 dataset [47]:

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

where sim(*i*,*j*) is a measure of the similarity between compound *i* and *j*, and *N* is the number of compounds in the dataset. The structural diversity of a dataset increases with decreasing DI value. In this work, sim(*i*,*j*) is computed by using the Tanimoto coefficient [48]:

$$\mathrm{sim}(i,j) = \frac{\sum_{d=1}^{l} x_{di}x_{dj}}{\sum_{d=1}^{l}(x_{di})^2 + \sum_{d=1}^{l}(x_{dj})^2 - \sum_{d=1}^{l} x_{di}x_{dj}}$$

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

DI values for current work with larger dataset and from previous ER binder studies [11,13] are 0.598 and 0.645, respectively. In addition, the DI values of three other ER-binder classes (flavanones, steroids with a phenolic ring, and flavones) [11] are 0.740, 0.771, 0.816, respectively, showing low structural diversity. The DI values suggest that the dataset of ER agonists used in this work is more diverse than those used in earlier studies for ER binders.

2.3. Construction of training and testing sets

ER agonists and non-agonists were further divided into training and testing sets by two different methods, five-fold cross validation and validation by an independent evaluation set. For five-fold cross-validation, a group of 243 ER agonists and that of 463 ER non-agonists was randomly divided into five subsets of approximately equal size respectively. Four of the subsets were used as the training set, and the remaining subset was used as the testing set for the ER agonists and non-agonists. This process was repeated five times such that every subset is used as the test set once. For the independent evaluation set, these compounds were divided into training and independent validation set based on their distribution in the chemical space, which is defined by the commonly used structural and chemical descriptors [49]. Compounds of similar structural and chemical features were evenly assigned into separate sets. For those compounds having insufficient number of structurally and chemically similar counterparts, they were assigned, in the order of priority, to the training and then the independent validation set, respectively. The training set was used for the development of the prediction system and the independent evaluation set is used for the assessment of the system. The training and independent evaluation set contains 626 compounds (216 ER agonists, 410 ER non-agonists), and 80 compounds (27 ER agonists, 53 ER non-agonists), respectively. An additional set of 11 ER agonists and 16 ER non-binders, obtained additional search of literatures such as newly published papers, were used to further evaluate the performance of SLM.

2.4. Molecular descriptors

Molecular descriptors are quantitative representations of structural and physicochemical features of molecules, which have been extensively used in the SAR [11,13], QSAR [12] and other statistical learning studies of pharmaceutical agents [16–20,22,23,50]. Major classes of molecular descriptors are simple molecular properties, molecular connectivity and shape, electro-topological state, quantum chemical and geometrical properties.

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 [18,36]. 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 electro-topological 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. The irrelevant and redundant descriptors to ER agonistic activities are further eliminated by using feature selection methods [29,31,51].

2.5. Feature selection method

The recursive feature elimination (RFE) method was used in this work as the feature selection method for selecting molecular descriptors associated to ER agonists. RFE has gained popularity due to its effectiveness for improving prediction performance and for discovering informative features associated with drug activity [29,31] and pharmacokinetic and toxicological properties [18,30]. An agent is represented by a vector $\mathbf{x}_i$, with its molecular descriptors (or features) as the components. The task of selecting appropriate molecular descriptors can be conducted by ranking and selecting those with higher contributions to a particular drug classification problem.

Descriptor ranking in RFE is based on the magnitude of the change of an objective function of a statistical learning model (which roughly measure the extent of contribution of each feature to the prediction capability of the statistical model) upon removing each descriptor [52]. The prediction capability of a statistical learning model is more significantly affected by a larger 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 zero, the change $DJ(i)$ in the cost function J is computed by $DJ(i) = \frac{1}{2}\frac{\partial^2 J}{\partial w_i^2}(Dw_i)^2$, where w_i is the weight of the feature I, and the change in weight $Dw_i = w_i$ corresponds to the removed descriptor x_i. One or more of descriptors with the smallest $DJ(i)$ can be eliminated in each iteration [31].

2.6. Statistical learning methods

2.6.1. Support vector machine (SVM)

Linear SVM constructs a hyperplane separating two different classes of feature vectors with a maximum margin [53]. This hyperplane is constructed by finding a vector $\mathbf{w}$ and a parameter b that minimizes$\|\mathbf{w}\|^2$ which satisfies the following conditions: $\mathbf{w}\cdot\mathbf{x}_i + b \geq +1$, for $y_i = +1$ (positive class) and $\mathbf{w}\cdot\mathbf{x}_i + b \leq -1$, for $y_i = -1$ (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, $|b|/\|\mathbf{w}\|$ is the perpendicular distance from the hyperplane to the origin and $\|\mathbf{w}\|^2$ is the Euclidean norm of $\mathbf{w}$. Nonlinear SVM projects feature vectors into a high dimensional feature space by using a kernel function such as$K(\mathbf{x}_i, \mathbf{x}_j) = e^{-\|\mathbf{x}_j-\mathbf{x}_i\|^2/2\sigma^2}$. 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[$(\mathbf{w}\cdot\mathbf{x}) + b$], a positive or negative value indicates that the vector $\mathbf{x}$ belongs to the positive or negative class, respectively.

2.6.2. k-NN

In k-NN, the Euclidean distance between an unclassified vector $\mathbf{x}$ and each individual vector $\mathbf{x}_i$ in the training set is measured [26,54]. A total of k number of vectors nearest to the unclassified vector $\mathbf{x}$ are used to determine the class of that unclassified vector. The class of the majority of the k nearest neighbors is chosen as the predicted class of the unclassified vector $\mathbf{x}$.

2.6.3. Probabilistic neural network (PNN)

PNN is a form of neural network that uses Bayes optimal decision rule for classification [27]. Traditional neural networks such as feed-forward back-propagation neural network rely on multiple parameters and network architectures to be optimized. In contrast, PNN only has a single adjustable parameter, a smoothing factor σ for the radial basis function in the Parzen's nonparameteric estimator. Thus the training process of PNN is usually orders of magnitude faster than those of the traditional neural networks.

