ARTICLES

Variable Selection and Interpretation in Structure−Affinity Correlation Modeling of Estrogen Receptor Binders

Federico Marini,[†,‡] Alessandra Roncaglioni,[†,§] and Marjana Novič*[,†]

National Institute of Chemistry, Ljubljana, Slovenia, University of Rome "La Sapienza", Rome, Italy, and Institute Mario Negri, Milan, Italy

Received April 29, 2005

A computational approach for the identification and investigation of correlations between a chemical structure and a selected biological property is described. It is based on a set of 132 compounds of known chemical structures, which were tested for their binding affinities to the estrogen receptor. Different multivariate modeling methods, i.e., partial least-squares regression, counterpropagation neural network, and error-back-propagation neural network, were applied, and the prediction ability of each model was tested in order to compare the results of the obtained models. To reduce the extensive set of calculated structural descriptors, two types of variable selection methods were applied, depending on the modeling approach used. In particular, the final partial least-squares regression model was built using the "variable importance in projection" variable selection method, while genetic algorithms were applied in neural network modeling to select the optimal set of descriptors. A thorough statistical study of the variables selected by genetic algorithms is shown. The results were assessed with the aim to get insight to the mechanisms involved in the binding of estrogenic compounds to the receptor. The variable selection on the basis of genetic algorithm was controlled with the test set of compounds, extracted from the data set available. To compare the predictive ability of all the optimized models, a leave-one-out cross-validation procedure was applied, the best model being the nonlinear neural network model based on error back-propagation algorithm, which resulted in $R^2 = 92.2\%$ and $Q^2 = 70.8\%$.

INTRODUCTION

Endocrine-disrupting chemicals (EDCs) present an increasing concern[1] because of their adverse effects on human and wildlife (http://www.who.int/pcs/emerg_site/edc/global_edc_TOC.htm). The effects in humans are not only reproductive and developmental disruptions. People exposed to EDCs can develop cancer; they can suffer from altered immune and nervous systems and thyroid function. Most of the studies found in the literature are focused on the receptor-mediated mechanism of action of EDCs, although it is well-known that other mechanisms of action, such as hormone synthesis, transport, and metabolism are equally important. The information about biological activity obtained from in vitro and in vivo tests is exploited in theoretical studies of the relationship between chemical structure and biological activity. Alternative methods, i.e., computer-based prediction and modeling systems for specific toxicological endpoints, based on the achievements of the structure−property relationship studies, are becoming widespread because of the increasing need for screening a large number of chemicals for their adverse biological activity. In particular, many of these various models have been developed to predict endocrine disruption activity.[2−6]

We have chosen a set of chemicals tested for their relative binding affinities (RBA) to rats' uteri estrogen receptor in order to study the relationship between this property and the compounds' chemical structure.[7] An automated experience-based predictive model was developed, which is able to predict the RBA on the basis of molecular structure of an unknown compound, provided that its structure is within the structural diversity encompassed by the studied set of compounds. Several modeling methods were applied (Partial Least Squares-Regression,[8−10] Counterpropagation Artificial Neural Network,[11,12] and Back-propagation Artificial Neural Network[13]), and the best method was chosen on the basis of cross-validation results. A common problem in structure−property relationship models is how to choose an optimal set of structural descriptors. A large number of structural descriptors[14] are usually generated by any of the existing software.[15−17] To avoid the noise introduced by the descriptors that do not influence the studied structure−property relationship and to improve the robustness and the generalization-ability of the constructed model, the variable selection method based on genetic algorithm was applied (see the Methods section for a detailed description).[18−20]

DATA

The data set of 132 compounds was compiled from the information obtained from literature[7] and available on the Internet (http://edkb.fda.gov/databasedoor.html). The values for RBA, i.e., relative binding affinities to the estrogen

* Corresponding author phone: 386 1 4760 253; fax: 386 1 4760 300; e-mail: marjana.novic@ki.si.
† National Institute of Chemistry.
‡ University of Rome "La Sapienza".
§ Institute Mario Negri.

10.1021/ci0501645 CCC: $30.25
Published on Web 10/21/2005

Table 1. Set of Molecules Used for Modeling with Observed[7] (for ID Labeling See Supporting Information) and Predicted LogRBA Values[a]

ID	LogRBA observed	LogRBA (PLS)	LogRBA (CP-ANN)	LogRBA (BP-ANN)	ID	LogRBA observed	LogRBA (PLS)	LogRBA (CP-ANN)	LogRBA (BP-ANN)
2	−0.36	−1.58	−1.87	−1.17	166	−3.13	−1.10	−3.13	−2.67
3	−0.82	−1.14	−0.82	−1.10	167	−1.51	−0.56	−0.15	−0.69
4	−1.65	−1.54	−1.24	−0.77	170	−3.67	−1.56	−3.67	−3.19
5	−2.98	−1.04	−2.98	−2.80	171	−3.04	−3.37	−3.24	−2.82
6	−2.34	−1.38	−2.25	−2.29	172	1.18	0.47	1.71	1.27
12	1.48	−0.36	0.64	0.31	173	−1.07	−1.34	−1.07	−0.39
15	−2.35	−1.50	−1.39	−2.55	174	0.14	0.49	−0.05	−0.25
18	−2.78	−2.27	−2.78	−2.78	175	−3.26	−2.77	−3.26	−2.76
23	−0.60	−1.75	−1.53	−0.73	177	−3.02	−2.29	−3.02	−2.69
24	−4.09	−1.55	−4.09	−4.25	203	−4.17	−4.30	−4.33	−4.05
26	−3.41	−2.28	−3.41	−2.78	267	−2.35	−1.40	−1.34	−1.31
30	−3.54	−2.90	−3.54	−3.82	269	−2.55	−2.18	−2.49	−2.02
31	−3.37	−2.81	−3.49	−3.63	332	−3.38	−4.30	−3.57	−3.48
43	−2.78	−1.91	−3.05	−3.75	346	−4.50	−5.90	−4.33	−4.07
44	−2.85	−1.63	−2.85	−3.14	411	−2.43	−2.24	−2.49	−2.41
50	−0.64	−1.74	−1.93	−0.87	420	−2.69	−2.01	−3.07	−2.29
54	−1.44	−2.69	−1.44	−1.86	421	−1.44	−1.96	−1.00	−0.99
55	−3.25	−1.98	−1.93	−3.68	442	−3.44	−3.43	−3.24	−3.18
56	−2.77	−2.96	−2.77	−2.40	481	−0.55	−1.79	−1.00	−0.72
57	−2.18	−2.74	−2.18	−2.52	495	−1.26	−1.92	−1.26	−1.41
62	−3.61	−3.38	−3.49	−3.54	500	−0.82	−1.62	−1.24	−1.13
71	−1.89	−3.43	−1.89	−2.19	514	1.16	−1.38	−0.47	1.17
73	−1.82	−1.94	−1.82	−2.50	527	1.31	−0.44	0.96	1.41
77	2.60	−0.76	2.21	2.06	531	−0.30	−0.38	0.43	−1.01
81	−2.11	−1.62	−0.47	−2.02	532	1.14	−0.66	0.43	1.08
82	−2.89	−1.31	−2.15	−3.05	601	−3.87	−4.39	−3.87	−3.95
83	2.00	−0.16	1.24	1.29	612	−0.89	−1.29	−0.89	−0.89
84	0.99	−0.19	1.03	0.81	613	−2.88	−1.32	−2.25	−2.76
85	0.86	−0.71	0.29	0.48	614	−0.64	−1.16	−0.64	−0.47
86	2.28	0.37	0.92	1.89	615	0.42	−1.65	0.42	0.23
87	1.57	−0.54	1.36	1.53	617	−2.65	−1.87	−3.05	−3.20
91	1.47	−0.02	1.03	0.90	624	−1.25	−0.08	−1.25	−1.81
92	−1.53	−1.20	−2.25	−1.96	625	−0.29	−0.87	0.29	0.99
93	1.82	0.05	1.03	0.80	626	−2.74	−0.11	−2.74	−2.68
94	1.14	0.74	0.92	1.01	627	0.97	−0.50	0.96	0.70
95	1.16	0.94	1.36	0.74	628	−2.20	−0.80	−2.20	−2.30
96	0.49	−0.15	1.24	1.10	629	−1.65	0.17	−1.56	−1.34
97	−2.67	−1.13	−2.15	−2.23	630	−1.48	0.37	−1.56	−1.66
98	−0.92	−1.14	−2.15	−2.17	631	0.60	−0.70	0.96	0.49
105	2.48	−0.77	2.21	1.46	632	−0.68	−1.66	−1.39	−1.65
106	1.57	−0.97	2.21	1.80	633	1.19	−0.51	−0.15	0.74
107	0.21	0.61	0.21	−0.13	634	1.11	−0.07	1.11	2.45
108	2.24	0.65	1.71	1.24	635	−0.05	−0.23	−0.05	0.15
109	−0.14	0.52	−0.05	−0.27	636	−0.02	−0.62	−0.02	0.08
110	−0.14	1.57	−0.05	0.30	637	−3.44	−2.16	−3.07	−3.35
111	−1.49	−1.33	−1.49	−2.11	638	−1.73	−0.90	−1.73	−1.69
113	−3.35	−1.98	−1.87	−3.19	639	−3.07	−1.56	−3.15	−2.66
114	0.35	0.66	0.35	0.07	640	−3.67	−4.05	−3.57	−3.76
115	−1.55	−1.58	−1.84	−2.29	641	−3.61	−2.67	−3.61	−3.76
116	−2.13	−1.52	−1.84	−1.87	642	−3.66	−4.21	−3.57	−3.40
117	−1.16	−1.42	−1.39	−1.66	646	−0.15	−0.10	1.03	0.39
118	−2.37	−1.17	−2.56	−2.44	647	−0.67	0.15	0.92	−1.30
119	0.32	−0.85	0.32	−0.16	654	−2.09	−0.92	−1.91	−1.85
121	−1.61	−1.48	−1.61	−1.29	656	−1.74	−0.81	−1.91	−0.89
122	−3.35	−1.55	−3.05	−2.98	660	−2.54	−2.68	−2.54	−2.50
123	−0.05	−1.36	−0.05	−0.88	661	−3.22	−1.93	−3.15	−3.00
124	−2.75	−1.68	−3.05	−3.26	662	−3.22	−2.55	−3.33	−3.34
132	1.63	−0.05	0.47	0.97	663	−3.44	−3.35	−3.33	−3.91
136	−2.61	−2.57	−2.61	−2.61	664	−2.74	−1.16	−2.56	−2.49
137	−1.87	−1.56	−1.87	−1.96	666	−2.82	−2.67	−2.82	−2.47
138	−3.25	−3.05	−3.16	−3.56	667	−0.19	−0.30	0.64	0.78
139	−3.07	−4.29	−3.16	−3.12	668	−0.69	−0.02	0.47	−0.14
155	−2.46	−2.73	−2.46	−2.81	675	−2.30	−1.92	−2.30	−2.65
159	−3.05	−1.69	−3.05	−3.31	705	−3.16	−2.32	−3.16	−3.25
160	−3.73	−1.62	−3.05	−3.00	708	−0.35	−1.69	−1.34	−0.64
165	−2.45	−1.92	−1.53	−2.01	2017	−2.74	−0.06	−2.74	−2.09

