Available online at www.sciencedirect.com
ScienceDirect
Journal of Hazardous Materials 151 (2008) 603–609
Journal of Hazardous Materials
www.elsevier.com/locate/jhazmat

# Prediction of toxicity of nitrobenzenes using ab initio and least squares support vector machines

Ali Niazi *, Saeed Jameh-Bozorghi, Davood Nori-Shargh

*Department of Chemistry, Faculty of Sciences, Azad University of Arak, Arak, Iran*

Received 8 January 2007; received in revised form 8 May 2007; accepted 10 June 2007
Available online 14 June 2007

**Abstract**

A quantitative structure–property relationship (QSPR) study is suggested for the prediction of toxicity ($IGC_{50}$) of nitrobenzenes. Ab initio theory was used to calculate some quantum chemical descriptors including electrostatic potentials and local charges at each atom, HOMO and LUMO energies, etc. Modeling of the $IGC_{50}$ of nitrobenzenes as a function of molecular structures was established by means of the least squares support vector machines (LS-SVM). This model was applied for the prediction of the toxicity ($IGC_{50}$) of nitrobenzenes, which were not in the modeling procedure. The resulted model showed high prediction ability with root mean square error of prediction of 0.0049 for LS-SVM. Results have shown that the introduction of LS-SVM for quantum chemical descriptors drastically enhances the ability of prediction in QSAR studies superior to multiple linear regression and partial least squares.
© 2007 Elsevier B.V. All rights reserved.

*Keywords:* Nitrobenzene; Toxicity; Ab initio; MLR; PLS; GA-PLS; LS-SVM

## 1. Introduction

Nitroaromatics are hazardous chemicals that display several manifestations of toxicity, including skin sensitization, immunotoxicity, germ cell degeneration, inhibition of liver enzymes and also a conjectured carcinogenicity [1]. Nitrobenzene toxicity to the aquatic ciliate *Tetrahymena pyriformis* has been extensively studied by several groups of workers [1–5] with the use of different quantitative structure–activity relationship (QSAR) methodologies. QSAR [6,7] as an important area of chemometrics has been the subject of a series of investigations. The main aim of QSAR studies is to establish an empirical rule or function relating the structural descriptors of compounds under investigation to bioactivities. This rule of function is then utilized to predict the same bioactivities of the compounds not involved in the training set from their structural descriptors. Whether the bioactivities can be predicted with satisfactory accuracy depends to a great extent on the performance of the applied multivariate data analysis method, provided the property being predicted is related to the descriptors.

Among the investigation of QSAR, one of the most important factors affecting the quality of the model is the method to build the model. Many multivariate data analysis methods such as multiple linear regression (MLR) [8–10], partial least squares (PLS) [7] and artificial neural network (ANN) [11] have been used in QSAR studies. MLR, as most commonly used chemometrics method, has been extensively applied to QSAR investigations. However, the practical usefulness of MLR in QSAR studies is rather limited, as it provides relatively poor accuracy. ANN offers satisfactory accuracy in most cases but tends to overfit the training data. The support vector machine (SVM) is a popular algorithm developed from the machine learning community. Due to its advantages and remarkable generalization performance over other methods, SVM has attracted attention and gained extensive applications [12,13]. As a simplification of traditional of SVM, Suykens and Vandewalle [14,15] have proposed the use of least-squares SVM (LS-SVM). LS-SVM encompasses similar advantages as SVM, but its additional advantage is that it requires solving a set of only linear equations (linear programming), which is much easier and computationally more simple.

A major step in constructing QSAR models is finding one or more molecular descriptors that represent variation in the structural property of the molecules by a number. A wide variety

* Corresponding author. Tel.: +98 8613663041; fax: +98 8613670017.
*E-mail address:* ali.niazi@gmail.com (A. Niazi).

0304-3894/$ – see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.jhazmat.2007.06.030

of descriptors have been reported to be used in QSAR analysis [7,9,11,16–18]. Recent progress in computational hardware and the development of efficient algorithms have assisted the routine development of molecular quantum chemical calculations. Quantum chemical calculations are thus an attractive source of new molecular descriptors, which can, in principle, highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital (LUMO) energies, molecular polarizability, dipole moments, and energies of molecule are examples of quantum chemical descriptors used in QSAR studies.

In the present paper, the MLR, PLS, GA-PLS and LS-SVM methods were applied in QSAR for modeling the relationship between the toxicity of 39 nitrobenzenes. Ab initio geometry optimization was performed at the B3LYP level, with a known basis set, 6–31$^{++}$G**. Local charges, electrostatic potential, dipole moment, polarizability, HOMO–LUMO energies, hardness, softness, electronegativity and electrophilicity were calculated for each compound.