2.6.4. C4.5 decision tree (DT)

C4.5 DT is a branch-test-based classifier [28]. A branch of the decision tree corresponds to a group of classes and a leaf represents a specific class. A decision node specifies a test on a single attribute value, with one branch and its subsequent classes as possible outcomes. C4.5 decision tree uses recursive partitioning to examine every attribute of the data and rank them according to their ability to partition the remaining data, thereby constructing a decision tree. A vector $\mathbf{x}$ is classified by starting at the root of the tree and moving through the tree until a leaf is encountered. At each non-leaf decision node, a test is conducted to move into a branch. Upon reaching the destination leaf, the class of the vector $\mathbf{x}$ is predicted to be that of the leaf.

2.7. Measurement of prediction accuracy

As in the case of all discriminative methods [55,56], the performance of statistical learning methods can be measured by the quantity of true positives TP (true ER agonists), true negatives TN (true non-agonists), false positives FP (false ER agonists), false negatives FN (false non-agonists). Sensitivity, SE = TP/(TP + FN) is the prediction accuracy for the ER agonists. Specificity, SP = TN/(TN + FP) is the prediction accuracy for the ER non-agonists. The overall prediction accuracy (Q) and Matthews correlation coefficient (C) [57] are used to measure the prediction accuracies and are given as:

$$Q = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{1}$$

$$C = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FN} \times \mathrm{FP}}{\sqrt{(\mathrm{TP} + \mathrm{FN})(\mathrm{TP} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})}} \tag{2}$$

2.8. Computational parameters and performance evaluation

Apart from the default parameters, there are no special parameters need to be optimized for DT. For SVM, k-NN and PNN, there is only one parameter in each classification system that needs to be optimized. For these four SLMs, there is only one possible classification system for a given set of parameters. This is different from stochastic methods, such as artificial neural networks (ANNs), which produce different classification models under the same set of parameters because of the use of a random number seed in generating the stochastic models.

The classification speed of these SLMs is in the order of a few thousands to hundreds of thousands of compounds per second [58]. DT has the fastest classification speed because it uses a simple set of rules to reach a decision leaf. The

Table 1
The accuracy of ER agonists and ER non-agonists derived from SVM without the use of a feature selection method (SVM) and from SVM with the use of the feature selection method RFE (SVM + RFE) by using five-fold cross validation

Method	Cross validation	ER agonists			ER non-agonists			*Q* (%)	*C*
		TP	FN	SE (%)	TN	FP	SP (%)		
SVM	1	39	13	75.0	78	13	85.7	80.1	0.607
	2	35	13	72.9	89	14	86.4	78.9	0.590
	3	36	7	83.7	79	11	87.8	86.5	0.700
	4	43	12	78.2	71	15	82.6	80.9	0.602
	5	38	7	84.4	80	13	86.0	85.5	0.684
Average				78.8			85.7	82.4	0.637
S.E.				±2.29			±0.85	±1.52	±0.02
SVM + RFE	1	42	10	80.8	82	9	90.1	86.7	0.712
	2	40	8	83.3	93	10	90.3	88.1	0.728
	3	38	5	88.4	83	7	92.2	91.0	0.797
	4	48	7	87.3	77	9	89.5	88.7	0.763
	5	41	4	91.1	87	6	93.5	92.8	0.837
Average				86.2			91.1	89.5	0.767
S.E.				±1.84			±0.75	±1.09	±0.02

The results are given in TP (true positive), FN (false negative), TN (true negative), FP (false positive), *Q* (overall accuracy), *C* (Matthews correlation coefficient), SE (sensitivity or prediction accuracy for ER agonists) and SP (specificity or prediction accuracy for ER non-agonists). Statistical significance is indicated by S.E. (standard error). The number of ER agonists or ER non-agonists is TP + FN or TN + FP.

classification speed of SVM is usually 25–55% faster than that of k-NN and PNN due to the fact that SVM typically uses 45–75% of the training set as support vectors for classification, whereas k-NN and PNN use the whole training set.

SLMs generally require a sufficient number of samples to develop a classification system. Irrelevant molecular descriptors may reduce the prediction accuracies of these classification systems [18,30,58–61]. SVM has been found to be least sensitive to data over-fitting, even in the cases when a large number of redundant and overlapping molecular descriptors are used [53]. This is because SVM is based on the structural risk minimization principle, which minimizes both training error and generalization error simultaneously.

SVM, k-NN, PNN, and DT do not explicitly provide information about the importance of each molecular descriptor. For SVM, this problem is further compounded when kernel function is used as there is no simple method to inversely map the solution back into the input space. Incorporation of feature selection methods [36,37] and regression methods [23] have been frequently used for extracting important molecular descriptors from these SML-based prediction systems.

3. Results and discussion

3.1. Overall prediction accuracies and merit of the statistical learning methods

The effect of feature selection method on the performance of SLMs for the prediction of ER agonists can be shown by comparing the computed accuracies of SVM with and without the use of RFE, which are shown in Table 1. The accuracies of SVM with RFE are 86.2% for ER agonists and 91.1% for ER non-agonists, which are substantially better than those of 78.8% for ER agonists and 85.7% for ER non-agonists derived from SVM without RFE. Similar prediction accuracies are found in two additional five-fold cross validation studies conducted by using training-testing sets separately generated from different random number seed parameters. This suggests that RFE is useful in selecting the proper set of molecular descriptors for the prediction of ER agonists as well as other classes of pharmaceutical agents [29–31]. The results show that selection of appropriate molecular descriptors is not only important for the improvement of prediction accuracy but more importantly to provide insight into physicochemical nature as well as the molecular mechanism of the action of ER agonists.

Table 2 gives the prediction accuracies of ER agonists and ER non-agonists derived from other three statistical learning methods k-NN, PNN and C4.5 DT by using the RFE selected descriptors and five-fold cross validation method. For comparison, those from SVM are also included in Table 2. The prediction accuracies from the other three methods are comparable to each other. For ER agonists, the accuracies of these methods are in the range of 66.3–86.2% with SVM giving the best accuracy at 86.2%. For ER non-agonists, the accuracies

Table 2
Comparison of the prediction accuracies of ER agonists and ER non-agonists derived from different statistical learning methods by using five-fold cross validation in this work

Method	ER agonists SE (%)	ER non-agonists SP (%)	*Q* (%)
C4.5 DT	66.3	83.8	77.8
PNN	83.6	76.0	78.7
k-NN	72.7	85.9	81.5
SVM + RFE	86.2	91.1	89.5

The methods include C4.5 DT (C4.5 decision tree), PNN (probabilistic neural network), k-NN (k nearest neighbor), SVM + RFE (support vector machine and recursive feature elimination).

