Journal of Coastal Development
ISSN: 1410-5217

Li F, et al., J Coast Dev 2014, 17:1
DOI: 10.4172/ 1410-5217.1000388

Research Article Open Access

Combined SVM-PLS Method for Predicting Estrogenic Activities of Organic Chemicals in the Coastal Water

Fei Li*, Lulu Cao, Huifeng Wu and Jianmin Zhao

Key Laboratory of Coastal Zone Environmental Processes, Yantai Institute of Coastal Zone Research (YIC), Chinese Academy of Sciences (CAS); Shandong Provincial Key Laboratory of Coastal Zone Environmental Processes, YICCAS, Yantai Shandong 264003, PR China

***Corresponding author:** Fei Li, Key Laboratory of Coastal Zone Environmental Processes, Yantai Institute of Coastal Zone Research, China, Tel: +86-535-210902 E-mail: fli@yic.ac.cn

Received date: May 14, 2014; **Accepted date:** June 28, 2014; **Published date:** July 06, 2014

Abstract

A data set of 517 natural, synthetic and environmental chemicals belonging to a broad range of structural classes have been tested for estrogenic activities (expressed as logREC10) to the estrogen receptor (ER) using a yeast two-hybrid assay. In this study, quantitative structure- activity relationships (QSARs) were determined using two methods, partial least square (PLS) and support vector machine (SVM). The Q^2_{cum} of the PLS model is 0.678, indicating high robustness and good predictive ability. The correlation coefficient (R) between the observed and the predicted values is 0.870, indicating the predicted values by the final QSAR models were in good agreement with the corresponding experimental values. Eight DRAGON descriptors were included in the PLS model, including Mor_{03p}, L_{3e}, R_{8p}, R_{Tv}^+, R_{8e}, R_{1p}^+, R_{7p}^+ and $HATS_v$, which implies that chemical estrogenic activities are related to atomic properties (atomic Sanderson electronegativities, polarizabilities and van der Waals volumes). Comparison of the results obtained from two models showed that the SVM method exhibited the best overall performances, with a RMS error of 0.145 $logREC_{10}$ units for the whole set. Moreover, three linear QSAR models were constructed for some specific families based on their chemical structures. These predictive models should be useful to rapidly identify potential estrogenic endocrine disrupting chemicals.

Key words:

Endocrine disrupting chemicals (EDCs); Estrogen receptor (ER); Quantitative structure activity relationship (QSAR); PLS; SVM

Introduction

Scientific and public concern heightens over the potential health effects of exposure to environmental pollutants with endocrine disruption potential [1]. Diminished or excessive production of estrogens is a major problem related to prostate cancer, spinal bular musular atrophy, and female pattern baldness. Consequently, there is a need to develop screening and testing procedures for endocrine disrupting chemicals (EDCs).

Considering the high number of potential EDCs, this remains a labor intensive and time-costing operation. It is crucial to develop efficient and economical alternative modeling approaches for the purpose of predicting the estrogenic activities of potential EDCs. Quantitative structure activity relationship (QSAR) methods are the most promising and successful tools to provide rapid and useful meanings for predicting the biological activity and chemical toxicity. They are considered as an important part of the priority setting process by the endocrine disruptor screening and testing advisory committee (EDSTAC) [2]. QSAR are widely applied for the understanding of the mechanism of chemicals' binding for the estrogen receptors (ER) [3-6], for androgen receptor (AR) and for several other members of the nuclear receptor family [7]. These include electrostatic models, comparative molecular field analysis (CoMFA) which considers the overall steric and electrostatic properties of the compound of interest, computer graphic and energy (electrostatic and van der Waals) based models for fit into DNA and common reactivity patterns (COREPA) which reflect the stereoelectronic features.

In this paper, a data set consisted of experimental values which were determined by Nishihara et al. [8], including 517 natural, synthetic, and environmental chemicals from a broad range of structural classes. The data set was used to construct global QSAR models for the whole data set and local models for specific well-known families. Some information descriptors were selected using Forward stepwise (FS) regression from the original 709 DRAGON calculated descriptors and were applied to construct an optimal model based on SVM. Another classical method, partial least square (PLS) [9] was utilized to establish QSAR model to compare the results with those obtained by SVM. In addition, some careful models for specific well-known families were examined in conjunction with knowledge of the recently reported ligand-ER crystal structures.

Data and Methods

Experiment and data set

Because it is expected that the major key target of EDCs is the nuclear hormone receptor, which binds specifically to the steroid hormone and regulates its gene expression, the yeast two-hybrid assay has been developed. Unlike yeast-based assay (YES) [10], another reporter gene assay

using a yeast two-hybrid cells, the method contains the coactivator, so that the system more closely resembles the mammalian hormonal system.A detailed description of the experimental methods is provided in Nishihara et al. [8].