## 2. Theory

Theory of LS-SVM has also been described clearly by Suykens et al. [14,15] and application of LS-SVM in quantification [19–21], classification [22,23] and QSAR [24,25] reported by some of the workers. So, we will only briefly describe the theory of LS-SVM. The LS-SVM [15] is capable of dealing with linear and nonlinear multivariate calibration and resolves multivariate calibration problems in a relatively fast way. In LS-SVM a linear estimation is done in kernel-induced feature space ($y = w^{\mathrm{T}}\phi(x) + b$). As in SVM, it is necessary to minimize a cost function ($C$) containing a penalized regression error, as follows:

$$C = \frac{1}{2}w^{\mathrm{T}}w + \frac{1}{2}\gamma\sum_{i=1}^{N}e_i^2 \tag{1}$$

such that:

$$y_i = w^{\mathrm{T}}\phi(x_i) + b + e_i \tag{2}$$

for all $i = 1, \ldots, N$, where $\phi$ denotes the feature map.

The first part of this cost function is a weight decay which is used to regularize weight sizes and penalize large weights. Due to this regularization, the weights converge to similar value. Large weights deteriorate the generalization ability of the LS-SVM because they can cause excessive variance. The second part of Eq. (1) is the regression error for all training data. The parameter $\gamma$, which has to be optimized by the user, gives the relative weight of this part as compared to the first part. The restriction supplied by Eq. (2) gives the definition of the regression error. Analyzing Eq. (1) and its restriction given by Eq. (2), it is possible to conclude that we have a typical problem of convex optimization [15] which can be solved by using the Lagrange multipliers method [26], as follows:

$$L = \frac{1}{2}\|w\|^2 + \gamma\sum_{i=1}^{N}e_i^2 - \sum_{i=1}^{N}\alpha\{w^{\mathrm{T}}\phi(x_i) + b + e_i - y_i\} \tag{3}$$

where

$$y_i = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}, \qquad e_i = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_N \end{bmatrix} \quad \text{and} \quad \alpha_i = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_N \end{bmatrix}$$

To obtain the optimum solution, one sets all corresponding partial first derivatives to zero; the weights obtained are linear combinations of the training data:

$$\frac{\partial L(w, b, e, \alpha)}{\partial w} = w - \sum_{i=1}^{N}\alpha_i\phi(x_i) = 0, \quad \therefore w = \sum_{i=1}^{N}\alpha_i\phi(x_i) \tag{4}$$

$$\frac{\partial L(w, b, e, \alpha)}{\partial e} = \sum_{i=1}^{N}\gamma e - \alpha = 0 \tag{5}$$

then:

$$w = \sum_{i=1}^{N}\alpha_i\phi(x_i) = \sum_{i=1}^{N}\gamma e_i\phi(x_i) \tag{6}$$

where a positive definite kernel is used as follows:

$$K(x_i, x_j) = \phi(x_i)^{\mathrm{T}}\phi(x_j) \tag{7}$$

An important result of this approach is that the weights ($w$) can be written as linear combinations of the Lagrange multipliers with the corresponding data training ($x_i$). Putting the result of Eq. (6) into the original regression line ($y = w^{\mathrm{T}}\phi(x) + b$), the following result is obtained:

$$y = \sum_{i=1}^{N}\alpha_i\phi(x_i)^{\mathrm{T}}\phi(x) + b = \sum_{i=1}^{N}\alpha_i\langle\phi(x_i)^{\mathrm{T}}, \phi(x)\rangle + b \tag{8}$$

for a point $y_i$ to be evaluated it is:

$$y_i = \sum_{i=1}^{N}\alpha_i\phi(x_i)^{\mathrm{T}}\phi(x_j) + b = \sum_{i=1}^{N}\alpha_i\langle\phi(x_i), \phi(x_j)\rangle + b \tag{9}$$

The $\alpha$ vector follows from solving a set of linear equations:

$$M\begin{bmatrix} \alpha \\ b \end{bmatrix} = \begin{bmatrix} y \\ 0 \end{bmatrix} \tag{10}$$

where $M$ is a square matrix given by:

$$M = \begin{bmatrix} K + \dfrac{I}{\gamma} & 1_N \\ 1_N^{\mathrm{T}} & 0 \end{bmatrix} \tag{11}$$

where $K$ denotes the kernel matrix with $ij$th element $K = (x_i, x_j) = \phi(x_i)^{\mathrm{T}}\phi(x_j)$ and $I$ denotes the identity matrix $N \times N$, $1_N = [1 \quad 1 \quad \ldots \quad 1]^{\mathrm{T}}$. Hence, the solution is given by:

$$\begin{bmatrix} \alpha \\ b \end{bmatrix} = M^{-1}\begin{bmatrix} y \\ 0 \end{bmatrix} \tag{12}$$

As can be seen from Eqs. (11) and (12), usually all Lagrange multipliers (the support vectors) are nonzero, which means that all training objects contribute to the solution. In contrast with standard SVM the LS-SVM solution is usually not sparse. However, as described by Suykens et al. [15] a sparse solution can be easily achieved via pruning or reduction techniques. Depending on the number of training data set either direct solvers can be used or an iterative solver such as conjugate gradients methods (for large data sets), in both cases with numerically reliable methods.