Table 3
Comparison of the ER agonists and ER non-agonists prediction accuracies by using SVM with two different validation method, five-fold cross validation and independent validation set

SVM with five-fold cross validation			SVM with independent validation set						
SE (%)	SP (%)	*Q* (%)	TP	FN	SE (%)	TN	FP	SP (%)	*Q* (%)
86.2	91.1	89.5	24	3	88.9	52	1	98.1	95.0

The results are given in TP (true positive), FN (false negative), TN (true negative), FP (false positive), *Q* (overall accuracy), SE (sensitivity) which is the prediction accuracy for ER agonists and SP (specificity) which is the prediction accuracy for ER non-agonists. The number of ER agonists or ER non-agonists is TP + FN or TN + FP.

from these methods are in the range of 76.0–91.1% with SVM giving the best accuracy at 91.1%.

A frequently used method for checking whether a prediction system is over-fitting is to compare the prediction accuracies determined by using cross validation methods with those determined by using independent validation sets [62]. Since descriptor selection was performed by using the cross validation method as the modeling testing sets, an over-fitted classification system is expected to have much higher prediction accuracy for the cross validation sets than that for the independent validation sets. As shown in Table 3, the prediction accuracies of the SVM systems based on the five-fold cross validation method and those based on independent validation sets are similar. This shows that the SVM classification systems in this work are unlikely to be over-fitted.

It had been shown that chance of correlations may occur during descriptor selection especially if the number of descriptors available for selection is large [63,64]. Y-randomization has been frequently used to determine the probability of chance of correlation during descriptor selection processes [65,66]. In y-randomization, a portion of ER agonists in the data set was randomly selected and converted to ER non-agonists. Another portion of ER non-agonists compounds was also randomly selected and converted to ER agonists. The ratios of ER agonists to ER non-agonists were kept unchanged during y-randomization. The ‘‘scrambled’’ data set was then used for the descriptor selection process. The process of scrambling of the data set and descriptor selection process was repeated for 20 times. This y-randomization analysis was conducted on the SVM model that consistently gives the better classification accuracies in this and other studies [67–69]. Unless over fitting is found in the SVM model, no further analysis on other models is to be conducted. The average Matthews correlation coefficient of these scrambled SVM classification systems derived by using the five-fold cross validation sets were found to be 0.331, which is significantly lower than that of the original SVM classification system, which is 0.767. This suggests that the original SVM classification system is relevant and unlikely to arise as a result of chance of correlation.

The performance of SVM classification system was further evaluated by two additional tests. One is the comparison of the ER binder and non-binder prediction accuracy of SVM with those of the tree-based model developed by Fang et al. [11] in a cited study in which all of the compounds have been provided [11]. The training set consists of 129 ER binders and 101 ER non-binders used by Fang et al. [11] and the testing set includes 56 ER binders and 354 ER non-binders used by Nishihara et al. [39]. The ER binder accuracy of SVM is 83.3% which is comparable to that of 87.1% from the tree-based model [11]. The ER non-binder and the overall prediction accuracy of SVM are 94.7% and 93.1%, respectively, which are significant higher than those of 81.8% and 82.5% from the tree based model [11].

The second test was conducted by using additionally searched ER agonists and non-binders from sources such as the most recent publications, which were carefully checked to ensure that they are not included in our dataset. Table 4 contains 11 ER agonists and 16 ER non-binders reported in the literatures after a comprehensive Medline search. These compounds were used to test the two SVM systems we developed, one developed by using our ER agonists/non-agonists dataset and the other by using the ER binder/non-binder dataset of Fang et al. [11]. It is found that the accuracies of the first SVM system are 90.9% for ER agonists and 100%

Table 4
Recently published ER agonists (+) and ER non-binders (−) searched from literatures

No.	Compound name	CAS no.	Class	Reference
1	Prochloraz	67747-09-5	+	16219411
2	Propamocarb	24579-73-5	+	16219411
3	PCB 74	32690-93-0	+	16203234
4	2-Methoxyestradiol	362-07-2	+	15755993
5	PPT	263717-53-9	+	15722404
6	4-ethoxymethyl phenol	57726-26-8	+	12732288
7	CHF 4056	437756-52-0	+	11861784
8	NNC 45-0781	207277-66-5	+	11738615
9	Indole-3-carbinol	700-06-1	+	16192472
10	Dichlofenthion	97-17-6	+	15064155
11	Chlorpyrifos	2921-88-2	+	15064155
12	PCB 138	35065-28-2	−	16203234
13	PCB 153	35065-27-1	−	16203234
14	PCB 170	35065-30-6	−	16203234
15	PCB 180	35065-29-3	−	16203234
16	PCB 187	52663-68-0	−	16203234
17	PCB 194	35694-08-7	−	16203234
18	PCB 199	52663-75-9	−	16203234
19	PCB 203	52663-76-0	−	16203234
20	Citalopram	59729-33-8	−	[78]
21	Mefloquine	53230-10-7	−	[78]
22	Zonisamide	68291-97-4	−	[78]
23	Escitalopram	128196-01-0	−	[78]
24	Mirtazapine	85650-52-8	−	[78]
25	Nitisinone	104206-65-7	−	[78]
26	Enterolactone	78473-71-9	−	15276617
27	Enterodiol	80226-00-2	−	15276617

Some of the references are given by their PubMed IDs.

for ER non-agonists, and those of the second SVM system are 72.7% for ER binders and 87.5% for ER non-binders. This suggests that SVM trained from a structural more diverse set of compounds can produce a significantly better prediction performance.

Some environmental estrogens, such as kepone (chlordecone), have been found to be estrogenic long after they are released into the environment and their molecular structures are very different from estradiol [70,71]. This makes it difficult to deduce estrogenic activity of these chemicals solely based on their molecular structures [72]. There are two such environmental estrogens in our independent validation set, kepone (chlordecone) and *op*′-DDT, which were correctly predicted by our SVM system (trained from our ER agonists/non-agonists dataset) as ER agonists, demonstrating its strength in identification of classically hard to characterize environmental estrogens at least for kepone and *op*′-DDT and its ability to deduce the estrogenic activity of chemicals based on the selected molecular descriptors.