The overall data set consisted of more than 500 organic chemicals, including natural substances, medicine, pesticides, and industrial chemicals. Table 1 shows a summary of the test compounds with the names of 55 positive compounds. Tested chemicals consisted of natural substances (metabolites, oxidation products, etc.), medicines, food additives, pesticides, and industrial chemicals (PCBs, PCDFs, PAHs, phenols, benzenes, phthalates and adipates, and others). The estrogenic activities to the ER, expressed as log unit of 10% relative effective concentration ($logREC_{10}$), are listed in Table 1.

			pre. $logREC_{10}$[b]	
No	Compounds	obs.logREC10[a]	PLS	SVM
A. natural products and related				
1	17α-Estradiol	3.125	3.461	2.905
2	Apigenin	6.523	5.999	6.303
3	Coumestrol	6.523	5.884	6.312
4	Daidzein	5.000	5.335	4.840
5	Dihydrogenistein	5.000	4.077	4.754
6	Equol	6.523	4.896	5.154
7	Estrone	2.000	3.802	3.367
8	Genistein	4.523	5.076	4.741
B. medicines, food additives, and related				
9	17α-Ethynylestradiol	1.824	2.711	3.504
10	β-Estradiol-17-acetate	5.222	3.063	4.098
11	Diethylstilbesterol (DES)	1.824	2.496	2.042
12	Ethyl 4-hydroxybenzoate	7.523	6.844	7.303
13	Methyl 4-hydroxybenzoate	8.125	7.041	7.907
14	n-Butyl 4-hydroxybenzoate	6.000	5.662	6.313
15	n-Propyl 4-hydroxybenzoate	6.523	6.091	6.433
C. PCBs, PCDFs, PAHs, and related				
16	2-Hydroxy benzo[a]pyrene	7.222	6.342	6.478
17	2-Hydroxy fluorene	7.523	7.476	7.589
18	3,8-Dihydroxy-2-chlorodibenzofuran	5.426	6.224	5.876
19	3-Hydroxy benzo[a]pyrene	6.523	7.216	6.431
20	4-Hydroxy-2',4',6'-trichlorobiphenyl	5.125	6.190	6.082
21	4-Hydroxy-2',4',6'-trichlorobiphenyl	6.301	6.190	6.082
22	8-Hydroxy-2,3,4-trichlorodibenzofuran	6.523	7.037	6.742
23	8-Hydroxy-2-monochlorodibenzofuran	6.523	7.220	6.673
24	8-Hydroxy-3,4,6-trichlorodibenzofuran	6.523	7.125	6.300
25	8-Hydroxy-3,4-dichlorodibenzofuran	6.222	7.165	6.401
26	8-Hydroxy-3-monochlorodibenzofuran	6.368	7.240	6.537
D. Phenols				

27	2,2-Bis(4-hydroxy-3-methylphenyl)propane	6.000	5.549	5.896
28	2,2-Bis(4-hydroxy-phenyl)butane	6.000	5.682	5.782
29	2,4-Dichlorophenol	7.125	7.180	6.918
30	3,4-Dichlorophenol	6.824	7.429	7.298
31	4,4'-Dihydroxybenzophenone	8.000	7.209	7.294
32	4,4'-Dihydroxybiphenyl	6.222	6.406	6.374
33	4,4'-Thiobiphenyl	6.000	5.984	5.780
34	4-Bromophenol	7.426	7.446	7.209
35	4-Chloro-3,5-xylenol	7.523	7.656	7.743
36	4-Chloro-3-methylphenol	7.222	7.358	7.251
37	4-Chlorophenol	7.824	7.537	7.605
38	4-Ethylphenol	7.000	7.331	7.324
39	4-Hydroxyacetophenone	7.824	7.595	7.915
40	4-Hydroxybiphenyl	7.000	7.462	7.482
41	4-Methylphenol (p-cresol)	8.000	7.523	7.604
42	4-n-Butylphenol	6.523	7.160	6.742
43	4-n-Hexylphenol	6.523	6.851	6.314
44	4-n-Pentylphenol	6.000	6.928	6.637
45	4-n-Propylphenol	7.368	7.114	6.881
46	4-sec-Butylphenol	6.523	7.159	6.981
47	4-tert-Butylphenol	7.000	6.201	6.184
48	4-tert-Octylphenol	4.824	5.805	5.615
49	4-tert-Pentylphenol	5.523	5.900	5.740
50	Bis(4-hydroxyphenyl)methane	6.824	6.981	7.101
51	Bisphenol A	6.000	5.707	6.023
E. Benzenes and heterocyclics				
52	cis-1,2-Diphenylcyclobutane	8.000	7.569	7.783
F. Phthalates and adipates				
53	Benzylbutyl phthalate (BBP)	8.222	7.128	7.833
54	Di-iso-propyl phthalate	8.824	7.837	8.605
55	Di-n-propyl phthalate	8.523	7.655	8.314

Table 1: Observed and predicted logREC10 for the global QSAR model. [a]Observed $logREC_{10}$ obtained from reference [1]. [b]Predicted $logREC_{10}$ obtained from PLS and SVM models, respectively.