In applications involving nonlinear regression it is enough to change the inner product $\langle\phi(x_i),\phi(x_j)\rangle$ of Eq. (9) by a kernel function and the $ij$th element of matrix $K$ equals $K_{ij}=\phi(x_i)^{\mathrm{T}}\phi(x_j)$. If this kernel function meets Mercer's condition [27] the kernel implicitly determines both a nonlinear mapping, $x\rightarrow\phi(x)$ and the corresponding inner product $\phi(x_i)^{\mathrm{T}}\phi(x_j)$. This leads to the following nonlinear regression function:

$$y=\sum_{i=1}^{N}\alpha_i K(x_i,x)+b \tag{13}$$

for a point $x_j$ to be evaluated it is:

$$y_j=\sum_{i}^{N}\alpha_i K(x_i,x_j)+b \tag{14}$$

The attainment of the kernel function is cumbersome and it will depend on each case. However, the kernel function more used is the radial basis function (RBF), $\exp(-(||x_i-x_j||^2)/2\sigma^2)$, a simple Gaussian function, and polynomial functions $\langle x_i,x_j\rangle^d$, where $\sigma^2$ is the width of the Gaussian function and $d$ is the polynomial degree, which should be optimized by the user, to obtain the support vector. For $\alpha$ of the RBF kernel and $d$ of the polynomial kernel it should be stressed that it is very important to do a careful model selection of the tuning parameters, in combination with the regularization constant $\gamma$, in order to achieve a good generalization model.

## 3. Materials and computational methods

### 3.1. Hardware and software

The computations were made with an AMD 2000 XP (512 MB RAM) microcomputer with the Windows XP operating system and with Matlab (Version 6.5, Mathwork Inc.). The PLS evaluations were carried out by using the PLS program from PLS-Toolbox Version 2.0 for use with Matlab from Eigenvector Research Inc. The LS-SVM optimization and model results were obtained using the LS-SVM lab toolbox (Matlab/C Toolbox for Least-Squares Support Vector Machines) [28]. Hyperchem (Version 6.03, Hyperchem Inc.) and Gaussian 98 software [29] were used for geometric optimization of the molecules and calculation of the quantum chemical descriptor.

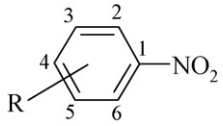


Fig. 1. The parent structure of nitrobenzenes.

### 3.2. Data set

The toxicities against the aquatic ciliate *T. pyriformis* were measured by Dearden et al. [3]. The parent structure of nitrobenzenes is shown in Fig. 1. The toxicity is expressed as log ($1/IGC_{50}$) where $1/IGC_{50}$ means 2-day (i.e. eight generations) static 50% inhibitory growth concentration. The *T. pyriformis* test is a 48 h (2-day) tests. The organism can reproduce itself in 6 h, so that in 48 h there are 48/6 = 8 generations. The structures of nitrobenzenes and their corresponding toxicities are

Table 1
The structures of nitrobenzenes studied in the present study and their toxicity [3]

| No. | Substituent, R | Log($1/IGC_{50}$) |
|---|---|---|
| 1[a] | H | 0.350 |
| 2[a] | 2-$NH_2$ | 0.077 |
| 3[a] | 2-OH | 0.770 |
| 4[b] | 2-$CH_3$ | 0.479 |
| 5[a] | 2-Cl | 0.676 |
| 6[b] | 2-Br | 0.863 |
| 7[a] | 2-$CH_2OH$ | −0.155 |
| 8[a] | 2-$C_6H_5$ | 1.300 |
| 9[a] | 2-$CONH_2$ | −0.721 |
| 10[b] | 2-$NO_2$ | 1.250 |
| 11[a] | 3-$NH_3$ | 0.026 |
| 12[a] | 3-OH | 0.506 |
| 13[a] | 3-$CH_3$ | 0.572 |
| 14[a] | 3-Cl | 0.836 |
| 15[a] | 3-CN | 0.451 |
| 16[a] | 3-$CH_2OH$ | −0.220 |
| 17[b] | 3-$C_6H_5$ | 1.570 |
| 18[a] | 3-$CONH_2$ | −0.193 |
| 19[a] | 3-CHO | 0.140 |
| 20[a] | 3-$NO_2$ | 0.762 |
| 21[b] | 3-$OCH_3$ | 0.670 |
| 22[a] | 4-$CH_3$ | 0.796 |
| 23[a] | 4-$C_2H_5$ | 0.804 |
| 24[a] | 4-$OCH_3$ | 0.544 |
| 25[b] | 4-$OC_2H_5$ | 0.829 |
| 26[a] | 4-$OC_4H_9$ | 1.420 |
| 27[a] | 4-F | 0.253 |
| 28[a] | 4-Cl | 0.559 |
| 29[a] | 4-Br | 0.461 |
| 30[a] | 4-$CH_2CN$ | 0.132 |
| 31[a] | 4-$CH_2Cl$ | 1.180 |
| 32[a] | 4-CH=NOH | 0.678 |
| 33[a] | 4-$NHC_6H_5$ | 1.890 |
| 34[b] | 4-$CH_2OH$ | 0.101 |
| 35[a] | 4-$COOCH_3$ | 0.398 |
| 36[a] | 4-$COOC_2H_5$ | 0.710 |
| 37[a] | 4-$CONH_2$ | −0.179 |
| 38[b] | 4-CHO | 0.203 |
| 39[b] | 4-$NO_2$ | 1.300 |

[a] Training set.
[b] Prediction set.

listed in Table 1. We randomly divided the 39 compounds into two subsets, a training set of 30 compounds and a test set of 9 compounds.