[a] Predictions are obtained by three different models reported in this study.

receptor, as well as the chemical structures were acquired for each molecule. The CAS numbers, LogRBA, and ID number of the molecules are given in Table 1. The MOPAC computational program was applied to optimize chemical structures using a well-documented AM1 semiempirical approach.[21] We used the CODESSA program[15] to calculate the descriptors of various types, constitutional, topological, geometrical, electrostatic, and quantum-chemical. With the applied software 280 descriptors were determined as numerical values, available for all the compounds in the data

set (Var_002 − Var_281). An additional parameter, LogP (Var_001), which reflects the molecular hydrophobic properties, was added. The LogP values[22] were obtained from the experimental database (http://www.syrres.com/esc/physprop.htm) or estimated with the KowWin program (http://www.syrres.com/esc/kowwin.htm). All descriptors were auto scaled (i.e. normalized with mean = 0 and standard deviation = 1). They are available from the authors upon request. The names of descriptors together with their numbers used in this study (Var_001-Var_281) are available in the Supporting Information.

METHODS

Partial Least Squares-Regression (PLS-R).[8−10] Partial least squares projection to latent structures is one of the most used regression techniques in QSAR. It consists of finding a subspace common to both the predictor (**X**) and the dependent variable (**Y**) blocks, by decomposing them in a single process

$$\mathbf{X} = \mathbf{TP}' + \mathbf{D}$$

$$\mathbf{Y} = \mathbf{UQ}' + \mathbf{F} \quad (1)$$

where **T** and **U** are the **X**- and **Y**-block score matrices, **P** and **Q** are the **X** and **Y** loadings, and **D** and **F** are the residuals. Then a linear inner relationship is modeled between the projections of the dependent and independent variables (**U** and **T**, respectively), according to

$$\mathbf{U} = \mathbf{BT} \quad (2)$$

Even if the linear model might appear an oversimplification of the actual relationship between the two blocks of variables, in many cases it leads to a high accuracy and predictive ability, as widely reported in the literature.[23]

Counterpropagation Artificial Neural Networks (CP-ANN).[11−12] Counterpropagation artificial neural network is a modeling tool, which comprises elements of both, supervised and unsupervised learning, consisting of two layers: a Kohonen and an output layer. The representation vector of each object ($\mathbf{x_s}$) is input to the network. In each iteration, the unit whose weight vector ($\mathbf{w_j}$) is most similar to the input ($\mathbf{x_s}$) is selected as the "winner". The weights of all the neighboring units within a topological distance d_{max} to the winning unit are updated according to

$$\Delta\mathbf{w_{ji}} = \eta\left(1 - \frac{d_r}{d_{max}+1}\right)(\mathbf{x_{si}} - \mathbf{w_{ji}}) \quad \text{for } d_r = 0, 1, ..., d_{max} \quad (3)$$

where η is the learning rate and d_r is the number of units separating the actual neuron from the winner. The maximal distance d_{max} decreases during the learning phase: while at the beginning of the training it covers the entire network, at the end of the learning it is limited only to the winning neuron ($d_{max} = 0$). As no information about the desired targets is used to correct this weight, the learning in the Kohonen layer appears to be unsupervised.

On the other hand, the weights in the output layer are updated on the basis both of the position of the winning neuron in the Kohonen layer and of the values of the output targets. In fact, the position of the winning unit in the Kohonen layer is used to define the neighborhood to be used in the correction of the output weights ($\mathbf{u_j}$), which is then carried out using a formula analogous to eq 3, but taking into account the target values ($\mathbf{t_s}$) instead of the inputs:

$$\Delta\mathbf{u_{ji}} = \eta\left(1 - \frac{d_r}{d_{max}+1}\right)(\mathbf{t_{si}} - \mathbf{u_{ji}}) \quad \text{for } d_r = 0, 1, ..., d_{max} \quad (4)$$

It is apparent from the last equation, that in the output layer the patterns are learned in a supervised way. Furthermore, the learning rate constant η is varied during the training phase from an initial maximum value η^{start} to a minimum, η^{final}, which is reached when the number of epochs n_{epochs} matches a prespecified maximum n_{tot}:

$$\eta = (\eta^{start} - \eta^{final})\left(1 - \frac{n_{epoch}}{n_{tot}}\right) + \eta^{final} \quad (5)$$

Multilayer Feedforward Artificial Neural Networks Trained by the Back-Propagation of Error Algorithm (BP-ANN).[13] The multilayer feedforward organization of the units is the most used neural network architecture for the applications in chemistry. It consists of computational units organized into three kinds of layers, *input*, *hidden*, and *output* layers. The units (neurons) in each layer all receive the same information—an output vector (**X**) from the previous layer, and in turn send their output vector as input to the neurons in the successive layer. The output of an individual neuron is calculated as a sigmoidal function of the input signals. In the *j*th layer the output (*sf*) is calculated as follows:

$$sf_j = \frac{1}{1 + \exp(-\sum_{i=1}^{m} w_{ji}x_i)} \quad (6)$$

The units of the input layer receive their input in the form of a data file, while the units of the output layer produce the output signal, which is the overall result of the network. This multilayer architecture is often used in conjunction with the back-propagation weight update rule, according to which a supervised form of learning is implemented. The error back-propagation algorithm (BP) is essentially an iterative weight update on the basis of a steepest descent criterion, so to minimize the root-mean square error (RMS) between the desired and the actual target of the network *E*. In mathematical terms, during the training phase each weight of the network is varied according to

$$\Delta\mathbf{w_{ji}}(t) = -\eta\frac{\partial E}{\partial \mathbf{w_{ji}}} + \mu\Delta\mathbf{w_{ji}}(t-1) \quad (7)$$

η being the learning rate and μ being an additional constant called the momentum. The rightmost term, which takes into account the update of the same weight in the previous iteration (t and $t-1$ represent the tth and $(t-1)$th iteration respectively) has been introduced to avoid the algorithm to be stuck into local minima and to damp the oscillations of the solution.

To compute the partial derivative of the error *E* with respect to the connection weights to the hidden layer(s) requires propagating backward the prediction error *E* using

the rules of chain derivation, hence the name "back-propagation": chain derivation acts as a way to "distribute" the error E between the neurons of the hidden layer(s) in order to apply the iterative weight adjustment, necessary for the learning of the network.

Genetic Algorithms (GA).[18−20] Genetic Algorithms (GA) is an advanced optimization technique that mimics the selection processes occurring in nature, Darwinian evolution: crossover, mutation, and selection. The main idea is the survival of individuals having the highest fitness score; under specified conditions they are the most likely to prevail in the next generation. Moreover, crossover and random mutations introduce genetic variety to the offspring, so that an increasingly wide solution space is spanned during the computing. In this study, genetic algorithms have been used, coupled to artificial neural networks, to select the input variables to be included in the models.

In particular, computational genetic algorithm consists of four steps:

(i) At first an initial population of chromosomes has to be generated. The chromosomes are represented as binary strings of bits (zeros and ones). Fifty chromosomes were randomly generated as the initial pool of each population used in this work. The code "1" in the *i*th position of the chromosome indicates the inclusion of the *i*th variable to the modeling procedure.