Descriptor generation and selection

Structures of chemicals were drawn with the Chem Draw computer program. These were then geometry-optimized with the PM3 Hamiltonian using the software package Chemoffice 6.0 program, and exported into a file format suitable for MOPAC analysis. The resulting geometry was transferred into the DRAGON software that was used to calculate molecular structural descriptors. Molecular descriptor meanings and their calculation procedure are summarized in the DRAGON software, and explained in detail, with related literature references, in the Handbook of Molecular Descriptors by Todeschini and Consonni [11].

Generally, more descriptors should be considered in QSAR study so as to better characterize molecular structures. However, if no significant relevant or irrelevant descriptors are included, the quality of prediction and robustness of the developed QSAR model may decrease, and its interpretation becomes more difficult. Hence, descriptors selection is necessary for QSAR study.

In this work, the FS regression was employed to select the optimal subset from an original set of 709 calculated descriptors, as also did and described in other studies [12]. As a result, 13 descriptors were obtained, which are listed in Table 2 with their physical-chemical meanings.

Descriptor	Chemical Meanings	Type*
Mor_{10e}	3D-MoRSE - signal 10/ weighted by atomic Sanderson electronegativities	3D-MoRSE
Mor_{03p}	3D-MoRSE - signal 03/weighted by atomic polarizabilities	3D-MoRSE
Mor_{23p}	3D-MoRSE - signal 23/weighted by atomic polarizabilities	3D-MoRSE
L_{3e}	3rd component size directional WHIM index/weighted by atomic Sanderson electronegativities	WHIM
G_{1e}	1st component symmetry directional WHIM index/weighted by atomic Sanderson electronegativities	WHIM
V_s	V total size index/weighted by atomic electrotopological states	WHIM
R_{8p}	R autocorrelation of lag 8/weighted by atomic polarizabilities	GETAWAY
R_{Tv}^+	R maximal index/weighted by atomic van der Waals volumes	GETAWAY
H_{4p}	H autocorrelation of lag 4/weighted by atomic polarizabilities	GETAWAY
R_{8e}	R autocorrelation of lag 8/weighted by atomic Sanderson electronegativities	GETAWAY
R_{1p}^+	R maximal autocorrelation of lag 1/weighted by atomic polarizabilities	GETAWAY
R_{7p}^+	R maximal autocorrelation of lag 7/weighted by atomic polarizabilities	GETAWAY
$HATS_v$	leverage-weighted total index/weighted by atomic van der Waals volumes	GETAWAY

Table 2: List of the molecular structural descriptors used in the model development and their physical-chemical meaning. *Type of descriptors: Geometry, Topology and Atom-Weights AssemblY (GETAWAY) descriptors, Wessex Head Injury Matrix (WHIM), and 3D-Molecule Representation of Structures based on Electron diffraction (3D-MoRSE).

PLS method

PLS regression was adopted here to develop QSAR model, for this method can analyze data with strongly collinear, noisy and numerous predictor variables [9,13]. PLS regression was carried out using the Simca-S package (Umetrics AB, Sweden). The conditions for computation were as follows: cross validation rounds=7, maximum iteration=200, missing data tolerance=50% and significance level (p) limit=0.05. Within Simca, the number of significant PLS components (model dimensionality) is determined by cross-validation. Cross-validation simulates how well a model predicts new data, and gives a statistic Q2cum (cumulative Q2, Q2 means the fraction of the total variation of the dependent variables that can be predicted by a component) for the final PLS model [14]. Q2cum is a good measure of the predictive power and robustness of the model. When Q2cum of a model is larger than 0.5, the model is believed to have a good predictive ability [15]. Besides Q2cum, model adequacy mainly was measured as the number of PLS components (A), the correlation coefficient between observed and predicted values (R), and the significance level (p). In addition, a general standard error (SE) was adopted to compare the prediction precision of different models. SE was defined like that in multiple regression analysis, i.e,

$$SE=\sqrt{\frac{\sum_{i=1}^{n}[\log REC10(observed)i-\log REC10(predicted)i]^2}{n-A-1}} \quad (1)$$

where n stands for the number of observations in the training set.

In PLS-regression modeling, a predictor variable may be important for the modeling of Y. Such variables are identified by large PLS-regression coefficients. However, a variable also may be important for modeling of X, which is identified by large loadings. A summary of the importance of an X-variable for both Y and X is given by a parameter, variable importance for the projection (VIP), which is a weighted sum of squares of the PLS-weights, with the weights calculated from the amount of Y-variance of each PLS component [9]. Therefore, terms with large values of VIP are the most relevant for explaining the dependent variable.