### 3.3. Quantum chemical descriptors calculation

The molecular structures of all the nitrobenzenes were built with Hyperchem software for structural chemistry. Gaussian 98 [29] was operated to optimize with the 6–31$^{++}$G** basis set for all atoms at the B3LYP level. No molecular symmetry constraint was applied; instead, full optimization of all bond lengths and angles was carried out at the B3LYP/6–31$^{++}$G** level. The calculated descriptors for each molecule are summarized in Table 2. Local charges (LC) and electrostatic potential (EP) [30] at each atom, highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital (LUMO) energies, molecular polarizabilities (MP) and molecular dipole moment (MDP) were calculated by Gaussian 98. Quantum chemical indices of hardness ($\eta$), softness ($S$), electronegativity ($\chi$), chemical potential ($\mu$) and electrophilicity ($\omega$) were calculated according to the method proposed by Thanikaivelan et al. [31].

## 4. Results and discussion

### 4.1. Principal component analysis (PCA) of the data set

In order to detect the homogeneities in the data set and identify possible outliers and clusters, PCA was performed within the calculated structure descriptors space for the whole data set. PCA is a useful multivariate statistical technique in which new variables (called principal components, PCs) are calculated as linear combinations of the old ones. These PCs are sorted by decreasing information content (i.e. decreasing variance) so that most of the information is preserved in the first few PCs. An important feature is that the obtained PCs are uncorrelated, and they can be used to derive scores which can be used to display most of the original variations in a smaller number of dimensions. These scores can also allow us to recognize groups of samples with similar behavior. The detailed description of the PCA can be found in Ref. [32].

Table 2
The calculated quantum chemical descriptors used in this study

| Descriptor name | Notation | Description |
|---|---|---|
| Local charges | $LC_i$ | The local charges at each atom of the base unit of nitrobenzenes |
| Electrostatic potential | $EP_i$ | The electrostatic potential at each atom of the base unit of nitrobenzenes |
| Molecular polarizability | MP | Total molecular polarizability |
| Dipole moment | DM | Total molecular dipole moment |
| HOMO | $E_{HOMO}$ | Highest occupied molecular orbital energy |
| LUMO | $E_{LUMO}$ | Lowest unoccupied molecular orbital energy |
| Electronegativity | $\chi$ | $-0.5(E_{HOMO} - E_{LUMO})$ |
| Hardness | $\eta$ | $0.5(E_{HOMO} + E_{LUMO})$ |
| Softness | $S$ | $1/\eta$ |
| Electrophilicity | $\omega$ | $\chi^2/2\eta$ |

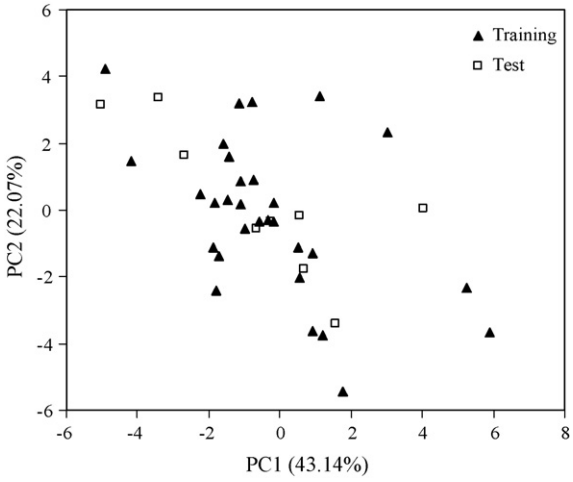


Fig. 2. Principal components analysis of the structural descriptors for the data set.

Here, PCA gives five significant PCs (eigenvalues > 1), which explains 83.44% of the variation in the data (43.14%, 22.07%, 9.23%, 5.47% and 3.53%, respectively). Fig. 2 shows the distribution of compounds over the two first components. As can be seen from Fig. 2, there is not a clear clustering between compounds. The data separation is very important in the development of reliable and robust QSPR models. The quality of the prediction depends on the data set used to develop the model. The toxicity of 39 specified nitrobenzenes were randomly classified into a training set (30 toxicity data) and a prediction set (nine toxicity data). As shown in Fig. 2, the distribution of the compounds in each subset seems to be relatively well-balanced over the space of the principal components. The data were centered to zero means and scaled to the unit variance.