Overall, our study suggests that SLMs, particularly SVM, are useful for facilitating the prediction of novel ER agonists from compounds with diverse structures. The prediction accuracy of these methods is at a comparable level as those of earlier studies for ER binders in which a substantially less diverse set of compounds were used. Another advantage of the SLMs studied in this work is that they do not require knowledge about the molecular mechanism or structure–activity relationship of a particular drug property.

3.2. Molecular descriptors associated with ER agonism

A total of 31 molecular descriptors were selected by RFE to be associated with the separation between ER agonists and non-agonists. These descriptors, given in Table 5, represent the structural and physicochemical properties associated with ER agonism, some of which are consistent with those used in earlier studies. For instance, earlier studies on the SAR for ER binders have indicated that aromatic ring structures or ring structures containing at least one hydrogen-bonding heteroatom are important features of estrogenic property [11]. The QSAR model of ER binders has been given as a linear function of molecular bulk, polarity, and hydrogen-bonding effects [12]. Eight of our selected molecular descriptors are related to these literature described features. These include molecular globularity (Gloty) and hydrophobic region (Shpb) that describe the bulk, S(10) that describes electro-topological state of the sp^2 atoms in an aromatic ring, $^3\chi_C$ and $^4\chi^v_{PC}$ that describe simple and valence molecular connectivity for a cluster and path of atoms, electrophilicity index (Ω) and electronegativity index (χ_{en}) that describe quantum mechanical properties related to

Table 5
Molecular descriptors selected from the RFE feature selection method for the classification of ER agonists and ER non-agonists

Descriptors	Description	Class
Nhet	Count of hetero atoms	Simple molecular property
$^3\chi_C$	Simple molecular connectivity Chi indices for cluster	Connectivity
$^4\chi^v_{PC}$	Valence molecular connectivity Chi indices for path/cluster	Connectivity
S(10)	Atom-type H Estate sum for :CH: (sp^2, aromatic)	Electrotopological state
S(16)	Atom-type Estate sum for –CH_3	Electrotopological state
Tcent	Centric index	
Tpeti	PetitJohn I2 index	Electrotopological state
Tiwie	Information Weiner	Electrotopological state
ε_b	Hydrogen bond acceptor basicity (covalent HBAB)	Quantum chemical properties
M	Molecular dipole moment	Quantum chemical properties
H	Absolute hardness	Quantum chemical properties
IP	Ionization potential	Quantum chemical properties
μ_{cp}	Chemical potential	Quantum chemical properties
χ_{en}	Electronegativity index	Quantum chemical properties
Ω	Electrophilicity index	Quantum chemical properties
$Q_{H,max}$	Most positive charge on H atom	Quantum chemical properties
$Q_{H,min}$	Most negative charge on H atom	Quantum chemical properties
$Q_{C,min}$	Most negative charge on C atom	Quantum chemical properties
$Q_{C,SS}$	Sum of squares of charges on C and all atoms	Quantum chemical properties
Mpc	Mean of positive charges	Quantum chemical properties
dis2	Length vectors (longest third atom)	Geometrical properties
dis3	Length vectors (4th atom)	Geometrical properties
Sapc	Sum of solvent accessible surface areas of positively charged atoms	Geometrical properties
Sanc	Sum of solvent accessible surface areas of negatively charged atoms	Geometrical properties
Sapcw	Sum of charge weighted solvent accessible surface areas of positively charged atoms	Geometrical properties
Svpc	Sum of van der Waals surface areas of negatively charged atoms	Geometrical properties
Rugty	Molecular rugosity	Geometrical properties
Gloty	Molecular globularity	Geometrical properties
Shpb	Hydrophobic region	Geometrical properties
Hiwpl	Hydrophilic integy moment	Geometrical properties
Hiwpa	Amphiphilic moment	Geometrical properties

polarity, and hydrogen bond acceptor basicity (ε_b) that describes quantum chemical properties associated with hydrogen bonding.

Structural studies from X-ray crystallography appear to lend further support to our selected molecular descriptors. Table S2 of the supplementary material gives the characteristics of several known structures of ligand–ER complexes. The major types of molecular interactions in these structures are hydrophobic, non-polar, aromatic, polar, and hydrogen bond, which is consistent with our selected molecular descriptors.

ER antagonists bind to the same binding site as agonists but induce slightly different molecular conformations. For instance, an antagonist raloxifene (RAL) binds to the same site as an agonist estradiol (E2) in the ligand binding domain of ERα. However, the binding of RAL causes the imidazole ring of His524 rotates and displaces 5.1 angstrom from the position occupied by E2 in order to form a favorable hydrogen-bonding position [34], as shown in Fig. 1A and B. This shows that atomic connectivity is important for ER agonism. Molecular descriptors associated with connectivity, such as our selected ${}^{3}\chi_{C}$ and ${}^{4}\chi_{PC}^{v}$, can be important for selecting ER agonists from ER binders. Hence, knowing certain molecular descriptors associated with ER binding may not always be sufficient for identifying ER agonists because both ER agonists and antagonists display similar binding modes to ER. Specific molecular descriptors such as connectivity and spatial charge distribution appear to be important for distinguishing agonists from antagonists. The use of these molecular descriptors in this study is likely one of the reasons for the good performance of SLMs in predicting ER agonists and non-agonists.

X-ray crystallography studies have indicated that ER antagonists such as OHT, RAL, ICI possess bulky side chains that protrude out of the opening lips of the ligand binding site and thus preventing helix12 from adopting the agonist-bound conformation [34,73–75]. The protruded side chains cause helix12 adopt an alternative conformation that occludes the binding of transcriptional coactivators. Our selected molecular descriptors in this study such as Gloty for molecular globularity, Shpb for hydrophobic region, and S(10) corresponding to the electro-topological state of sp^2 atoms in the aromatic ring correlate well with the bulky property of ER antagonists. Moreover, RAL and OHT are antagonists of ERα that contain basic amine side chains forming hydrogen bonds with Asp351 in the crystal structures. These basic side chains of antagonists displace the AF2 helix from a conformation that favor the recruitment of transcriptional coactivators to that unfavorable to such recruitment [73]. Our selected molecular descriptor ε_b for hydrogen bond acceptor basicity is also consistent with the fact that some ER antagonists such as RAL and OHT (which belong to one class of ER non-agonists) contain basic amine side chains of ER antagonists.