All the predictor variables are not necessarily to be included in a PLS model. Inclusion of redundant variables may lead to PLS models with low statistical significance [5]. Accordingly, the following PLS analysis process was followed to obtain an optimal model. First, a PLS model with all the predictor variables was calculated. Then each variable was eliminated and new PLS analysis was performed, leading to a series of new PLS models. The one with the largest Q2cum was selected. If there were several models with the same Q2cum, the model was selected that eliminated

the variable for which the VIP was the lowest in the previous model. This procedure was repeated until two predictor variables were left. Finally, the model with the largest Q2cum was selected as the optimal PLS model.

SVM method

SVM is a new and very promising classification and regression method developed by Vapnik et al. [16]. A detailed description of the theory of SVM can be referred in several excellent books and tutorials [17,18]. SVMs are originally developed for classification problems; they can also be extended to solve nonlinear regression problems by the introduction of Vapnik's ε-insensitive loss function. The SVM method has a number of interesting properties, including an effective avoidance of overfitting, which improves its ability to build models using large numbers of molecular proterty descriptors with relatively few experimental results in the training set. The application of SVM in regression can be expressed in the following way [14,19,20]:

Suppose the training data,

$$T=\left\{(x^1,y^1),(x^2,y^2),\cdots,(x^k,y^k)\right\},x\in R^n,y\in R \qquad (2)$$

where xm is the independent variables assembly of No. m sample (n-dimensional); ym is the independent variables assembly of No. m sample which is a measured value; k is the total number of training set. The kernel idea of SVM algorithms is to make a regression hyper plane, which can do the best to fit samples in space. The linear function is formulated in the high dimensional feature space, with the form of function:

$$y=f(x)=w\Phi(x)+b \qquad (3)$$

where Φ(x) is the high dimensional feature space, which is nonlinearly mapped from the input space x, w is the weight vector to be identified in the function, and b is the threshold.

The optimal regression function is given by the minimum of the functional,

$$\Phi(w,\xi)=\frac{1}{2}\|w\|^2+C\sum_i\xi_i^-+\xi_i^+ \qquad (4)$$

where C is a pre-specified value, and ξ_i^-,ξ_i^+ are slack variables representing upper and lower upper and lower constraints on the outputs of the system.

$$L_\varepsilon(y)=\begin{cases}0 & for\left|f(x)-y\right|<\varepsilon\\ \left|f(x)-y\right|-\varepsilon & otherwise\end{cases} \qquad (5)$$

(5)In this work, an ε -insensitive loss function was used which can be presented in Figure 1.

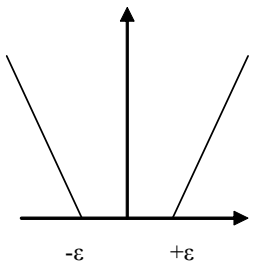

Figure 1 ε-insensitive

Figure 1: ε-insensitive.

Based on the ε-insensitive loss function and Lagrange function, the original fitting problems can be transformed as the corresponding dual Lagrangian form, which can be given by,

$$\max_{a,a^*}W(a,a^*)=\max_{a,a^*}-\frac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{k}(a_i-a_i^*)(a_j-a_j^*)\langle x_i,x_j\rangle+\sum_{i=1}^{k}a_i(y_i-\varepsilon)-a_i^*(y_i+\varepsilon) \qquad (6)$$

Or alternatively,

$$\overline{a},\overline{a}^*=\underset{a,a^*}{\operatorname{argmin}}\frac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{k}(a_i-a_i^*)(a_j-a_j^*)\langle x_i,x_j\rangle-\sum_{i=1}^{k}(a_i-a_i^*)y_i+\sum_{i=1}^{k}(a_i+a_i^*)\varepsilon \qquad (7)$$

With constrains,

$$0 \le \alpha_i, \alpha_i^* \le C, i = 1, \ldots, k$$

$$\sum_{i=1}^{k} (\alpha_i - \alpha_i^*) = 0 \quad (8)$$

Solving Eq. (6) with constraints Eq. (8) determines the Lagrange multipliers: α_i, α_i^*, and the regression function is given by Eq.(3), where

$$\overline{w} = \sum_{i=1}^{k} (\alpha_i - \alpha_i^*) x_i \quad (9)$$

$$\overline{b} = -\frac{1}{2} \langle \overline{w}, (x_r + x_s) \rangle$$

Finally, considering kernel function K(x, xi), the space transformation of inner product operation can be realized. By introducing Lagrange multipliers and exploiting the optimality constraints, decision function can take the following form:

$$f(x) = \sum_{i=1}^{m} (\alpha_i^* - \alpha_i) K(x \cdot x_i) + b^* \quad (9)$$

Where α_i^* and α_i are the introduced Lagrange multipliers.According to Karush–Kuhn–Tucker (KKT) conditions, only a number of coefficients among α_i^* and $\alpha_i$will be nonzero, and the data points corresponding to them could be defined as support vectors, which can determine the hyper plane [20]. In this equation, K(x, xi) refers to kernel function, including linear, polynomial, radial basis function (RBF), and sigmoid function.