(ii) The performance of each chromosome must be evaluated. With the indicated variables, a model was generated, and the fitness score has been computed as the predictive ability over the test set expressed in terms of the r_{test}, the correlation coefficient between the actual and the predicted property (target). The chromosomes are ranked according to a decreasing value of the fitness score.

(iii) In a successive stage, the best chromosome is "protected" and copied without any further modification to the next generation, while the remaining chromosomes of the child offspring are created by mutually exchanging a selected part in pairs of randomly selected chromosomes (crossover).

(iv) A random mutation is introduced to modify a single position of a chromosome, by changing the value of one of its bits.

The cycles (ii)−(iv) are repeated until a stopping criterion is satisfied; in this work a maximum number of generation condition has been used.

RESULTS AND DISCUSSION

PLS-R Model. In the first stage of our study, Partial Least Squares Regression was used to model the relationship between the logarithm of the Relative Binding Affinity (LogRBA) and the 281 predictor variables, descriptors of chemical structure. Leave-one-out cross-validation (LOO-CV) has been used to evaluate the predictive ability of the model, and both dependent and predictor variables have been auto scaled to zero mean and unit variance. Using all the 281 input variables as descriptors, initially only one significant component was extracted from the data set on the basis of the Q^2 value resulting in a model with an R_X^2 and R_Y^2 of 0.28 and 0.36, respectively, and with $Q^2 = 0.31$. Here, Q^2 is the fraction of the total variation of the dependent variable Y that can be predicted by a component according to the cross-validation, and $R_X^2$and R_Y^2 are fractions of the sums of squares (SS) of all the predictor and dependent variables explained by the chosen component (http://www.umetrics.com/download/ SIMCA-P/simcapplus10ug.pdf). Despite the insignificant second component, the use of a three-component model increased the Q^2 to 0.41, with R_X^2 and R_Y^2 rising correspondingly to 0.59 and 0.55, respectively. The correlation between the observed and the predicted values of the relative binding affinity for each of the samples is reported in Figure 1a. The inspection of the scaled regression coefficients for this model built on all the input variables did not allow a clear interpretation of the results, as many variables appeared to contribute significantly to the projection and correlate with y.

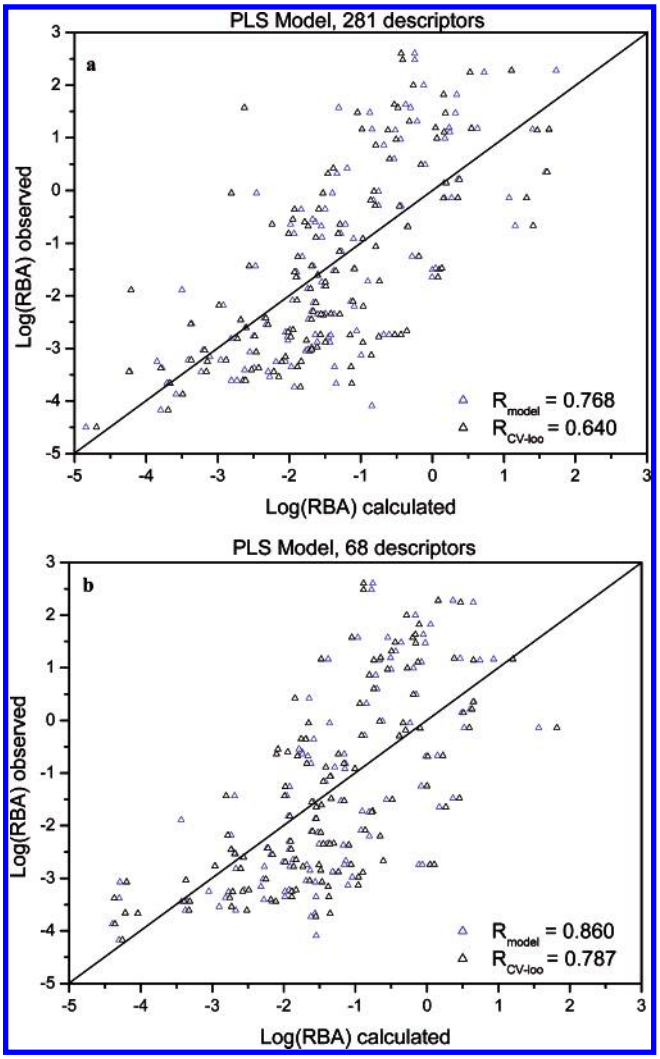

Figure 1. PLS modeling. Correlation of calculated vs observed values for the three-components PLS model built using all the variables (a) and 68 descriptors selected according to their VIP value (b). Predicted and cross-validates values are ploted in blue and black colors, respectively.

PLS-R − Variable Selection. Based on these considerations, in a successive stage, to get rid of less informative and/or redundant variables, a variable selection procedure has been used. In the procedure, the columns with a negligible variable importance in projection (VIP) are excluded from the matrix of variables (predictor set). The VIP index accounts for the contribution of each individual descriptor for explaining the y variable, summed over the model dimension, and has been computed according to

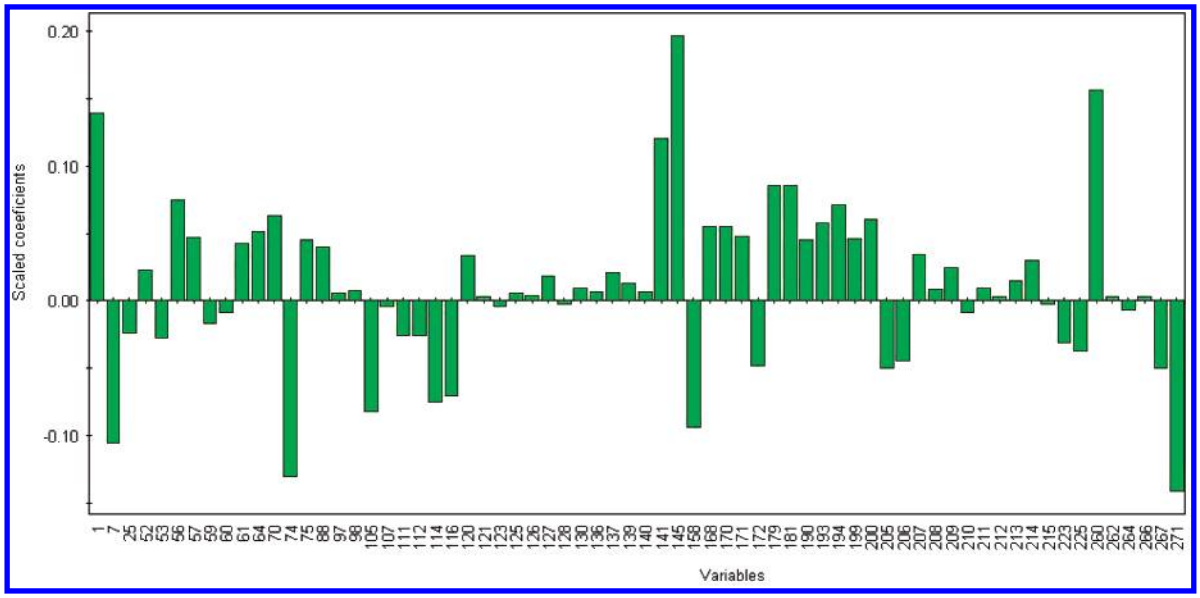

Figure 2. PLS modeling. Scaled regression coefficients for the model built on the 68 variables selected on the basis of their VIP values.

the formula reported in the SIMCA-P User's Guide Umetrics AB (http://www.umetrics.com/download/SIMCA-P/simcapplus10ug.pdf). The descriptors with a VIP value larger than 1 are the most relevant to model the variation in the dependent variable, so at first all the variables whose contribution was less than 1 were ruled out. Successively, the other variables have been iteratively excluded until an optimal model (in terms of Q^2) was obtained. At last, 68 descriptors (reported in Table 2 together with their individual VIP score) have been retained to build the optimal PLS model. The resulting three-component model showed an improved modeling ($R_X^2 = 0.74$ and $R_Y^2 = 0.64$) and predictive ($Q^2 = 0.62$) ability. For detailed predictive performance of the model see Figure 1b in which the predicted and cross-validated RBA values are compared to the observed ones.

PLS-R − Interpretation of Variables. Reducing the number of the variables has allowed higher predictive accuracy, and, moreover, also the regression coefficients are more easily interpretable. From Figure 2, in which the scaled coefficients of all the variables are plotted, the most effective descriptors can be easily identified. In particular, three factors seem to influence significantly the value of the relative binding affinity for the investigated set of compounds. The high positive coefficient for LogP (var_001) and the corresponding negative contribution from the absolute number of oxygen atoms (var_007) suggests that the polarity of the sample is involved in modulating the binding of the endocrine disrupter to the receptor site: specifically, less polar molecules, characterized by an high octanol/water partition coefficient and by a small number of oxygen atoms, are supposed to be more favored with respect to highly polar ligands. Furthermore, the highly negative coefficients corresponding to the moments of inertia along an orthogonal axis C (var_74) and the first principal axis A (var_271) indicate that small and flexible molecules are privileged over bulky ones. Last, a significant positive correlation is observed between the dependent variable and the final heat of formation of the molecule (var_141), the energy of the HOMO-1 (var_145), and the maximum total interaction for a C−C bond (var_260). This issue suggests that probably the binding of the ligand to the receptor site involves a certain degree of electron transfer. The higher the HOMO-1 energy (the second highest energy of valence electrons), the higher the binding affinity, which indicates that less bound valence electrons enhance the ligand−protein complex formation. On the other hand, the van der Waals interactions, which are present and important for the binding mechanism, are accounted for by the polarity terms described before.