The data set of 39 nitrobenzenes includes recent data on toxicity [3] as summarized in Table 1. The calculated descriptors for each molecule are summarized in Table 2. For the evaluation of the predictive ability of a different model, the root mean square error of prediction (RMSEP) and relative standard error of prediction (RSEP) can be used:

$$\mathrm{RMSEP} = \sqrt{\frac{\sum_{i=1}^{n}(y_{i,\mathrm{pred}} - y_{i,\mathrm{obs}})^2}{n}} \quad (15)$$

$$\mathrm{RSEP}\,(\%) = 100 \times \sqrt{\frac{\sum_{i=1}^{n}(y_{i,\mathrm{pred}} - y_{i,\mathrm{obs}})^2}{\sum (y_{i,\mathrm{obs}})^2}} \quad (16)$$

where $y_{i,\mathrm{pred}}$ is the predicted toxicity using different model, $y_{i,\mathrm{obs}}$ the observed value of the toxicity and $n$ is the number of samples in the prediction set.

### 4.2. MLR analysis

Among the descriptors mentioned in Section 3.3, the most significant molecular descriptors were identified using multiple linear regression analysis with a stepwise forward selection

method. The best equation obtained for the toxicity of the nitrobenzenes derivatives was:

$$\log\left(\frac{1}{IGC_{50}}\right) = 24.71 + 0.58LC1 + 11.74LC2 + 9.42LC3 + 13.89LC4 + 3.25EP2 + 4.32EP3 + 3.89EP4 + 3.21S + 2.08MP + 1.69DM + 0.53\omega$$

where LC1, LC2, LC3, LC4, EP2, EP3, EP4, $S$, MP, DM and $\omega$ are the local charges and electrostatic potentials on carbon atom C1, C2, C3 and C4, softness, molecular polarizability, dipole moment and electrophilicity, respectively. In this model, the highly correlated descriptors were not considered. As seen, the resulting model has eleven significant descriptors (correlation coefficient > 0.5). Table 3 shows the descriptors coefficients, the standard error of coefficients, the $t$-values for null hypothesis, and their related $P$-values.

### 4.3. PLS analysis

The factor-analytical multivariate calibration method is a powerful tool for modeling, because it extracts more information from the data and allows building more robust models [33–38]. According to toxicity data (Table 1), data randomly classified to training and prediction sets. The PLS model was run twice. In first run (run a), all calculated descriptors (Section 3.3) were considered in modeling; while in the second run (run b), after finding the effective descriptors by the genetic algorithm (GA) procedure, only the effective descriptors were considered. A GA is a stochastic method to solve optimization problems defined a fitness criterion applying the evolution hypothesis of Darvin and different genetic functions, i.e. crossover and mutation. Leardi et al. [39–42] demonstrated that genetic algorithm after suitable modifications, produces more interpretable results, since the selected variables are less dispersed than in other methods. The algorithm used in this paper is an evolution of the algorithm described in Refs. [39,40], whose parameters are reported under Table 4.

Table 3
Results of multiple linear regression analysis

| Descriptor | Coefficient | S.E.[a] of coefficient | $t$-Value | $P$-Value |
|---|---|---|---|---|
| Intercept | 24.71 | 3.78 | 8.42 | 0.0001 |
| LC1 | 0.58 | 0.41 | 1.08 | 0.001 |
| LC2 | 11.74 | 2.64 | 3.78 | 0.001 |
| LC3 | 9.42 | 2.11 | 3.26 | 0.001 |
| LC4 | 13.89 | 3.04 | 3.91 | 0.001 |
| EP2 | 3.25 | 1.21 | 1.89 | 0.031 |
| EP3 | 4.32 | 1.32 | 2.96 | 0.032 |
| EP4 | 3.89 | 1.14 | 2.74 | 0.030 |
| $S$ | 3.21 | 0.89 | 2.53 | 0.001 |
| MP | 2.08 | 0.76 | 2.05 | 0.001 |
| DM | 1.69 | 0.68 | 1.86 | 0.0001 |
| $\omega$ | 0.53 | 0.39 | 0.87 | 0.001 |

[a] Standard error.

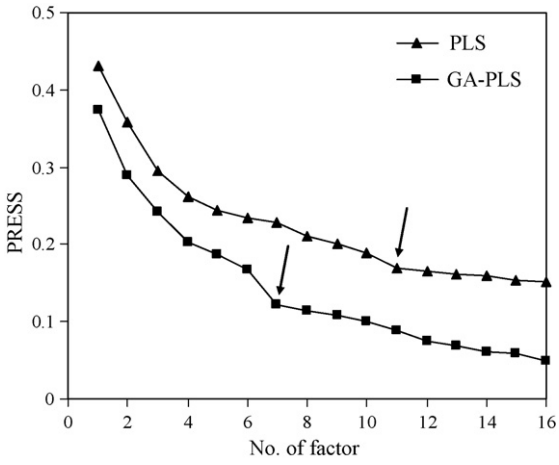


Fig. 3. Plots of PRESS vs. number of factors by PLS and GA-PLS.

The optimum number of factors (latent variables) to be included in the calibration model was determined by computing the prediction error sum of squares (PRESS) fro cross-validated models using a high number of factors (half of the number of total training set + 1) [43]. The cross-validation method employed was to eliminate only one compound at a time and then PLS calibrated the remaining of training set. The toxicity of the left-out sample was predicted by using this calibration. This process was repeated until each compound in the training set had been left out once. According to Haaland suggestion [43], the optimum number of factor was selected. In Fig. 3, the PRESS obtained by optimizing the training set of the descriptor data with PLS and GA-PLS models are shown. Table 4 shows the optimum number of factor and PRESS value. However, modeling with GA-PLS, is a little better than of PLS.