It is noted that our study is focused on the prediction of ER agonists and non-agonists without further classification of ER antagonists and ER non-binders. Therefore, our prediction systems are not intended for identification of ER antagonists. Non-the-less, as ER antagonists form an important group of ER non-agonists in our dataset, some of the molecular characteristics specific to ER antagonism are likely included in our selected molecular descriptors. Shiau et al. have shown that, while it is an agonist of ERα, THC also acts as a ''passive'' ERβ antagonist due to the lack of a bulky side chain [35]. Upon binding to ERβ, THC destabilizes helix 12 of ERβ from recruiting transcriptional coactivators [35]. Based on our definition of ER agonists, THC was included in the ER agonist group. Because of the lack of ''passive'' ER antagonists in our dataset, it is unlikely that ''passive'' antagonism is adequately covered in our selected molecular descriptors. Further works for extending our models for predicting ER antagonism and for

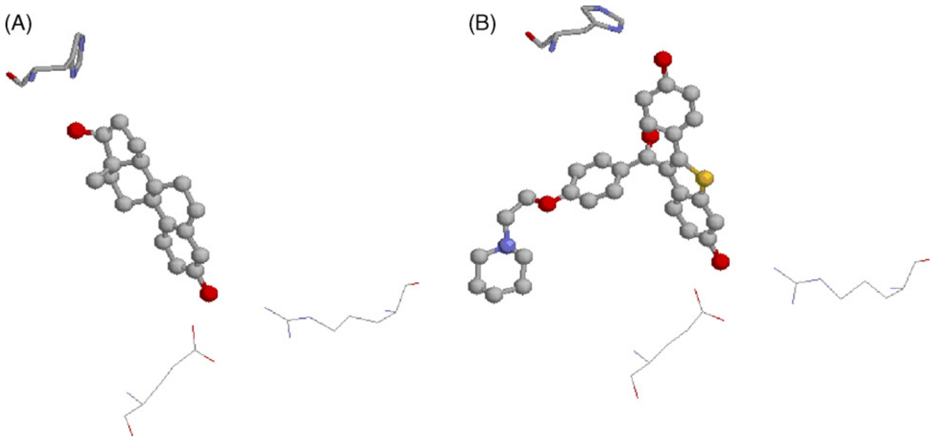

Fig. 1. A and B Binding of agonist (E2) and antagonist (RAL) to ERα. (A) Shows binding of agonist E2 and (B) for antagonist RAL to ERα from X-ray crystallography by Brzozowski et al. [34]. Only three residues (Glu353, Arg394, and His524) in the ligand binding domain (LBD) are shown for clarity. From the figure, both E2 and RAL bind to LBD with similar types of interactions (hydrophobic, polar, and hydrogen bonds), His524 is rotated 5.1 angstrom away from the position occupied by E2 in order to form a favorable hydrogen bond with RAL although both E2 and RAL form hydrogen bonds with Glu353 and Arg394 in similar molecular conformation. This shows that slight variation in rotamer orientation of residues in LBD can cause dramatic effect as agonist or antagonist activities. The variation of these activities cannot be determine from ER binding affinity since both ER agonists and ER antagonists exhibit similar interaction types. Hence, SLMs together with feature selection method are used to characterize estrogen receptor associated molecular descriptors.

Table 6
Average values of the descriptors most relevant to distinguishing ER agonists from ER non-agonists

Descriptor	Descriptor class	Average value[a]	
		ER agonists	ER non-agonists
Ω (electrophilicity index)	Electrostatic	48.16 (±0.64)	53.04 (±0.73)
Nhet (count of hetero atoms)	Hydrogen bonding	3.89 (±0.17)	3.77 (±0.14)
Hiwpl (hydrophilic integy moment)	Hydrophobicity	14.23 (±0.27)	12.29 (±0.29)
dis2 (length vectors of the longest third atom)	Size	15.44 (±0.25)	13.61 (±0.25)
Gloty (molecular globularity)	Shape	2.13 (±0.02)	1.98 (±0.02)

The averages are taken over all of the ER agonists and ER non-agonists, respectively.

[a] Values in parentheses are the standard error.

distinguishing between "active" and "passive" antagonism are needed for fully studying different types of ER binding and for the design of ER antagonist-based drugs.

To provide further insight into the correlation between our selected molecular descriptors and ER agonism, these molecular descriptors are classified into five major classes of interaction types: electrostatic interactions, hydrogen bonding, hydrophobicity, size and shape. The computed average values of the most relevant descriptor in each of the five classes over all ER agonists and those over all ER non-agonists are given in Table 6. It is found that the molecular properties for ER agonists are generally larger in size, higher in molecular globularity, more hydrophilic, lower in the electrophilicity index, and exhibiting higher hydrogen bonding potential than non-agonists. These properties are consistent with the observation that the ligand binding site of ER is generally "plastic" to various structurally distinct compounds with relatively large molecular volume [73], and to the finding that the basic side chains of antagonists displace the AF2 helix thereby disfavoring the recruitment of transcriptional coactivators [73].

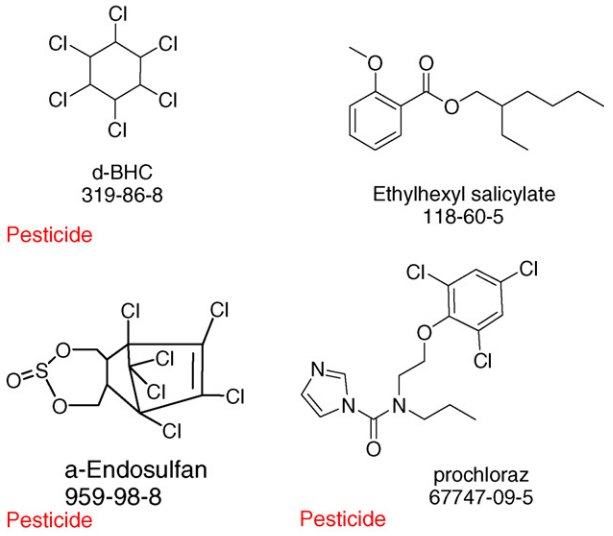

Fig. 2. Structures of misclassified ER agonists in the independent validation set. Chemical names, relevant Chemical Abstracts Service (CAS) number of compound and structure type are shown in the figure. Hetero atoms (oxygen) are marked.