Nonlinear ANN Models. To find additional confirmation of the binding mechanism hypothesis and to look for an improved performance in predicting estrogen receptor binding activity, in the successive stage two different kinds of artificial neural networks, namely counterpropagation and error-back-propagation, have been used to model the relationship between the input descriptors and the Log(RBA) value. Artificial neural networks, contrarily to PLS, are nonlinear models and, in case of nonlinear relationships, can result a more effective and flexible regression tool. In the application of both kind of neural networks, the counterpropagation (CP-ANN) and back-propagation artificial neural network (BP-ANN), initially the model was built using all the computed descriptors and validated with the leave-one-out procedure. Successively, genetic algorithms have been used to reduce the number of variables included in the model.

CP-ANN Model. Several counterpropagation networks have been tested, differing in the dimension of the Kohonen layer, the number of training epochs, and in the initial learning rate value. The networks with the best training (R) and cross-validation (Q) predictive ability are reported in Table 3, whose squared values can be compared with the R_Y^2 and Q^2 from the PLS-R model. The measure for the model cross-validation Q was obtained as the correlation between the actual and predicted Log(RBA), if the objects predicted by the leave-one-out procedure were considered.

It is apparent that many of the best networks reported in Table 3 perform significantly better than the PLS models, and some are also better in cross-validated prediction. In particular, the best CP-ANN architecture − 12 × 12 Kohonen layer, $\eta^{start} = 0.5$, $\eta^{end} = 0.1$, 300 learning epochs − led to a R_Y^2 value of 0.88 and $Q^2 = 0.62$. The

Table 2. 68 Variables (Structural Descriptors) of the PLS-R Model, Selected on the Basis of Their Importance in Prediction (VIP)

var_ no.	description	VIP
145	HOMO-1 energy	1.696
56	information content (order 1)	1.627
271	principal moment of inertia A	1.625
70	bonding information content (order 2)	1.599
267	Tot molecular 2-center exchange energy	1.598
64	information content (order 2)	1.584
168	PPSA-1 partial positive surface area [semi-MO PC]	1.584
74	moment of inertia C	1.572
52	complementary information content (order 0)	1.570
60	complementary information content (order 1)	1.563
98	WPSA-1 weighted PPSA (PPSA1*TMSA/1000) [Zefirov's PC]	1.562
260	max total interaction for a C−C bond	1.559
207	count of H-donors sites [semi-MO PC]	1.541
120	count of H-donors sites [Zefirov's PC]	1.538
1	LogP	1.538
75	XY shadow	1.523
170	DPSA-1 difference in CPSAs (PPSA1-PNSA1) [semi-MO PC]	1.522
59	average complementary information content (order 1)	1.520
114	RPCG relative positive charge (QMPOS/QTPLUS) [Zefirov's PC]	1.512
105	WPSA-2 weighted PPSA (PPSA2*TMSA/1000) [Zefirov's PC]	1.510
194	FHDSA fractional HDSA (HDSA/TMSA) [semi-MO PC]	1.508
116	RNCG relat. negative charge (QMNEG/QTMINUS) [Zefirov's PC]	1.506
200	FHDCA fractional HDCA (HDCA/TMSA) [semi-MO PC]	1.505
112	WPSA-3 weighted PPSA (PPSA3*TMSA/1000) [Zefirov's PC]	1.504
193	HDSA H-donors surface area [semi-MO PC]	1.502
25	relative number of aromatic bonds	1.502
171	FPSA-1 fractional PPSA (PPSA-1/TMSA) [semi-MO PC]	1.501
172	FNSA-1 fractional PNSA (PNSA-1/TMSA) [semi-MO PC]	1.498
199	HDCA H-donors charged surface area [semi-MO PC]	1.486
111	FNSA-3 fractional PNSA (PNSA-3/TMSA) [Zefirov's PC]	1.477
137	HACA-1/TMSA [Zefirov's PC]	1.476
57	average structural information content (order 1)	1.471
107	PPSA-3 atomic charge weighted PPSA [Zefirov's PC]	1.471
266	Tot molecular 2-center resonance energy/# of atoms	1.463
214	HA dependent HDCA-1/TMSA [semi-MO PC]	1.462
53	average bonding information content (order 0)	1.447
141	final heat of formation	1.445
61	average bonding information content (order 1)	1.432
97	FNSA-1 fractional PNSA (PNSA-1/TMSA) [Zefirov's PC]	1.431
139	HACA-2/TMSA [Zefirov's PC]	1.425
127	HA dependent HDCA-1/TMSA [Zefirov's PC]	1.420
209	HA dependent HDSA-1/TMSA [semi-MO PC]	1.411
181	WNSA-2 weighted PNSA (PNSA2*TMSA/1000) [semi-MO PC]	1.410
140	HACA-2/SQRT(TMSA) [Zefirov's PC]	1.406
213	HA dependent HDCA-1 [semi-MO PC]	1.400
130	HA dependent HDCA-2/SQRT(TMSA) [Zefirov's PC]	1.393
136	HACA-1 [Zefirov's PC]	1.389
211	HA dependent HDSA-2/TMSA [semi-MO PC]	1.388
264	Tot molecular 1-center E−E repulsion/# of atoms	1.383
208	HA dependent HDSA-1 [semi-MO PC]	1.382
262	Tot molecular 1-center E−N attraction/# of atoms	1.381
125	HA dependent HDSA-2/SQRT(TMSA) [Zefirov's PC]	1.373
126	HA dependent HDCA-1 [Zefirov's PC]	1.362
121	HA dependent HDSA-1 [Zefirov's PC]	1.359
212	HA dependent HDSA-2/SQRT(TMSA) [semi-MO PC]	1.353
128	HA dependent HDCA-2 [Zefirov's PC]	1.333
215	HA dependent HDCA-2 [semi-MO PC]	1.315
123	HA dependent HDSA-2 [Zefirov's PC]	1.307
158	avg 1-electron react. index for a C atom	1.296
210	HA dependent HDSA-2 [semi-MO PC]	1.294
7	number of O atoms	1.282
223	HACA-1 [semi-MO PC]	1.280
225	HACA-2 [semi-MO PC]	1.270
205	min(#HA, #HD) [semi-MO PC]	1.254
206	count of H-acceptor sites [semi-MO PC]	1.251
190	RPCS rel. positive charged SA (SAMPOS*RPCG) [semi-MO PC]	1.244
179	FNSA-2 fractional PNSA (PNSA-2/TMSA) [semi-MO PC]	1.241
88	polarity parameter (Qmax-Qmin)	1.240

corresponding observed vs predicted plots are reported in Figure 3a.

To look for a chemical interpretation of the result obtained by the use of counterpropagation network, both the top-map and the weight maps have been examined. The top-map is a 2D representation of the samples projected onto the neurons organized in a quadratic neighborhood called the Kohonen layer and can account for the "quality" of the projection into

Table 3. Counterpropagation Artificial Neural Networks: Modeling and Predictive ability (LOO-CV) of the Best CP-ANNs ($\eta^{start} = 0.5$, $\eta^{end} = 0.01$)

		number of neurons ($N_x \times N_y$)				
		10 × 10	11 × 11	12 × 12	13 × 13	15 × 15
100 epochs	R^2	0.85	0.85	0.86		0.86
	Q^2	0.55	0.56	0.59		0.59
250 epochs	R^2			0.88		
	Q^2			0.55		
300 epochs	R^2	0.79	0.83	**0.88**	0.92	0.94
	Q^2	0.55	0.56	**0.62**	0.58	0.58
500 epochs	R^2			0.90		0.94
	Q^2			0.56		0.58
1000 epochs	R^2			0.88		
	Q^2			0.59		