The regression performance of SVM depends on the combination of several parameters [19]. They are penalty factor C, ε of the ε-insensitive loss function, the kernel type, and its corresponding parameters. The penalty factor C is a regularization parameter that controls the tradeoff between maximizing the margin and minimizing the training error. If C is too small, hen insufficient stress will be placed on fitting the training data. If C is too large, then the algorithm will over fit the training data. The optimal value for ε depends on the type of noise present in the data, which is usually unknown. Even if enough knowledge of the noise is available to select an optimal value for ε, there is the practical consideration of the number of resulting support vectors. "ε-insensitivity" prevents the total training set meeting boundary conditions and so allows for the possibility of sparsely in the dual formulation's solution. So, choosing the appropriate value of ε is critical from theory. The kernel type is another important one. In our study, the Gaussian radial basis function is selected, because it has only one kernel parameter and has been commonly used in regression, shown as below:

$$K(x, x_i) = \exp\left(-\frac{\|x - x_i\|^2}{2\sigma^2}\right) \quad (11)$$

The kernel parameter σ controls the amplitude of the Gaussian function and controls the generalization ability of SVM. We have to optimize σ and find the optimal one. So we should take effective and reliable measures to set the three parameters in RBF-SVM. In this study, Random Search Technique is proposed.

The overall performance of SVM is evaluated in terms of R and a root-mean-squared error (RMSE) according to the equation below

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_k - \hat{y}_k)^2}{n-1}}$$

(12)Where y_k is the desired output, $\hat{y}_k$ is the actual output of the SVM model, and n is the number of compounds in analyzed set.

The SVM model in our present study was implemented using the software LibSVM that is efficient software for classification and regression developed by Chih-Chang and Chih-Jen Lin [21]. All the algorithms used in this study were written in Matlab 7.0 and run on a personal computer (Intel Celeron-420 processor /1.66GHz 512MB RAM).

Results

PLS model

After the FS regression, thirteen molecular structural descriptors are obtained and are listed in Table 2 with their physical-chemical meanings.

The PLS regression was used to perform regression analysis, with logREC10 as a dependent variable and thirteen selected 3D descriptors as independent variables. According to the variable selection procedure mentioned above, 8 descriptors (Mor03p, L3e, R8p, RTv+, R8e, R1p+, R7p+ and HATSv) were obtained. The predicted logREC10 values by the PLS method are given in Table 1, and the statistical values of Q2cum, SE and correlation coefficient R are shown in Table 3. The linear function [3] was built as follows, with parameters defined in Table 2:

logREC10=8.253 + 0.282Mor03p - 1.820L3e - 0.988R8p + 3.284RTv+ - 2.774R8e

+ 0.567R1p+ - 11.915R7p+ - 0.149HATSv (3)

n=55, A=2, R2X(adj)(cum)=0.587, R2Y(adj)(cum)=0.757

Where A is the number of PLS components, R2X(adj)(cum) and R2Y(adj)(cum) stand for cumulative variance of all the predictor variables and dependent variable, respectively, explained by all extracted components. So it can be concluded that two PLS components were selected in the QSAR model (Eq. 3), and the two PLS components explained 58.7% of the variance of the independent variables, and 75.7% of the variance of the dependent variable. R2 Y (adj)(cum) should act as a criterion for optimal variable selection since they describe the performance of models.

	Model development			Model validation		
	Q^2_{cum}	r^2	SE	Q^2_{cum}	r^2	SE
PLS	0.678	0.757	0.765	0.664	0.733	0.870
SVM	-	0.888	0.527	-	0.875	0.743

Table 3: The statistical values of Q^2_{cum}, SE and correlation coefficient R.

Q2cum value of our universal QSAR model is as high as 0.678, indicating that the model has good predictive ability and robustness. Figure 1 shows these predicted values of $logREC_{10}$ versus experimental values. Concerning the goodness of fit of the model, the correlation coefficient (R) between the observed and the predicted logREC10 with multiple correlation coefficients is 0.870. All the absolute residuals are less than 3×SE and it indicated that there were not outliers (Figure 2). Therefore it can be concluded that the fitting results are satisfactory.

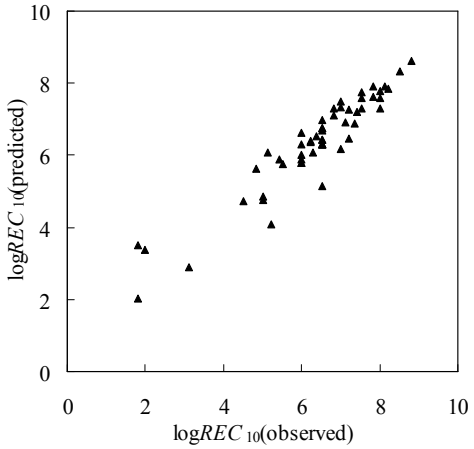

Figure 2: Plot of observed logREC10 values vs the values predicted by Eq. (3).