3.3. Misclassified ER agonists and non-agonists from independent test sets

Figs. 2 and 3 show the four misclassified ER agonists and one misclassified ER non-agonist in our independent evaluation set and the set of recently discovered ER agonists/non-binders. The misclassified ER agonists are d-BHC (CAS No. 319-86-8), ethyhexyl salicylate (CAS No. 118-60-5), α-endosulfan (CAS No. 959-98-8), and prochloraz (CAS No. 67747-09-5). The misclassified ER non-agonist is 4-hydroxybenzaldehyde (CAS No. 123-08-0). From Figs. 2 and 3, the misclassified compounds are mainly polychlorinated pesticides with multiple chlorine atoms attached in same ring structure. This seems to suggest that our currently used molecular descriptors may not be sufficient to properly represent these types of structures, and further improvement and refinement of our molecular descriptors may be needed.

SLMs are subjected to some degree of error due to such factors as dataset 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 biological activity and physiological effect [76]. However, biological activity of a compound is an induced response that is influenced by numerous factors dictated by the many levels of biological complexity. The relationship between structure and activity is thus more implicit and thereby requires

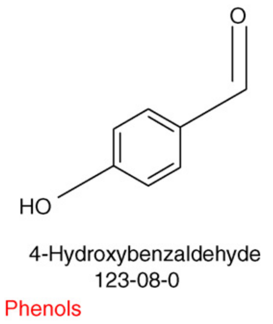

Fig. 3. Structure of misclassified ER non-agonists in the independent validation set. Chemical names, relevant Chemical Abstracts Service (CAS) number of compound and structure type are shown in the figure. Hetero atoms (oxygen) are marked.

a more thorough investigation and rigorous validation [77]. Hence, the choice for better descriptors is still under investigation.

4. Conclusion

ER agonists are important in regulation of a wide range of biological processes such as development and oncogenesis. Some environmental compounds disrupt normal functions of endocrine system due to their ER agonism. Thus, identification of novel ER agonists from structurally diverse compounds is important for drug discovery and environmental safety evaluation. Works on QSAR for ER binders have identified types of molecular interactions important for ER binding. However, knowledge of ER binding is insufficient for determining ER agonism, which is more relevant to drug discovery and safety evaluation. Hence, SLMs were explored as tools for predicting ER agonists and for characterizing molecular descriptors associated with ER agonism. This study shows that SLMs such as SVM, k-NN and PNN are useful for performing these tasks. Among the SLMs, SVM shows the highest classification accuracy. By incorporating feature selection methods such as RFE into SLMs, molecular descriptors relevant to ER agonistic activities can be identified. Some of these selected molecular descriptors are consistent to those used in previous studies and with the findings from X-ray crystallography studies. Further works on the improvement and refinement of feature selection methods as well as molecular descriptors are needed in order to improve the capability of SLMs for accurately identifying ER agonists and the related molecular characteristics.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmgm.2006.01.007.

References

[1] M.J. Tsai, B.W. O'Malley, Molecular mechanisms of action of steroid/thyroid receptor superfamily members, Annu. Rev. Biochem. 63 (1994) 451–486.

[2] P.S. Danielian, R. White, J.A. Lees, M.G. Parker, Identification of a conserved region required for hormone dependent transcriptional activation by steroid hormone receptors, EMBO J. 11 (3) (1992) 1025–1033.

[3] J.I. MacGregor, V.C. Jordan, Basic guide to the mechanisms of antiestrogen action, Pharmacol. Rev. 50 (2) (1998) 151–196.

[4] K. Paech, P. Webb, G.G. Kuiper, S. Nilsson, J. Gustafsson, P.J. Kushner, T.S. Scanlan, Differential ligand activation of estrogen receptors ERalpha and ERbeta at AP1 sites, Science 277 (5331) (1997) 1508–1510.

[5] B.H.J. Coelingh, Are all estrogens the same? Maturitas 47 (4) (2004) 269–275.

[6] W.K. Oh, The evolving role of estrogen therapy in prostate cancer, Clin. Prostate Cancer 1 (2) (2002) 81–89.

[7] C. Behl, Oestrogen as a neuroprotective hormone, Nat. Rev. Neurosci. 3 (4) (2002) 433–442.

[8] L.W. Lissin, J.P. Cooke, Phytoestrogens and cardiovascular health, J. Am. Coll. Cardiol. 35 (6) (2000) 1403–1410.

[9] S.H. Safe, L. Pallaroni, K. Yoon, K. Gaido, S. Ross, B. Saville, D. McDonnellc, Toxicology of environmental estrogens, Reprod. Fertil. Dev. 13 (4) (2001) 307–315.

[10] B. Hileman, Hormone disrupter research expands, Chem. Eng. News 75 (34) (1997) 24.

[11] H. Fang, W. Tong, L.M. Shi, R. Blair, R. Perkins, W. Branham, B.S. Hass, Q. Xie, S.L. Dial, C.L. Moland, D.M. Sheehan, Structure–activity relationships for a large diverse set of natural, synthetic, and environmental estrogens, Chem. Res. Toxicol. 14 (2001) 280–294.

[12] J.Y. Hu, T. Aizawa, Quantitative structure–activity relationships for estrogen receptor binding affinity of phenolic chemicals, Water Res. 37 (6) (2003) 1213–1222.

[13] W. Tong, Q. Xie, H. Hong, L. Shi, H. Fang, R. Perkins, Assessment of prediction confidence and domain extrapolation of two structure–activity relationship models for predicting estrogen receptor binding activity, Environ. Health Perspect. 112 (12) (2004) 1249–1254.

[14] A.H. Asikainen, J. Ruuskanen, K.A. Tuppurainen, Consensus kNN QSAR: a versatile method for predicting the estrogenic activity of organic compounds in silico. A comparative study with five estrogen receptors and a large, diverse set of ligands, Environ. Sci. Technol. 38 (24) (2004) 6724–6729.