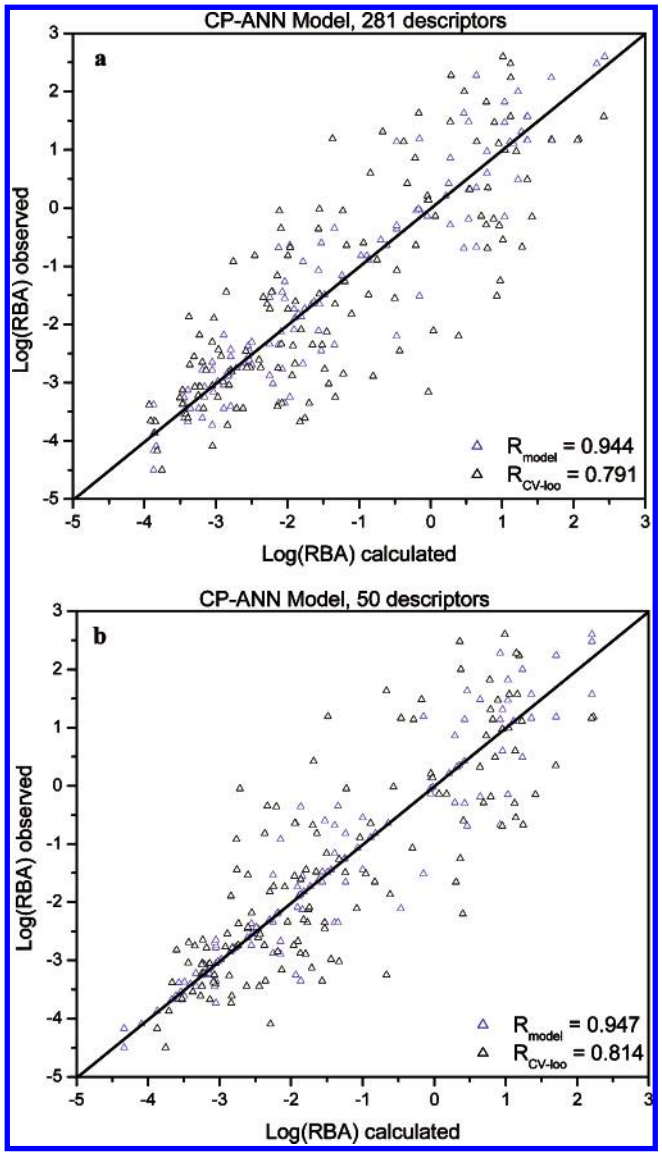

Figure 3. CP-ANN modeling. Correlation of calculated versus observed values of the Log(RBA) for the CP-ANN model based on the basis of a complete set of variables (a) and a reduced set of 50 variables (b). Predicted and cross-validates values are plotted in blue and black colors, respectively.

the Kohonen layer. In Figure 4 the top-map for the best counterpropagation network (a) and the corresponding output layer (b) is reported. The compounds have been marked by their ID numbers (see Table 1 and the Supporting Information). Not all neurons are occupied, 70 out of 144 neurons

203, 332 346, 601 640, 642	30, 31 62 175		57 171 442	666	18 641	44	71	50 55		108 172	109 110 174
		420 675	637			54 56			614		107
639, 661 662 663				481		615	23 165	612			624
	660	26 269 411			177 421				527		705
123		136 155			3		81 173	77	627 631		638
111	550	4			43, 159 160 617		106 514	105	634		6 92 613
113 115 495	267 708	2	118 664	5							73
121	116					635		654 656	531 532 628		
15 122 632	117		138 139		137	999	626				97 98
124						636		85 625	629 630		82
24			166 170					83 96			
87 95			167 633		12 119 667	132 668		84, 91 93, 646		94	86 114 647

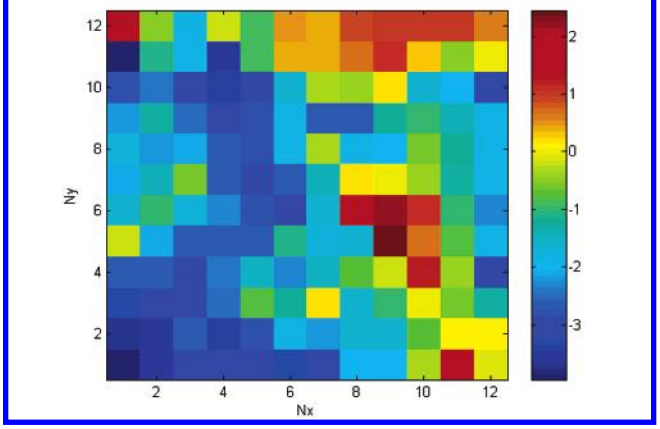

Figure 4. Top-map of the best CP-ANN network (upper plot) and the corresponding response surface of the output layer (lower plot).

are empty, while half of the remaining are occupied by more than one sample, giving rise to clusters of compounds which share a similarity in their chemical structure or, more precisely, in structural descriptors representing the chemical structure of each compound. For example, the compounds mapped onto the first two neurons in Figure 4a are all short chain alkylphenol derivatives. On the neuron in the top-left corner (position: 1, 1) there are 4-methylphenol, 3-ethylphenol, 4-ethylphenol, 4-chloro-2-methylphenol, 2-chloro-4-methylphenol, and 4-chloro-3-methylphenol, while on the neighboring neuron (position 2, 1) there are 2-*sec*-butylphenol, 4-*sec*-butylphenol, 4-*tert*-butylphenol, and 4-*tert*-amylphenol. The same consideration can be extended to the whole network.

On the other hand, the inspection of the weight maps, i.e. the distribution of the values of each input and of the output along the Kohonen layer, it is possible to inspect the correlation between each descriptor and the dependent variable. A careful inspection of all the weights has confirmed the considerations reported above on the PLS section. In Figure 5 we can compare the distribution of weights in the output layer (response surface, Figure 5f) with the weight maps shown in Figure 5a−e, which correspond to the variables var_001 (LogP), var_007 (the number of O

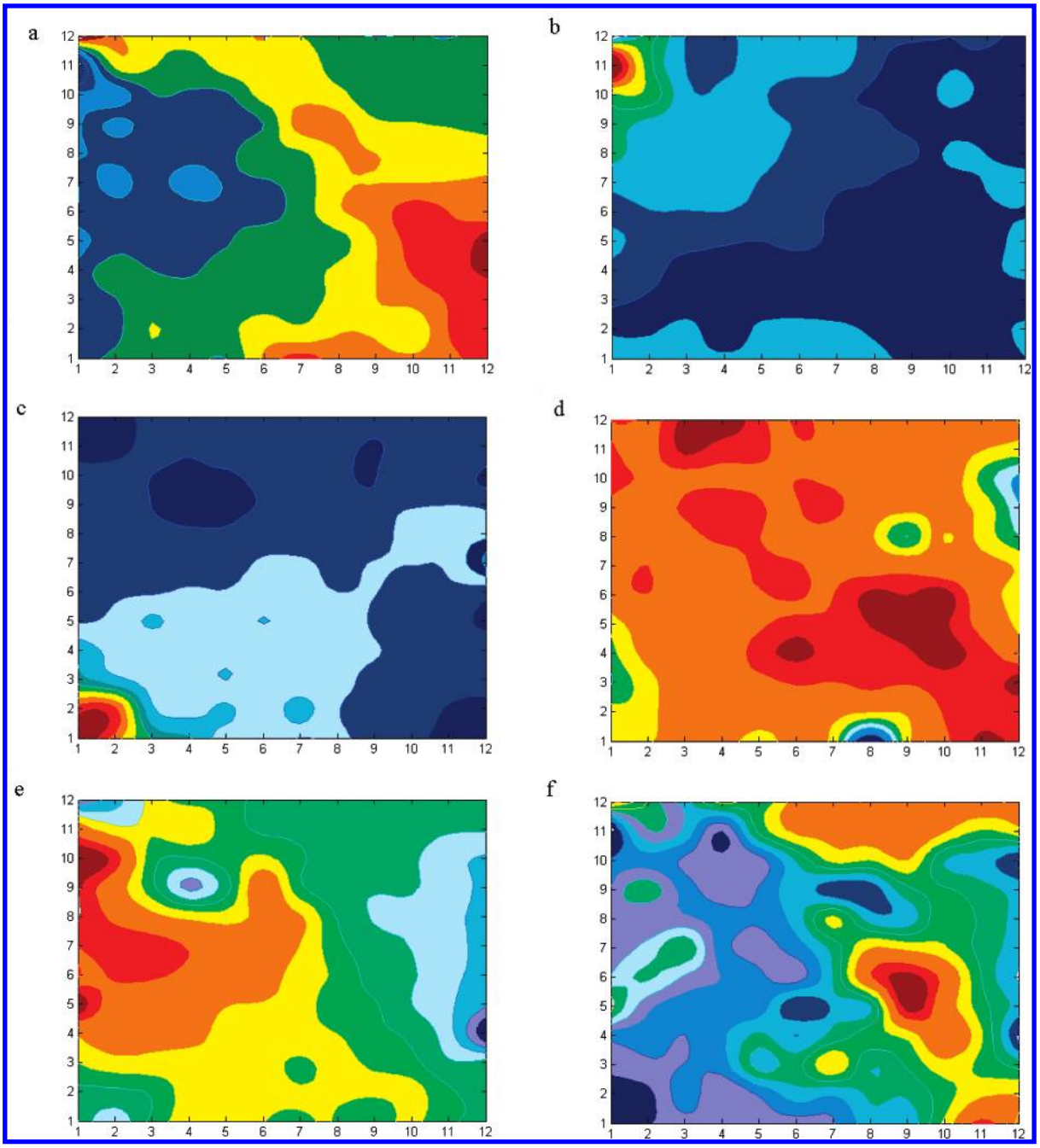

Figure 5. Weight maps of CP-ANN model corresponding to five arbitrarily chosen variables and the output layer: LogP (a), number of O atoms (b), moment of inertia along the first principal axis A (c), energy of the HOMO-1 (d), maximum total interaction for a C−C bond (e), and the output layer with the response surface of LogRBA (f).

atoms), var_271 (moment of inertia along the first principal axis A), var_145 (the energy of the HOMO-1), and var_260 (maximum total interaction for a C−C bond). Most of the correlations observed in the interpretation of the PLS regression coefficients are "conserved" also in the counterpropagation model. In particular, the high positive correlation between the LogP value and the relative binding affinity (the distribution of the weight values along the maps (plot (a)) and (plot (f)) show very similar trends) and the corresponding negative correlation of the dependent variable with the number of oxygen atoms (plot (b)) seem to confirm the hypothesis for the mechanism of action obtained from the PLS-R model. Also the correlation between HOMO energy terms (plot (d)) and relative binding affinity (plot (f)) is obvious, and the contours around high values on both plots coincide, which shows that the ligand with high RBA values have high HOMO energy.