### 4.4. LS-SVM analysis

The all descriptors were used as the input to develop nonlinear model by LS-SVM. The quality of LS-SVM for regression depend on $\gamma$ and $\sigma^2$ parameters. In this work, LS-SVM was performed with radial basis function (RBF) as a kernel function. To determine the optimal parameters, a grid search was performed based on leave-one-out cross-validation on the original training set for all parameter combinations of $\gamma$ and $\sigma^2$ from 1 to 200 and 1 to 100, respectively, with increment steps of 1. Table 4 shows the optimum $\gamma$ and $\sigma^2$ parameters for the LS-SVM and RBF kernel, using the calibration sets for 30 toxicity data.

### 4.5. Prediction of toxicity of nitrobenzenes

The predictive ability of these methods (MLR, PLS, GA-PLS and LS-SVM) were determined using nine toxicity data (their structure are given in Table 1). The results obtained by MLR, PLS, GA-PLS and LS-SVM methods are listed in Tables 4 and 5. Table 4 also shows RMSEP, RSEP and the percentage error for prediction of toxicity of nitrobenzenes. As can be seen, the percentage error was also quite acceptable only for LS-SVM. Good results were achieved in LS-SVM model with percentage

Table 4
Actual and predicted values of toxicity for nitrobenzenes using MLR, PLS, GA-PLS and LS-SVM models

| Substituent, R (Fig. 1) | Actual log(1/IGC$_{50}$) | Predicted log(1/IGC$_{50}$) | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | MLR | Error (%) | PLS | Error (%) | GA-PLS[a] | Error (%) | LS-SVM | Error (%) |
| 2-$CH_3$ | 0.479 | 0.427 | −10.86 | 0.438 | −8.56 | 0.431 | −10.02 | 0.481 | 0.42 |
| 2-Br | 0.863 | 0.784 | −9.15 | 0.794 | −8.00 | 0.807 | −6.49 | 0.860 | −0.35 |
| 2-$NO_2$ | 1.250 | 1.348 | 7.84 | 1.358 | 8.64 | 1.362 | 8.96 | 1.247 | −0.24 |
| 3-$C_6H_5$ | 1.570 | 1.712 | 9.04 | 1.641 | 4.52 | 1.618 | 3.06 | 1.565 | −0.32 |
| 3-$OCH_3$ | 0.670 | 0.503 | −24.93 | 0.511 | −23.73 | 0.563 | −15.97 | 0.658 | −1.79 |
| 4-$OC_2H_5$ | 0.829 | 0.741 | −10.62 | 0.724 | −12.67 | 0.789 | −4.83 | 0.827 | −0.24 |
| 4-$CH_2OH$ | 0.101 | 0.114 | 12.87 | 0.112 | 10.89 | 0.108 | 6.93 | 0.101 | 0.00 |
| 4-CHO | 0.203 | 0.251 | 23.65 | 0.237 | 16.75 | 0.219 | 7.88 | 0.204 | 0.49 |
| 4-$NO_2$ | 1.300 | 1.078 | −17.08 | 1.103 | −15.15 | 1.191 | −8.38 | 1.296 | −0.31 |
| NF[b] | | | | 11 | | 7 | | | |
| PRESS | | | | 0.1702 | | 0.1230 | | | |
| $\gamma$ | | | | | | | | 94 | |
| $\sigma^2$ | | | | | | | | 14 | |
| RMSEP | | 0.1184 | | 0.0769 | | 0.0152 | | 0.0049 | |
| RSEP (%) | | 12.6509 | | 8.2197 | | 1.6218 | | 0.5187 | |

[a] Parameters for genetic algorithm. Population size: 30 chromosomes; probability of mutation: 1%; windows size for smoothing: 3.
[b] Number of factor.

Table 5
Comparison of the statistical parameters by different QSPR models for the prediction of the log(1/IGC$_{50}$)

| Methods | Data set | $R^2$ | $Q^{2a}$ |
|---|---|---|---|
| MLR | Training | 0.9743 | |
| | Test | 0.9403 | |
| PLS | Training | 0.9678 | 0.8123 |
| | Test | 0.9412 | 0.8022 |
| GA-PLS | Training | 0.9875 | 0.8486 |
| | Test | 0.9718 | 0.8124 |
| LS-SVM | Training | 0.9999 | 0.9326 |
| | Test | 0.9995 | 0.9214 |

[a] Coefficient for the model validation by leave-one-out.

error ranges from −1.79 to 0.24 for toxicity of nitrobenzenes. The plots of the predicted toxicity versus actual values are shown in Fig. 4 for each model (line equations and $R^2$-values are also shown). The correlation coefficients ($R^2$) for LS-SVM model were better than other models and close to one. Also, it is possible to see that LS-SVM presents excellent prediction abilities when compared with other regression.