[15] H. Hong, W. Tong, H. Fang, L. Shi, Q. Xie, J. Wu, R. Perkins, J.D. Walker, W. Branham, D.M. Sheehan, Prediction of estrogen receptor binding for 58,000 chemicals using an integrated system of a tree-based model with structural alerts, Environ. Health Perspect. 110 (1) (2002) 29–36.

[16] V.V. Zernov, K.V. Balakin, A.A. Ivaschenko, N.P. Savchuk, I.V. Pletnev, Drug discovery using support vector machines. The case studies of drug-likeness, agrochemical-likeness, and enzyme inhibition predictions, J. Chem. Inf. Comput. Sci. 43 (6) (2003) 2048–2056.

[17] E. Byvatov, U. Fechner, J. Sadowski, G. Schneider, Comparison of support vector machine and artificial neural network systems for drug/nondrug classification, J. Chem. Inf. Comput. Sci. 43 (6) (2003) 1882–1889.

[18] Y. Xue, C.W. Yap, L.Z. Sun, Z.W. Cao, J.F. Wang, Y.Z. Chen, Prediction of *p*-glycoprotein substrates by support vector machine approach, J. Chem. Inf. Comput. Sci. 44 (4) (2004) 1497–1505.

[19] S. Doniger, T. Hofman, J. Yeh, Predicting CNS permeability of drug molecules: comparison of neural network and support vector machine algorithms, J. Comput. Biol. 9 (6) (2002) 849–864.

[20] L. He, P.C. Jurs, L.L. Custer, S.K. Durham, G.M. Pearl, Predicting the genotoxicity of polycyclic aromatic compounds from molecular structure with different classifiers, Chem. Res. Toxicol. 16 (12) (2003) 1567–1580.

[21] R.D. Snyder, J.W. Green, A review of the genotoxicity of marketed pharmaceuticals, Mutat. Res.: Rev. Mutat. 488 (2) (2001) 151–169.

[22] C.W. Yap, C.Z. Cai, Y. Xue, Y.Z. Chen, Prediction of torsade-causing potential of drugs by support vector machine approach, Toxicol. Sci. 79 (1) (2004) 170–177.

[23] C.W. Yap, Y.Z. Chen, Quantitative structure–pharmacokinetic relationships for drug distribution properties by using general regression neural network, J. Pharm. Sci. 94 (1) (2004) 153–168.

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

[25] C.J.C. Burges, A tutorial on support vector machines for pattern recognition, Data Min. Knowl. Disc. 2 (2) (1998) 127–167.

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

[27] D.F. Specht, Probabilistic neural networks, Neural Netw. 3 (1) (1990) 109–118.

[28] J.R. Quinlan, C4. 5: Programs for Machine Learning, Morgan Kaufmann, San Mateo, CA, 1993.

[29] H. Yu, J. Yang, W. Wang, J. Han, Discovering compact and highly discriminative features or feature combinations of drug activities using support vector machines, in: IEEE Computer Society Bioinformatics Conference (CSB'03), Stanford, CA, August 11–14, (2003), pp. 220–228.

[30] Y. Xue, Z.R. Li, C.W. Yap, L.Z. Sun, X. Chen, Y.Z. Chen, 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 (5) (2004) 1630–1638.

[31] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene selection for cancer classification using support vector machines, Mach. Learn. 46 (1–3) (2002) 389–422.

[32] W.S. Branham, S.L. Dial, C.L. Moland, B.S. Hass, R.M. Blair, H. Fang, L. Shi, W. Tong, R.G. Perkins, D.M. Sheehan, Phytoestrogens and mycoestrogens bind to the rat uterine estrogen receptor, J. Nutr. 132 (4) (2002) 658–664.

[33] H. Kojima, E. Katsura, S. Takeuchi, K. Niiyama, K. Kobayashi, Screening for estrogen and androgen receptor activities in 200 pesticides by in vitro reporter gene assays using Chinese hamster ovary cells, Environ. Health Perspect. 112 (5) (2004) 524–531.

[34] A.M. Brzozowski, A.C. Pike, Z. Dauter, R.E. Hubbard, T. Bonn, O. Engstrom, L. Ohman, G.L. Greene, J.A. Gustafsson, M. Carlquist, Molecular basis of agonism and antagonism in the oestrogen receptor, Nature 389 (1997) 753–758.

[35] A.K. Shiau, D. Barstad, J.T. Radek, M.J. Meyers, K.W. Nettles, B.S. Katzenellenbogen, J.A. Katzenellenbogen, D.A. Agard, G.L. Greene, Structural characterization of a subtype-selective ligand reveals a novel mode of estrogen receptor antagonism, Nat. Struct. Biol. 9 (5) (2002) 359–364.

[36] H. Li, C.W. Yap, C.Y. Ung, Y. Xue, Z.W. Cao, Y.Z. Chen, 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 (5) (2005) 1376–1384.

[37] C.W. Yap, Y.Z. Chen, Prediction of cytochrome P450 3A4, 2D6, and 2C9 inhibitors and substrates by using support vector machines, J. Chem. Inf. Model. 45 (4) (2005) 982–992.

[38] R.M. Blair, H. Fang, W.S. Branham, B.S. Hass, S.L. Dial, C.L. Moland, W. Tong, L. Shi, R. Perkins, D.M. Sheehan, The estrogen receptor relative binding affinities of 188 natural and xenochemicals: Structural diversity of ligands, Toxicol. Sci. 54 (2000) 138–153.

[39] T. Nishihara, J. Nishikawa, T. Kanayama, F. Dakeyama, K. Saito, M. Imagawa, S. Takatori, Y. Kitagawa, S. Hori, H. Utsumi, Estrogenic activities of 517 chemicals by yeast two-hybrid assay, J. Health Sci. 46 (4) (2000) 282–298.

[40] CambridgeSoft Corporation. ChemDraw. In., 7.0.1 ed. Cambridge, MA 02140 USA, 2002.

[41] Accelrys. DS ViewerPro. In., 5.0 ed., California, USA.

[42] R.S. Pearlman, CONCORD User's Manual. In. St. Louis, MO: Tripos.