CP-ANN Model − Variable Selection and Interpretation. Genetic algorithms have been employed to reduce the number of variables to be included in the counterpropagation model. Three independent genetic algorithms runs from different random origins have been performed, considering a population of 50 chromosomes evolved for 300 generations and looking for the combination of variables which led to the better predictive ability, as evaluated by the test set defined in the training/test splitting procedure.[24] The samples to be included in each set have been chosen according to a Kohonen-based intelligent selection, by mapping the 132 samples onto a 12 × 12 self-organizing map and selecting at least one sample from each of the occupied units, having

Table 4. CP-ANN Modeling and Predictive Ability of the Best Models Obtained from Three Runs of GA Applied for the Selection of Variables

variable set	no. of descriptors	R^{training} [a]	R^{test} [b]	RMS^{test} [c]
GA1	50	0.938	0.800	0.494
GA2	44	0.949	0.812	0.525
GA3	31	0.938	0.781	0.529

[a] Correlation coefficient between observed and predicted Log(RBA) values of 79 compounds from the training set. [b] Correlation coefficient between observed and predicted Log(RBA) values of 53 compounds from the test set. [c] Predictive error of the test compounds.

care that the whole experimental range was spanned. Additionally, samples with a high leave-one-out CV prediction error have been added to the training set in order to include the information of their unique structure−property relationship into the model, so that they could be modeled as good as possible. The best model of each of the three independent GA runs is reported in Table 4 together with their modeling and predictive ability for the training and test sets of 79 and 53 compounds, respectively. In Figure 3b the calculated and predicted (leave-one-out cross-validation) versus observed Log(RBA) values are shown.

A close inspection of the variables involved in the best models shows that most of the relevant PLS variables are "conserved" in the GA-selected set of descriptors. This could result in an additional confirmation of the proposed interpretation of our results.

Furthermore, an additional variable-selection experiment was done, based on the projection of the variables[25] from the 132-dimensional sample space to a 5 × 5 two-dimensional Kohonen network to select on this basis a 50-descriptors data set (the nearest and the farthest variables from the centroid of each unit has been included in the set), also resulted in most of the variables already significant for the PLS discrimination to be included in the optimal classifier. Anyway, the modeling and predicting ability of the resulting counterpropagation network appeared to be slightly lower ($R^2 = 0.85$ and $Q^2 = 0.55$) than those of the corresponding 281-descriptors model or reduced representation models obtained with GA variable selection.

BP-ANN Model. To make a comparison between the PLS-R and different nonlinear ANN models, the multilayer feedforward artificial neural networks trained by the back-propagation of errors algorithm have also been used to build a suitable predictive model. As for the other multivariate methods, in a first stage a model containing all the descriptors has been built, having care to choose the dimensionality of the problem (particularly the number of hidden neurons) in a way not to risk overfitting. In this respect, the final number of adjustable parameters (i.e. the connection weights) has been kept significantly lower than the number of independent data available. Different network architectures have been tested, varying the number of hidden neurons (15−20), training epochs (200−1645), and the values of the parameters η (0.10−0.15) and μ (0.15−0.90), to find the best regression model. The optimal BP-ANN model constructed with the complete descriptor set resulted to be a 281−18−1 architecture (5095 connection weights including bias nodes) trained for 225 epochs with learning rate $\eta = 0.10$ and momentum $\mu = 0.15$. Further reduction of variables is necessary for the BP-ANN models in order to avoid over determination of the model due to a large number of adjustable parameters. As given above, there are 5095 adjustable parameters in our BP-ANN model, in comparison with the number of experimental data this is not too high (132 compounds represented by 281 variables each, i.e., 37092 independent values), but for a robust model it is better to have this ratio even larger, thus a variable reduction procedure was performed in the next step.

BP-ANN Model − Variable Selection and Interpretation. A genetic algorithms-based variable selection technique was again applied to reduce the number of the descriptors to be included in the regression model in order to improve its predictive ability. Ten independent genetic algorithm runs have been performed, considering a 50-chromosomes population, which was evolving for 300 generations, using the predictive ability over the test set as the fitness criterion. The test set was chosen on the basis of self-organizing maps from the Kohonen neural network,[24] the so-called Kohonen-based intelligent selection, the same as reported in the case of the CP-ANN model. The best 15 chromosomes of each independent GA run have then been considered, and their predictive ability has been finally evaluated using the leave-one-out cross-validation. The percentage of occurrences of each descriptor in the best 150 models (15 best chromosomes of each of the 10 independent runs), weighted by the fitness value scored by each of the models were calculated and are available for all 281 descriptors in the Supporting Information. The most frequently chosen descriptors were variables number 1, 117, 136, 76, 162, 228, 177, 115, 139, and 40, to list only those with more that 50% of occurrence in average. In Table 5 the selected set of descriptors of the optimal model is displayed, including the top 10 descriptors listed above. The best overall regression model was found to be a 49−10−1 network (511 weighted connections including bias nodes) trained for 2500 epochs with learning rate $\eta = 0.15$ and momentum $\mu = 0.15$. This optimal BP-ANN model performed better that the PLS-R or CP-ANN models, with the $R^2 = 0.92$ and $Q^2 = 0.71$. This result is demonstrated in Figure 6, where the observed vs CV-predicted value plot for the best GA-selected model is reported.

The inspection of the 49 selected descriptors used in the optimal BP-ANN model shows some overlapping features, if confronted with the best models resulting from the use of the other modeling techniques considered in this work. In particular, we can observe the selection of the logarithm of the partition coefficient LogP as a relevant variable present in all the reduced sets of descriptors used for modeling the binding affinity of the examined samples. Moreover, the selection of the XY and YZ shadows and of the moment of inertia A confirms a possible influence of the shape and flexibility of the molecule on the binding affinity, while the inclusion of the max and min atomic charge and of other terms accounting for the electronic distribution of the inspected molecules is indicative of a correlation between the polarity of the molecules and the value of the dependent variable.

LogP is the most frequently selected descriptor, being present on average in more than 80% of the best models. Moreover, terms accounting for each of the contributions to the relative binding affinities suggested by previous models (PLS-R and CP-ANN) are present among these 10 variables: charge, polarity, and van der Waals effects (var_117,

Table 5. 49 Variables Selected in the Best GA-BP-ANN Model

Var_no	descriptor	Var_no	descriptor
1	LogP	133	HASA-2 [Zefirov's PC]
3	number of C atoms	136	HACA-1 [Zefirov's PC]
4	relative number of C atoms	139	HACA-2/TMSA [Zefirov's PC]
5	number of H atoms	153	min electroph. react. index for a C atom
19	relative number of single bonds	160	min net atomic charge for a C atom
21	relative number of double bonds	161	max net atomic charge
35	Randic index (order 0)	162	min net atomic charge
36	Randic index (order 1)	166	image of the Onsager-Kirkwood solvation energy
37	Randic index (order 2)	168	PPSA-1 partial positive surface area [semi-MO PC]
38	Randic index (order 3)	175	PPSA-2 total charge weighted PPSA [semi-MO PC]
39	Kier&Hall index (order 0)	177	DPSA-2 difference in CPSAs (PPSA2-PNSA2) [semi-MO PC]
40	Kier&Hall index (order 1)	179	FNSA-2 fractional PNSA (PNSA-2/TMSA) [semi-MO PC]
44	Kier shape index (order 2)	188	WNSA-3 weighted PNSA (PNSA3*TMSA/1000) [semi-MO PC]
45	Kier shape index (order 3)	189	RPCG relative positive charge (QMPOS/QTPLUS) [semi-MO PC]
52	complementary information content (order 0)	205	min(#HA, #HD) [semi-MO PC]
72	moment of inertia A	208	HA dependent HDSA-1 [semi-MO PC]
76	XY shadow/XY rectangle	212	HA dependent HDSA-2/SQRT(TMSA) [semi-MO PC]
77	YZ shadow	219	HASA-1/TMSA [semi-MO PC]
100	PPSA-2 total charge weighted PPSA [Zefirov's PC]	227	HACA-2/SQRT(TMSA) [semi-MO PC]
106	WNSA-2 weighted PNSA (PNSA2*TMSA/1000) [Zefirov's PC]	228	min atomic orbital electronic population
		232	max PI−PI bond order
112	WPSA-3 weighted PPSA (PPSA3*TMSA/1000) [Zefirov's PC]	235	min valency of a C atom
		239	max bond order of a C atom
115	RPCS relative positive charged SA (SAMPOS*RPCG) [Zefirov's PC]	244	max e−n attraction for a C atom
		247	min resonance energy for a C−C bond
117	RNCS Relative negative charged SA (SAMNEG*RNCG) [Zefirov's PC]	257	min Coulombic interaction for a C−C bond