According to the results, quantum chemical descriptors are suitable descriptors for describing the toxicity of nitrobenzene derivatives. In MLR and GA-PLS methods, in which more effective descriptors are used, it is seen that $LC_i$ and $EP_i$ have larger effects on the toxicities of nitrobenzenes in atoms number 2, 3 and 4. And also when LS-SVM method with all descriptors is used, prediction of toxicity in test step, with a small error is possible, which is improved in comparison with other methods

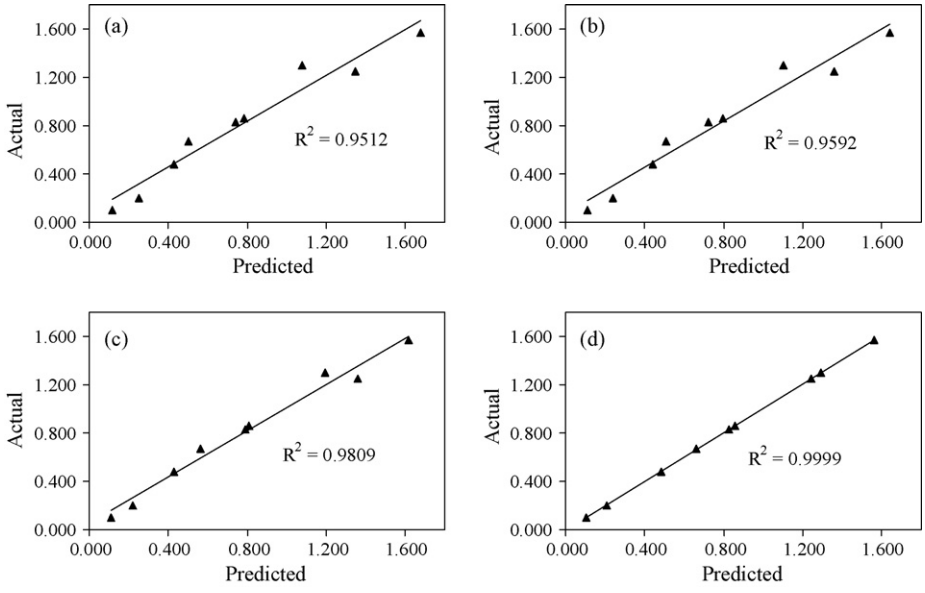


Fig. 4. Plots of predicted vs. actual toxicity for nitrobenzenes with (a) MLR, (b) PLS, (c) GA-PLS and (d) LS-SVM.

(MLR, PLS and GA-PLS), which shows that by using all chemical quantum descriptors and also LS-SVM method, the toxicity of nitrobenzene is predicted with satisfactory results.

## 5. Conclusion

LS-SVM was established to predict the toxicity of some nitrobenzenes. A suitable model with high statistical quality and low prediction errors was obtained. The model can accurately predict toxicity of nitrobenzenes that do not exist in the modeling procedure. The quantum chemical descriptors concerning all the molecular properties and those of individual atoms in the molecule were found to be important factors controlling the toxicity behavior. In this study, the results obtained by LS-SVM, are compared with results obtained by MLR, PLS and GA-PLS. The results show that, LS-SVM is more powerful in prediction of toxicity of nitrobenzenes than MLR, PLS and GA-PLS.