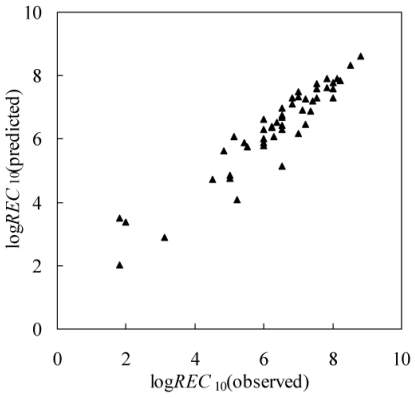

Figure 3: Plot of predicted logREC10 values vs the residuals by Eq. (3).

All the predictor variables are listed in Table 4. The VIP values indicate the significance of the variable in explaining the variance of the dependent variable.

Descriptors	VIP	w*c[1]	w*c[2]
L_{3e}	1.478	-0.548	-0.338
Mor_{03p}	1.336	0.495	-0.029
R_{7p}^{+}	1.068	-0.380	0.249
R_{8e}	0.989	-0.329	-0.590
$HATS_{v}$	0.941	-0.314	0.380
R_{Tv}^{+}	0.668	0.248	0.148
R_{8p}	0.650	-0.206	-0.442
R_{1p}^{+}	0.397	0.023	0.444

Table 4: VIP values and PLS weights. a) The bold-faced numerical values are larger than 0.3, indicating the PLS components are mainly loaded on the corresponding variables.

SVM model

c=64.0, g=0.0078125, p=0.0625

R2=0.888, MSE=0.021, SE=0.527

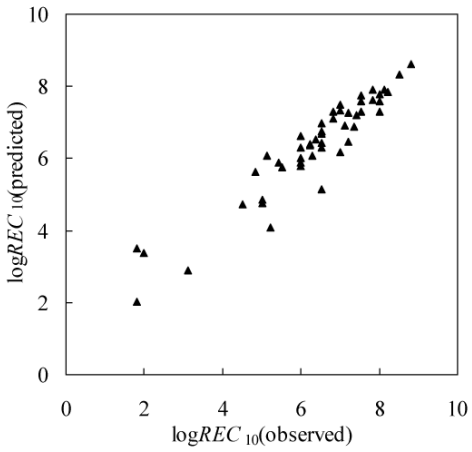

Figure 4: Plot of observed $logREC_{10}$ values vs the values predicted by SVM.

Model validation

Model validation is one of the most important processes of QSAR development [22]. Any model needs to be validated before it is used for "understanding" or for predicting new events such as the biological activities of new compounds or the yield and impurities at other process conditions. Many researchers apply the leave-one-out (LOO) or leave-some-out (LSO) cross-validation procedures. Another widely used approach to estimate model robustness is the so called y-randomization [23]. However, some researches suggested that the only way to estimate the true predictive power of a QSAR model is to compare the predicted and observed activities of an external test set of compounds that were not used in the model development [23-25].

To validate the developed QSAR model, approximately 60% of the compounds under study were selected randomly and used to develop a new PLS model using the same descriptors, then the new model was used to predicate the log RP values of the remaining 40% compounds. The procedure was repeated 10 times, and the final results are shown in Table 3. From the results, it indicates that the developed QSAR models have good robustness and predictive ability.

Discussion

Comparison of the results

Figure 4 gives the comparison between the results obtained by PLS and SVM based on the SE. As shown in Table 3, the SVM model gives the highest correlation coefficient R. It indicates that the SVM performed better than the PLS method. It also showed the better generalization ability. The reason may be that the SVM method embodies the structural risk minimization principle which minimizes an upper bound of the generalization error rather than minimizes the training error. This eventually leads to better generalization than neural networks, which implement the empirical risk minimization principle and may not converge to global solutions.

Mechanism interpretation

From a practical point of view, interpreting the descriptors used in the models could provide some insight into factors that are likely to govern the ER binding of EDCs and help us to understand which interactions may play an important role in the binding process.

Model (3) extracts two PLS components that are relevant to 8 predictor variables. The factors governing logREC10 can be interpreted by PLS weights of the variables included in model (3). The respective weights of the 8 calculated descriptors retained in the PLS model are shown in Table 4. From the PLS weights, one could see how much one descriptor contributes to the interpretation of the variance of estrogenic activity and how they relate to each other.