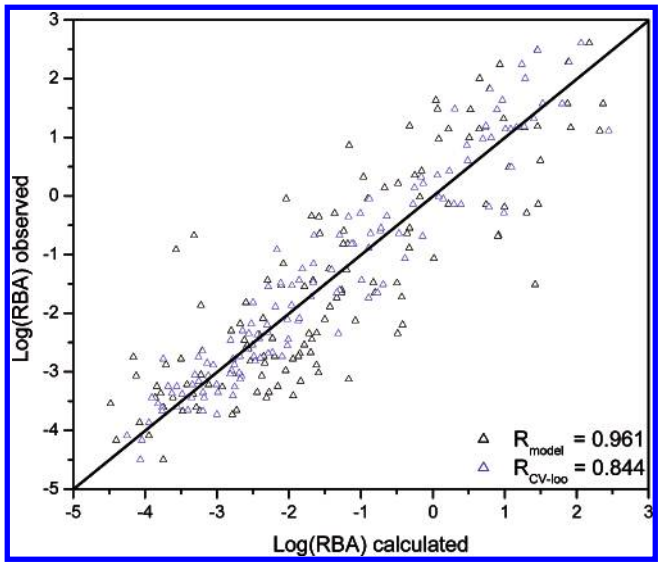

Figure 6. Performance of the best BP-ANN (back-propagation artificial neural network) model based on a reduced set of 49 descriptors. Predictions obtained for 132 compounds are tested (a) and leave-one-out cross-validated predictions (b) are shown.

var_136, var_162, var_177, var_115, var_139), shape (var_76), and orbital (var_228).

DISCUSSION

It has been shown that different modeling methods (PLS-R, CP-ANN, BP-ANN) were successfully employed to model RBA with different degrees of prediction ability. It is important that the model validation procedure is strict and comparable for all investigated modeling methods. The most objective and independent of the chosen modeling method of relatively small data sets is the complete cross-validation or leave-one-out method;[26] it has been shown by Hua Gao et al.[27] that it may be also applied for larger data sets, however, not in the procedure of variable selection, but for final comparison of different models. This was also our modeling strategy presented here. The leave-one-out cross-validation has been applied to assess the predictive ability of all models obtained by the three modeling methods investigated, while in the variable selection process using the genetic algorithm, the division of the data set into the training and the test sets was performed; the prediction of the test compounds was considered for the fitness function on the basis of which the genetic algorithm was able to optimize the selection of a reduced set of descriptors (see paragraphs for variable selection in CP-ANN and BP-ANN modeling).

It is obvious that for different modeling methods, different variable selection/reduction procedures are appropriate. In the case of QSAR models based on a large number of descriptors used as molecular structure representation vectors it is not feasible to expect a unique optimal set of selected descriptors for all modeling methods. This is specially true due to the fact that the descriptors are actually grouped into several clusters, each of them being descriptive for certain structural features or particularities, which are reflected in a whole group of descriptors. Having obtained the optimal sets of descriptors for each of the three applied modeling methods, we are enabled to compare them and find clusters of descriptors overlapping in the three optimal sets.

The best modeling results were obtained with the BP-ANN model; however, it was not the only aim of the presented research to obtain the best model. A comparison of which descriptors have been chosen for the best models in all three different modeling methods accounts for the essence of information about structure−property relationship. We mapped the three sets of variables, selected as individual optimal descriptors for the three confronted methods, onto the maps from Kohonen neural network (Koh-map) and compared them with the map of original, nonreduced set of 218 descriptors. The resulting 2-D distributions of descriptors are

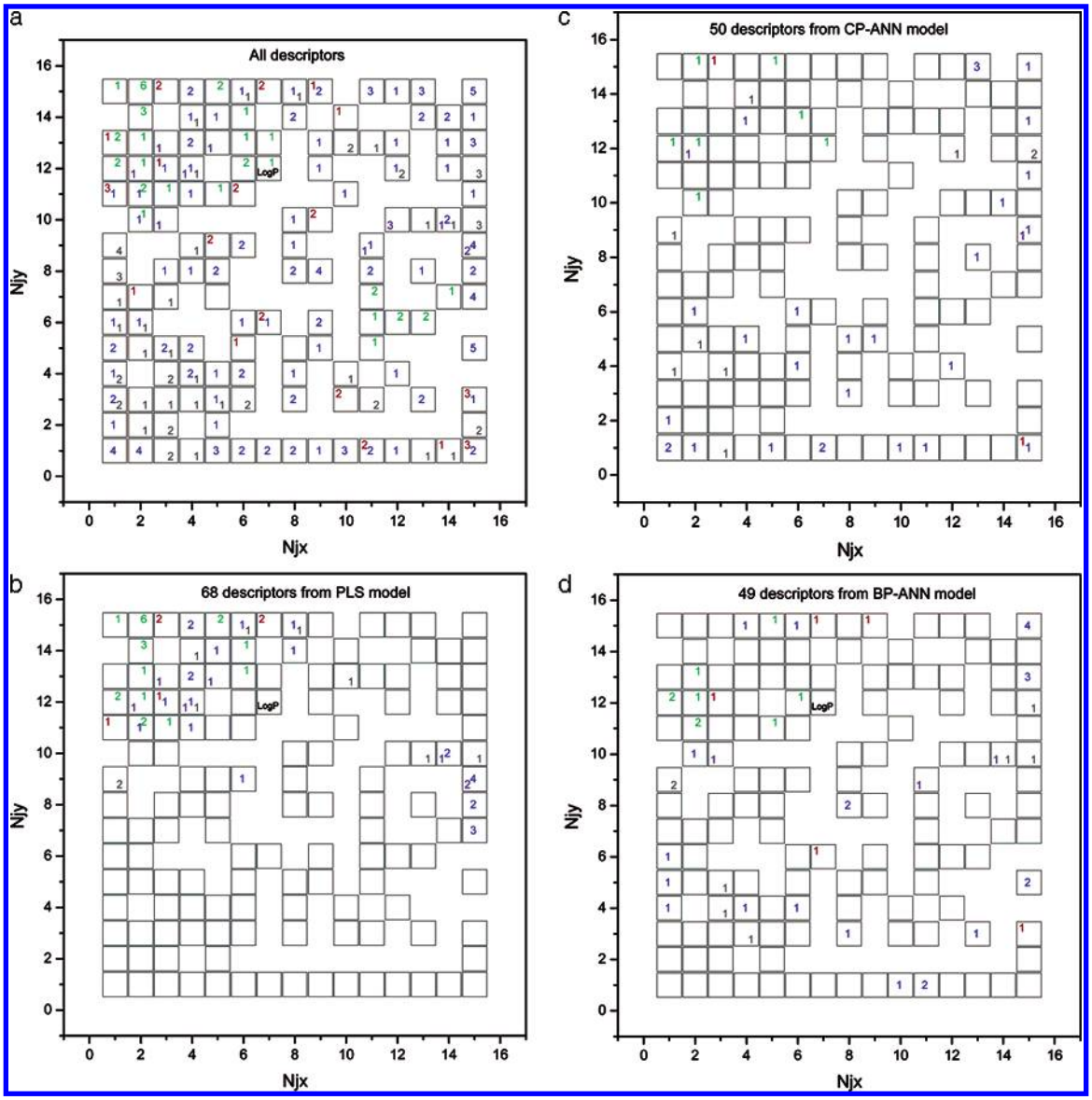

Figure 7. Top-map of the self-organizing Kohonen neural network with 15 × 15 neurons (Koh-map) of all 281 descriptors (a), of the descriptors selected in the PLS-R model (b), the CP-ANN model (c), and the BP-ANN model (d). Only neurons occupied by at least one descriptor in (a) are shown. The number of descriptors from 5 classes are indicated by different colors and positions within each square: constitutional (red, up-left position); topological (green, up-right position); geometrical (purple, low-left position); electrostatic (gray, low-right position); quantum chemical descriptors (cyan, center). Descriptors shown in (b), (c), and (d) are projected into the same Koh-map as shown in (a).

shown in Figure 7a−d. As an example, let us examine the selection of LogP as one of the descriptors for the reduced set in different modeling methods. The LogP (Var_001) was present in the reduced sets obtained with the PLS-R and BP-ANN modeling methods and not with the CP-ANN method. In Figure 7a, in which the Koh-map of nonreduced set of descriptors is shown, we can find the Var_001 (LogP) at the position $Nj_x = 7$, $Nj_y = 12$; the same situation can be observed in the case of reduced sets obtained in PLS-R (Figure 7b) and BP-ANN (Figure 7c) models. With the inspection of the Koh-map of the reduced set associated with the CP-ANN modeling method (see Figure 7d) it can be observed that this position is occupied by the Var_067 (complementary information content − order 2). Comparing with Figure 7a, in which the complete set of descriptors is shown in the Koh-map, we can see that the descriptors Var_001 and Var_067 hit the same neuron. If two descriptors occupy the same neuron, i.e., position in the Koh-map, they must be similar enough to be indistinguishable in the mapping procedure. Of course the conclusion of the similarity of two inspected descriptors is only valid for the investigated set of molecules and providing that the descriptors are normalized as described in the "Data" section.