[43] M.J.S. Dewar, E.G. Zoebisch, E.F. Healy, J.P.P. Steward, AM1: a new general purpose quantum mechanical molecular model, J. Am. Chem. Soc. 107 (1985) 3902–3909.

[44] S.O. Mueller, J.A. Katzenellenbogen, K.S. Korach, Endogenous estrogen receptor beta is transcriptionally active in primary ovarian cells from estrogen receptor knockout mice, Steroids 69 (10) (2004) 681–686.

[45] http://pubchem.ncbi.nlm.nih.gov/.

[46] http://chemfinder.cambridgesoft.com/.

[47] J.J. Perez, Managing molecular diversity, Chem. Soc. Rev. 34 (2) (2005) 143–152.

[48] P. Willett, J.M. Barnard, G.M. Downs, Chemical similarity searching, J. Chem. Inf. Comput. Sci. 38 (1998) 983–996.

[49] R. Todeschini, V. Consonni, Handbook of Molecular Descriptors, Wiley-VCH, Weinheim, 2000.

[50] R.D. Snyder, G.S. Pearl, G. Mandakas, W.N. Choy, F. Goodsaid, I.Y. Rosenblum, Assessment of the sensitivity of the computational programs DEREK, TOPKAT, and MCASE in the prediction of the genotoxicity of pharmaceutical molecules, Environ. Mol. Mutagen. 43 (3) (2004) 143–158.

[51] S. Degroeve, B. De Baets, Y. Van de Peer, P. Rouzé, Feature subset selection for splice site prediction, Bioinformatics 18 (2002) S75–S83.

[52] R. Kohavi, G.H. John, Wrappers for feature subset selection, Artif. Intell. Med. 97 (1997) 273–324.

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

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

[55] P. Baldi, S. Brunak, Y. Chauvin, C.A. Andersen, H. Nielsen, Assessing the accuracy of prediction algorithms for classification: an overview, Bioinformatics 16 (5) (2000) 412–424.

[56] J.E. Roulston, Screening with tumor markers, Mol. Pharmacol. 20 (2002) 153–162.

[57] B.W. Matthews, Comparison of the predicted and observed secondary structure of T4 phage lysozyme, Biochim. Biophys. Acta 405 (2) (1975) 442–451.

[58] 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 methods, Chem. Res. Toxicol. 18 (6) (2005) 1071–1080.

[59] H. Yu, J. Yang, W. Wang, J. Han, Discovering compact and highly discriminative features or feature combinations of drug activities using support vector machines, in: Proceedings of the IEEE computer society bioinformatics conference (CSB), 2003, pp. 220–228.

[60] R. Kohavi, G.H. John, Wrappers for feature subset selection, Artif. Intell. 97 (1–2) (1997) 273–324.

[61] J.M. Sutter, J.H. Kalivas, Comparison of forward selection, backward elimination, and generalized simulated annealing for variable selection, Microchem. J. 47 (1–2) (1993) 60–66.

[62] D.M. Hawkins, The problem of overfitting, J. Chem. Inf. Comput. Sci. 44 (1) (2004) 1–12.

[63] J.G. Topliss, R.P. Edwards, Chance factors in studies of quantitative structure–activity relationships, J. Med. Chem. 22 (10) (1979) 1238–1244.

[64] D. Jouan-Rimbaud, D.L. Massart, O.E. de Noord, Random correlation in variable selection for multivariate calibration with a genetic algorithm, Chemometr. Intell. Lab. 35 (2) (1996) 213–220.

[65] B.F.J. Manly, Randomization Bootstrap and Monte Carlo Methods in Biology, 2nd ed., Chapman & Hall, London, 1997.

[66] R. Leardia, A.L. González, Genetic algorithms applied to feature selection in PLS regression: how and when to use them, Chemometr. Intell. Lab. 41 (2) (1998) 195–207.

[67] R. Burbidge, M. Trotter, B. Buxton, S. Holden, Drug design by machine learning: support vector machines for pharmaceutical data analysis, Comput. Chem. 26 (1) (2001) 5–14.

[68] R. Czerminski, A. Yasri, D. Hartsough, Use of support vector machine in pattern classification: application to QSAR studies, Quant. Struct.: Act. Rel. 20 (3) (2001) 227–240.

[69] D. Meyer, F. Leischa, K. Hornik, The support vector machine under test, Neurocomputing 55 (1–2) (2003) 169–186.

[70] D. Kupfer, Effects of pesticides and related compounds on steroid metabolism and function, Crit. Rev. Toxicol. 4 (1) (1975) 83–124.

[71] P.S. Guzelian, Comparative toxicology of chlordecone (Kepone) in humans and experimental animals, Annu. Rev. Pharmacol. Toxicol. 22 (1982) 89–113.

[72] A.M. Soto, K.L. Chung, C. Sonnenschein, The pesticides endosulfan, toxaphene, and dieldrin have estrogenic effects on human estrogen-sensitive cells, Environ. Health Perspect. 102 (4) (1994) 380–383.

[73] A.K. Shiau, D. Barstad, P.M. Loria, L. Cheng, P.J. Kushner, D.A. Agard, G.L. Greene, The structural basis of estrogen receptor/coactivator recognition and the antagonism of this interaction by tamoxifen, Cell 95 (7) (1998) 927–937.

[74] A.C. Pike, A.M. Brzozowski, R.E. Hubbard, T. Bonn, A.G. Thorsell, O. Engstrom, J. Ljunggren, J.A. Gustafsson, M. Carlquist, Structure of the ligand-binding domain of oestrogen receptor beta in the presence of a partial agonist and a full antagonist, EMBO J. 18 (1999) 4608–4618.

[75] A.C. Pike, A.M. Brzozowski, J. Walton, R.E. Hubbard, A.G. Thorsell, Y.L. Li, J.A. Gustafsson, M. Carlquist, Structural insights into the mode of action of a pure antiestrogen, Structure 9 (2) (2001) 145–153.

[76] M. Johnson, G.M. Maggiora, Concepts and Applications of Molecular Similarity, Wiley, 1990.

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

[78] C.F. Lacy, L.L. Armstrong, M.P. Goldman, L.L. Lance, 10th Anniversary ed., Drug Information Handbook, vol. 2003–2004, Lexi-Comp, Hudson, Cleveland, 2004.