The first PLS component is loaded primarily on the five descriptors, L3e, Mor03p, R7p+, R8e and HATSv (Mor03p on the positive side, L3e, R7p+, R8e and HATSv on the negative). 3D-MoRSE descriptors appearing in the model are important because they take into account the 3D arrangement of the atoms without ambiguities (in contrast with those coming from chemical graphs), and also because they do not depend on the molecules with great structural variance and being a characteristic common to all of them. This type of indices are based on the idea of obtaining information from the 3D atomic coordinates by the transform used in electron diffraction studies for preparing theoretical scattering curves [26]. In order to take into account the specific contributions of the atoms to the property being studied, different atomic properties can be employed as weighting schemes. Mor03p corresponds to signal 03 and is weighted by atomic polarizabilities. On the other hand, for L3e, R7p+, R8e and HATSv, W*[1] and the corresponding coefficient in model (3) are both negative. L3e is a WHIM descriptor weighted by atomic Sanderson electronegativities, and it remarkably governs logREC10, as indicated by its VIP, the largest among the predictor variables. R7p+ and R8e are R-GETAWAY descriptors, which are derived by combines the information provided by the molecular influence matrix with geometric interatomic distances in the molecule [27]. The negative PLS weights W*[1] and coefficient of R7p+ and R8e in the model (3) indicate the negative correlation relationship between them and logREC10. This type of elaborated 3D descriptors is able to determine the shape and size of the inhibitor. The descriptor R7p+ is R maximal autocorrelation of lag 7 weighted by atomic polarizabilities and R8e is autocorrelation of lag 8 weighted by atomic Sanderson electronegativities. HATSv is an H-GETAWAY descriptor, which encodes both the geometrical information given by the molecular influence matrix H and the topological information given by the molecular graph, weighted by selected atomic weights. The selected descriptor HATSv is a leverage-weighted total index weighted by atomic van der Waals volumes.

The second PLS component that also extract five descriptors, L3e, R8e, HATSv, R8p and R1p+. The negative PLS weights W* [2] and coefficient of L3e, R8e and R8p in the model (3) also indicate the negative correlation relationship between them and logREC10. R1p+ is also an R-GETAWAY descriptor weighted by atomic polarizabilities.

In conclusion, the molecular descriptors most frequently selected by FS regression can be used to predict the logREC10 value of organic chemicals. The estrogenic activity is related to distributed atomic Sanderson electronegativities, atomic polarizabilities and atomic van der Waals volumes.

Local QSAR models

As shown in Table 1, the chemicals were divided into six "families" based on their structural characters. They were: (1) natural products and related compounds; (2) medicines, food additives, and related compounds; (3) PCBs, PCDFs, PAHs, and related compounds; (4) Phenols; (5) Benzenes and heterocyclics; (6) Phthalates and adipates. To increase our understanding of the structural requirements for a chemical's binding to ER, three linear QSAR models for three of the families were developed based on the PLS regression method, which are listed in the Table 5.

(1) natural products and related compounds

$$logREC_{10}=29.602 + 24.545E_{1e} - 29.320H_{1v} - 30.144\ R_{4e} \quad (5)$$

$n=12$, $Q^2_{cum}=0.877$, $R=0.957$, $SE=0.549$, $p < 0.0001$

(2) medicines, food additives, and related compounds:

$$logREC_{10}=7.913 + 1.910\ RDF_{020m} - 6.706L_{3p} \quad (6)$$

$n=7$, $Q^2_{cum}=0.934$, $R=0.989$, $SE=0.420$, $p < 0.0001$

(3) phenols:

$$logREC_{10}=16.474 + 0.040MWC_{01} - 1.846GATS_{4v} - 1.710GATS_{1e} - 1.423Mor_{02m}$$

$+ 0.807Mor_{21e} + 0.695E_{2e} - 8.037A_s + 0.009G_m - 0.668V_m - 0.878HATS_{6p}$ (7)

$n=25$, $Q^2_{cum}=0.915$, $R=0.983$, $SE=0.157$, $p < 0.0001$

Inspecting the knowledge obtained above, it is possible to gain some information about what factors are likely to govern the ER binding ability for a specific family. This is beneficial for developing a credible model for prediction.

Conclusion

The two methods, PLS and SVM, were used to develop linear and nonlinear QSARs to predict estrogenic activities of 55 structurally diverse organic chemicals. Eight descriptors, which represent the features of the compounds responsible for the binding ability to estrogen receptor, were selected to develop global QSAR models. Inspection of the PLS model indicates that atomic Sanderson electronegativities, polarizabilities and van der Waals volumes may be most relevant factor controlling the binding behavior, affecting the space-matching between the ER protein and the ligand. The two resultant QSAR models were further compared with respect to statistical measures from a leave-some-out process, with SVM yielding the best model performance in terms of self-consistency and ability to predict the activity of the test chemicals. Additionally, 55 chemicals were divided into six well-known "families" according to their chemical structures. Three local QSAR models were developed, which gave us insight into the factors that govern the binding behavior for these specific chemicals.