## References

[1] A.R. Katrizky, P. Oliferenko, A. Oliferenko, A. Lomaka, M. Karelson, J. Phys. Org. Chem. 16 (2003) 811–817.
[2] M.T.D. Cronin, B.W. Gregory, T.W. Schultz, Chem. Res. Toxicol. 11 (1998) 902–908.
[3] J.C. Dearden, M.T.D. Cronin, T.W. Lin, D.T. Lin, Quant. Struct. Act. Relat. 14 (1995) 427–432.
[4] M.T.D. Cronin, T.W. Schultz, Chem. Res. Toxicol. 14 (2001) 1284–1295.
[5] Y.P. Zhou, J.H. Jiang, W.Q. Lin, H.Y. Zou, H.L. Wu, G.L. Shen, R.Q. Yu, Eur. J. Pharm. Sci. 28 (2006) 344–353.
[6] E.L. Willigghagen, R. Wehrens, L.M.C. Buydens, Crit. Rev. Anal. Chem. 36 (2006) 189–198.
[7] A. Niazi, S. Jameh-Bozorghi, D. Nori-Shargh, Turk. J. Chem. 30 (2006) 619–628.
[8] M. Kompany-Zareh, Acta Chim. Slov. 50 (2003) 259–273.
[9] M. Shamsipur, B. Hemmateenejad, M. Akhond, H. Sharghi, Talanta 54 (2001) 1113–1120.
[10] J.M. Luco, F.H. Ferretti, J. Chem. Inform. Sci. 37 (1997) 392–401.
[11] B. Hemmateenejad, M.A. Safarpour, F. Taghavi, J. Mol. Struct. (TheoChem.) 635 (2003) 183–190.
[12] A.I. Belousov, S.A. Verzakov, J. Von Frese, J. Chemometr. Intell. Syst. 64 (2002) 15–25.
[13] R. Burbidge, M. Trotter, B. Buxton, S. Holden, Comput. Chem. 26 (2001) 5–14.
[14] J.A.K. Suykens, J. Vandewalle, Neural Process Lett. 9 (1999) 293–300.
[15] J.A.K. Suykens, T. van Gestel, J. de Brabanter, B. de Moor, J. Vandewalle, Least-squares Support Vector Machines, World Scientifics, Singapore, 2002.
[16] F.A. Molfetta, A.T. Bruni, F.P. Rosselli, A.B.F. da Silva, Struct. Chem. 18 (2007) 49–57.
[17] O. Isayev, B. Rasulev, L. Gorb, J. Leszczynski, Mol. Divers. 10 (2006) 233–245.
[18] E. Zvinavashe, A.J. Murk, J. Vervoort, A.E.M.F. Sofferes, A. Freidig, I.M.C.M. Rietjens, Environ. Toxicol. Chem. 25 (2006) 2313–2321.
[19] U. Thissen, B. Ustun, W.J. Melssen, L.M.C. Buydens, Anal. Chem. 76 (2004) 3099–3105.
[20] A. Borin, M.F. Ferrao, C. Mello, D.A. Maretto, R.J. Poppi, Anal. Chim. Acta 579 (2006) 25–32.
[21] A. Niazi, J. Ghasemi, A. Yazdanipour, Spectrochim. Acta A, 2007, doi:10.1016/2006.12.022.
[22] N. Acir, Neural Comput. Appl. 14 (2005) 299–309.
[23] Y. Ke, C. Yiyu, Chin. J. Anal. Chem. 34 (2006) 561–564.
[24] J. Shi, X. Liu, J. Appl. Polym. Sci. 101 (2006) 285–289.
[25] J. Li, H. Liu, X. Yao, M. Liu, Z. Hu, B. Fan, Anal. Chim. Acta 581 (2007) 333–342.
[26] G.B. Arfken, H.J. Weber, Mathematical Methods for Physicists, 4th ed., Academic Press, New York, 1995.
[27] J. Mercer, Philos. Trans. Roy. Soc. Lond. A 209 (1909) 415.
[28] V. Vapnik, in: J.A.K. Suykens, J. Vandewalle (Eds.), Nonlinear Modeling: Advanced Black-box techniques, Kluwer Academic Publishers, Boston, 1998.
[29] M.J. Frisch, G.W. Trucks, H.B. Schlegel, G.E. Scuseria, M.A. Robb, J.R. Cheeseman, V.G. Zakrzewski, J.A. Montgomery Jr., R.E. Stratmann, J.C. Burant, S. Dapprich, J.M. Millam, A.D. Daniels, K.N. Kudin, M.C. Strain, O. Franks, J. Tomasi, V. Barone, M. Cossi, R. Cammi, B. Mennucci, C. Pomelli, C. Adamo, S. Clifford, J. Ochterski, G.A. Petersson, P.Y. Ayala, Q. Cui, K. Morokuma, D.K. Malick, A.D. Rabuck, K. Raghavachari, J.B. Foresman, J. Cioslowski, J.V. Ortiz, B.B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. Gomperts, R.L. Martin, D.J. Fox, T. Keith, M.A. Al-Laham, C.Y. Peng, A. Nanayakkara, C. Gonzalez, M. Challacombe, P.M.W. Gill, B. Johanson, W. Chen, M.W. Wong, J.L. Andres, C. Gonzalez, M. Head-Gordon, E.S. Replogle, J.A. Pople, GAUSSIAN 98, Gaussian Inc., Pittsburg, PA, 1998.
[30] I.N. Levine, Quantum Chemistry, 5th ed., Prentice Hall, 2000.
[31] P. Thanikaivelan, V. Subramanian, J.R. Rao, B.U. Nair, Chem. Phys. Lett. 323 (2000) 59.
[32] S. Sharma, Applied Multivariate Techniques, John Wiley, Singapore, 1996.
[33] J. Ghasemi, A. Niazi, Talanta 65 (2005) 1168–1173.
[34] J. Ghasemi, A. Niazi, Anal. Chim. Acta 533 (2005) 169–177.
[35] A. Niazi, J. Ghasemi, A. Yazdanipour, Anal. Lett. 38 (2005) 2377–2392.
[36] A. Niazi, J. Braz. Chem. Soc. 17 (2006) 1020–1026.
[37] A. Niazi, Croat. Chim. Acta 79 (2006) 573–579.
[38] A. Niazi, A. Soufi, M. Mobarakabadi, Anal. Lett. 39 (2006) 2359–2372.
[39] R. Leardi, R. Boggia, M. Terrile, J. Chemometr. 6 (1992) 267–281.
[40] R. Leardi, J. Chemometr. 8 (1994) 65–79.
[41] J. Ghasemi, A. Niazi, R. Leardi, Talanta 59 (2003) 311–317.
[42] J. Ghasemi, D.M. Ebrahimi, L. Hejazi, R. Leardi, A. Niazi, J. Anal. Chem. 61 (2006) 92–98.
[43] D.M. Haaland, E.V. Thomas, Anal. Chem. 60 (1988) 1193–1202.