CONCLUSIONS

The correlation between chemical structures and estrogen receptor binding affinities of 132 compounds have been studied during different attempts at constructing a model, which would be able to predict the estrogen receptor binding affinity of an unknown compound on the basis of its chemical structure. Among the three different modeling methods tested, (i) linear model based on partial least-squares regression (PLS-R model), (ii) counterpropagation neural network model (CP-ANN model), and (iii) multilayer feedforward

artificial neural networks trained by the back-propagation of errors algorithm (BP-ANN model), the last one has shown the best predictive ability. The results were compared with similar studies from the literature[2−5,28−34] and taking into account that it is not possible to have an absolute measure of comparison because of different sets of compounds, our modeling results were comparable with the most successful models reported (BP ANN with reduced set of selected descriptors, $R^2 = 0.92$ and $Q^2 = 0.71$).

Reduction of variables employed in each modeling procedure enabled us to select the descriptors, which were the most relevant in the studied structure−property relationship and therefore provided the best models. Interestingly, the logarithm of the partition coefficient LogP resulted as a relevant variable present in all the reduced sets of descriptors in the two modeling methods, PLS-R and BP-ANN, while in the CP-ANN modeling method LogP was not selected, However, another descriptor ("complementary information content − order 2") was present, which, as far as the studied compounds are concerned, has been shown to provide the same kind of information because it was located at the same position as LogP in the Kohonen map of descriptors.

The influence of shape and flexibility of the molecule on the binding affinity is hypothesized on the basis of the descriptors such as XY and YZ shadows or the moment of inertia chosen in the variable reduction procedure. The selection of the descriptors accounting for the electronic distribution of the inspected molecules (such as max and min atomic charge) indicates a correlation between the polarity of the molecules and the modeled property.

ACKNOWLEDGMENT

The authors acknowledge partial financial support of the Ministry of Higher Education and Science of Slovenia for financing the research through the project grants P104-507 and P1-017 as well as the European Union 5th framework program scheme for partial financial support through the program Marie Curie Host Training Site no. HPMT-CT-2001-00240 and the EASYRING project (contract no.: QLK4-CT-2002-02286). We thank Prof. A. Katritzky (University of Florida) and Prof. M. Karelson (University of Tartu) for the use of CODESSA.

Supporting Information Available: Chemicals in the data set (names and CAS numbers) as well as the names of descriptors together with their numbers used in this study (Var_001-Var_281). This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) ICPS-WHO. *International Programme on Chemical Safety, Global Assessment of the State-of-the-Science of Endocrine Disruptors*; Damstra, T., Barlow, S., Bergman, A., Kavlock, R., Van Der Kraak, G., Eds.; 2000.

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

(3) Schultz, T. W.; Sinks, G. D.; Cronin, M. T. D. Structure−activity relationships for gene activation oestrogenicity: Evaluation of a diverse set of aromatic chemicals. *Environ. Toxicol.* **2002**, *17*, 14−23.

(4) de Voogt, P.; van Hattum, B. Critical factors in exposure modelling of endocrine active substances. *Pure Appl. Chem*. **2003**, *75*, 1933−1948.

(5) Tong, W. D.; Fang, H.; Hong, H. X.; Xie, Q.; Perkins, R.; Anson, J.; Sheehan, D. M. Regulatory application of SAR/QSAR for priority setting of endocrine disruptors: A perspective. *Pure Appl. Chem.* **2003**, *75*, 2375−2388.

(6) Fukushima, S.; Freyberger, A. Simple, rapid assays for conventional definite testing of endocrine disruptor hazard: Summary and recommendations. *Pure Appl. Chem*. **2003**, *75*, 2479−2482.

(7) Blair, R. M.; Fang, H.; Branham, W. S.; Hass, B. S.; Dial, S. L.; Moland, C. L.; Tong, W.; Shi, L.; Perkins, R.; Sheehan, D. M. The Estrogen Receptor Relative Binding Affinities of 188 Natural and Xenochemicals: Structural Diversity of Ligands. *Toxicol. Sci.* **2000**, *54*, 138−153.

(8) Martens, H.; Martens, M. *Multivariate Analysis of Quality. An Introduction*; J. Wiley & Sons, Ltd.: 2000.

(9) Geladi, P.; Kowalski, B. R. Partial least-squares regression: a tutorial. *Anal. Chim. Acta* **1986**, *185*, 1−17.

(10) Eriksson, L.; Johansson, E.; Kettaneh-Wold, N.; Wold, S. *Introduction to Multi and Megavariate Data Analysis using Projection Methods (PCA & PLS)*; Umetrics AB: Umeå, Sweden, 1999.

(11) Dayhof, J. *Neural Network Architectures, An Introduction*; Van Nostrand Reinhold: New York, 1990; p 192.

(12) Zupan, J.; Novič, M.; Ruisanchez, I.; Kohonen and counter-propagation artificial neural networks in analytical chemistry. *Chemom. Intell. Lab. Syst.* **1997**, *38*, 1−23.

(13) Rumelhart, D. E.; Hinton, G. E.; Williams, R. J. Learning internal representations by error propagation. In *Microstructures of Cognition, 1*; Rumalhart, D. E., McClelland, J. L., Eds.; MIT Press: Cambridge, 1986; pp 318−362.

(14) Todeschini, R. Consonni, V. *The Handbook of Molecular Descriptors, Series of Methods and Principles in Medicinal Chemistry − Vol. 11*; Mannhold, R., Kubinyi, H., Timmerman, G., Eds.; Wiley-VCH: New York, 2000.

(15) Katritzky, A. R.; Lobanov, V. S.; Karelson, M. Copyright 1994−1996, *CODESSA 2.0, Comprehensive Descriptors for Structural and Statistical Analysis*; University of Florida, U.S.A.

(16) Dragon, Copyright 1997−2004 *TALETE srl − Milano, Italy*; Developed by Todeschini, R.; Consonni, V.; Mauri, A.; Pavan, M. Milano Chemometrics and QSAR Research Group, Department of Environmental Sciences − University of Milano.

(17) Sadowski, J.; Gasteiger, J.; Klebe, G. Comparison of Automatic 3-Dimensional Model Builders Using 639 X-Ray Structures. *J. Chem. Inf. Comput. Sci*. **1994**, *34*, 1000−1008.

(18) Davis, L. *Handbook of Genetic Algorithms*; New York, 1991.

(19) Hibbert, B. Genetic Algorithm in Chemistry. *Chemom. Intell. Lab. Syst.* **1993**, *19*, 277−293.

(20) Zupan, J.; Novič, M. Optimisation of structure representation for QSAR studies. *Anal. Chim. Acta* **1999**, *388*, 243−250.

(21) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. The Development and Use of Quantum-Mechanical Molecular-Models. 76. Am1 − A New General-Purpose Quantum-Mechanical Molecular-Model. *J. Am. Chem. Soc*. **1985**, *107*, 3902−3909.

(22) Hansch, L.; Leo, A. *Exploring QSAR Fundamentals and Applications in Chemistry and Biology*; American Chemical Society: Washington, DC, 1995.

(23) Osborne, C. Statistical Calibration − A Review. *Int. Stat. Rev.* **1991**, *59*, 309−336.

(24) Novič, M.; Zupan, J. Investigation of Infrared Spectra-Structure Correlation Using Kohonen and Counter-Propagation Neural Network. *J. Chem. Inf. Comput. Sci*. **1995**, *35*, 454−466.

(25) Roncaglioni, A.; Novič, M.; Vračko, M.; Benfenati, E. Classification of potential endocrine disrupters on the basis of molecular structure using a nonlinear modelling method. *J. Chem. Inf. Comput. Sci*. **2004**, *44*, 300−309.

(26) Martens, H. A.; Dardenne, P. Validation and verification of regression of small data sets. *Chemom. Intell. Lab. Syst.* **1998**, *44*, 99−121.

(27) Gao, H.; Lajiness, M. S.; Van Drie, J. Enhancement of binary QSAR analysis by a GA-based variable selection method. *J. Mol. Graph. Model.* **2002**, *20*, 259−268.

(28) Shi, L. M.; Fang, H.; Tong, W.; Wu, J.; Perkins, R.; Blair, R. M.; Branham, W. S.; Dial, S. L.; Moland, C. L.; Sheehan, D. M. QSAR Models Using a Large Diverse Set of Estrogens. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 186−195.

(29) Asikainen, A.; Ruuskanen, J.; Tuppurainen, K. 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*. **2004**, *38*, 6724−6729.

(30) Ghafourian, T.; Cronin, M. T. D. The impact of variable selection on the modelling of oestrogenicity. *SAR QSAR Environ. Res*. **2005**, *16*, 171−190.

(31) Klopman, G.; Chakravarti, K. S. Structure−activity relationship study of a diverse set of estrogen receptor ligands (I) using MultiCASE expert system. *Chemosphere* **2003**, *51*, 445−459.

(32) Waller, C. L. A comparative QSAR study using CoMFA, HQSAR, and FRED/SKEYS paradigms for estrogen receptor binding affinities

of structurally diverse compounds. *J. Chem. Inf. Comput. Sci*. **2004**, *44*, 758−765.

(33) Wolohan, P.; Reichert, D. E. CoMFA and docking study of novel estrogen receptor subtype selective ligands. *J. Comput.-Aided. Mol. Des*. **2003**, *17*, 313−328.

(34) Pasha, F. A.; Srivastava, H. K.; Singh, P. P. QSAR study of estrogens with the help of PM3-based descriptors. *Int. J. Quantum Chem*. **2005**, *104*, 87−100.

CI0501645