References

1. Zacharewski T (1997) In vitro bioassays for assessing estrogenic substances. Environmental Science & Technology 31: 613-623.
2. Fang H, Tong WD, Branham WS, Moland CL, Dial SL, et al. (2003) Study of 202 natural, synthetic, and environmental chemicals for binding to the androgen receptor. Chem. Res. Toxicol 16: 1338-1358.
3. Waller CL, Oprea TI, Chae K, Park HK, Korach KS, et al. (1996) Ligand-based identification of environmental estrogens. Chem Res Toxicol 9: 1240-1248.
4. Tong WD, Perkins R, Xing L, Welsh WJ, Sheehan DM (1997) QSAR models for binding of estrogenic compounds to estrogen receptor alpha and beta subtypes. Endocrinology 138: 4022-4025.
5. Tong W, Lowis DR, Perkins R, Chen Y, Welsh WJ, et al. (1998) Evaluation of quantitative structure-activity relationship methods for large-scale prediction of chemicals binding to the estrogen receptor. J Chem Inf Comput Sci 38: 669-677.
6. Wiese TE, Polin LA, Palomino E, Brooks SC (1997) Induction of the estrogen specific mitogenic response of MCF-7 cells by selected analogues of estradiol-17 beta: A 3D QSAR study. J Med Chem 40: 3659-3669.
7. Costantino G, Macchiarulo A, Camaioni E, Pellicciari R (2001) Modeling of poly(ADP-ribose)polymerase (PARP) inhibitors. Docking of ligands and quantitative structure-activity relationship analysis. J Med Chem 44: 3786-3794.
8. Nishihara T, Nishikawa J, Kanayama T, Dakeyama F, Saito K, et al. (2000) Estrogenic activities of 517 chemicals by yeast two-hybrid assay. Journal of Health Science 46: 282-298.
9. Wold S, Sjostrom M, Eriksson L (2001) PLS-regression: a basic tool of chemometrics. Chemometr Intell Lab 58: 109-130.
10. Coldham NG, Dave M, Sivapathasundaram S, McDonnell DP, Connor C, et al. (1997) Evaluation of a recombinant yeast cell estrogen screening assay. Environ Health Perspect 105: 734-742.
11. Todeschini R (2000) Handbook of molecular descriptors. Wiley-VCH, Weinheim, Germany, 2000.
12. Morales AH, Duchowicz PR, Perez MAC, Castro EA, Cordeiro MNDS, et al. (2006) Application of the replacement method as a novel variable selection strategy in QSAR. 1. Carcinogenic potential. Chemometr Intell Lab 81: 180-187.
13. Ding GH, Chen JW, Qiao XL, Huang LP, Lin J et al. (2006) Quantitative relationships between molecular structures, environmental temperatures and solid vapor pressures of PCDD/Fs. Chemosphere 62, 1057-1063.
14. Yao XJ, Panaye A, Doucet JP, Zhang RS, Chen HF, et al. (2004) Comparative study of QSAR/QSPR correlations using support vector machines, radial basis function neural networks, and multiple linear regression. J Chem Inf Comput Sci 44: 1257-1266.
15. Golbraikh A, Tropsha A (2002) Beware of q(2)! J Mol Graph Model 20: 269-276.
16. Vapnik (1998) Statistical Learning Theory. John Wiley & Sons: New York.
17. Cristianini NS (2000) An Introduction to Support Vector Machines. Cambridge University Press: Cambridge, UK.
18. Schölkopf BS, Alexander JS (2002) Learning with Kernels. MIT Press: Cambridge, MA.
19. Liu HX, Xue CX, Zhang RS, Yao XJ, Liu MC, et al. (2004) Quantitative prediction of logk of peptides in high-performance liquid chromatography based on molecular descriptors by using the heuristic method and support vector machine. J Chem Inf Comput Sci 44: 1979-1986.
20. Xue CX, Zhang RS, Liu HX, Yao XJ, Liu MC, et al. (2004) QSAR models for the prediction of binding affinities to human serum albumin using the heuristic method and a support vector machine. J Chem Inf Comput Sci 44: 1693-1700.
21. Chang CCL, CJ LIBSV- A Library for Support Vector Machines.
22. Golbraikh A, Shen M, Xiao Z, Xiao YD, Lee KH, (2003) Rational selection of training and test sets for the development of validated QSAR models. J Comput Aided Mol Des 17: 241-253.
23. Golbraikh A, Tropsha A (2002) Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. 16: 357-369.
24. Norinder U (1996) Single and domain mode variable selection in 3D QSAR applications. Journal of Chemometrics 10: 95-105.
25. Zefirov NS, Palyulin VA (2001) QSAR for boiling points of "small" sulfides. Are the "high-quality structure-property-activity regressions" the real high quality QSAR models? J Chem Inf Comput Sci 41: 1022-1027.
26. Duchowicz PR, Fernández M, Caballero J, Castro EA, Fernández FM (2006) QSAR for non-nucleoside inhibitors of HIV-1 reverse transcriptase. Bioorganic & Medicinal Chemistry 14: 5876-5889.

27. Consonni V, Todeschini R, Pavan M (2002) Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors. J Chem Inf Comput Sci 42, 682-692.