Journal of Hazardous Materials 177 (2010) 344–351

Contents lists available at ScienceDirect

Journal of Hazardous Materials

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

QSAR modeling of toxicity of diverse organic chemicals to *Daphnia magna* using 2D and 3D descriptors

Supratik Kar, Kunal Roy*

Drug Theoretics and Cheminformatics Laboratory, Department of Pharmaceutical Technology, Jadavpur University, Raja S C Mullick Road, Kolkata 700032, India

ARTICLE INFO

Article history:
Received 3 September 2009
Received in revised form 4 December 2009
Accepted 6 December 2009
Available online 31 December 2009

Keywords:
QSAR
QSTR
REACH
Daphnia magna
Chemometric tools
Validation

ABSTRACT

One of the major economic alternatives to experimental toxicity testing is the use of quantitative structure–activity relationships (QSARs) which are used in formulating regulatory decisions of environmental protection agencies. In this background, we have modeled a large diverse group of 297 chemicals for their toxicity to *Daphnia magna* using mechanistically interpretable descriptors. Three-dimensional (3D) (electronic and spatial) and two-dimensional (2D) (topological and information content indices) descriptors along with physicochemical parameter log $K_{o/w}$ (n-octanol/water partition coefficient) and structural descriptors were used as predictor variables. The QSAR models were developed by stepwise multiple linear regression (MLR), partial least squares (PLS), genetic function approximation (GFA), and genetic PLS (G/PLS). All the models were validated internally and externally. Among several models developed using different chemometric tools, the best model based on both internal and external validation characteristics was a PLS equation with 7 descriptors and three latent variables explaining 67.8% leave-one-out predicted variance and 74.1% external predicted variance. The PLS model suggests that higher lipophilicity and electrophilicity, less negative charge surface area and presence of ether linkage, hydrogen bond donor groups and acetylenic carbons are responsible for greater toxicity of chemicals. The developed model may be used for prediction of toxicity, safety and risk assessment of chemicals to achieve better ecotoxicological management and prevent adverse health consequences.

1. Introduction

On a global scale, the enormous annoying effect of hazardous chemicals and pollutants on the ecosystem is an issue of great concern considering that though large numbers of chemical compounds are in commercial use, relatively few of these have been subjected to adequate assessment for their perilous environmental properties [1]. The global production of chemicals has increased from 1 million tons in 1930 to 400 million tons in the 21st century. Around 100,000 different chemicals are registered in the European market of which 10,000 are marketed in volumes of more than 10 tons, and a further 20,000 are marketed at 1–10 tons per year [2]. Testing and assessing their risks to human health and the environment according to the European Commission Directive 67/548 [3] are required before marketing in volumes above 10 kg per year. For higher volumes, more in-depth testing and focusing on long-term and chronic effects are required [3]. In contrast, more than 99% of the total volumes of all substances in the market are not subject to the same testing requirements. Some of them have never been tested at all.

There is an essential need to use computation-based quantitative structure–activity relationship (QSAR) modeling for providing information about the physicochemical properties of chemicals and their environmental fate as well as their human health effects [4]. Advanced predictive models are being designed and tested by regulatory agencies to assess physical, chemical, and biological properties of individual chemical entities using applications specific for decision-making frameworks in safety assessments. The use of QSAR modeling for toxicological predictions would help determine the potential adverse effects of chemical entities in risk assessment, chemical screening, and priority setting [5].

In addition, the recent European Union REACH (Registration, Evaluation and Authorisation of Chemicals) legislation requires toxicological hazard and risk assessments for all new and existing chemicals, and QSAR will play an important role in this endeavour [6]. In the European Union, the use of in silico methods is explicitly encouraged and even required in the REACH regulation [3], which

* Corresponding author. Tel.: +91 98315 94140; fax: +91 33 2837 1078.
E-mail address: kunalroy_in@yahoo.com (K. Roy).
URL: http://sites.google.com/site/kunalroyindia (K. Roy).

0304-3894/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.jhazmat.2009.12.038

Table 1
Categorical list of descriptors used in the development of QSAR models.

Category of descriptors	Name of the descriptors
Topological	Balaban (Jx), kappa shape index ($^1\kappa$, $^2\kappa$, $^3\kappa$, $^1\kappa_{am}$, $^2\kappa_{am}$, $^3\kappa_{am}$), flexibility (φ), subgraph count (SC-0, SC-1, SC-2, SC-3_P, SC-3_C), connectivity index ($^0\chi$, $^1\chi$, $^2\chi$, $^3\chi_p$, $^3\chi_c$, $^0\chi^v$, $^1\chi^v$, $^2\chi^v$, $^3\chi_p^v$, $^3\chi_c^v$), Wiener, Zagreb, electrotopological state fragment type (S_sCH$_3$, S_ssCH$_2$, S_aaCH, S_sssCH, S_dssC, S_aasC, S_ssssC, S_dsN, S_sssN, S_sOH, S_ddssS, S_dO, S_ssO, S_ssS, S_dssS, S_sF, S_sCl, S_sBr)
Structural	MW, Rotlbonds, H-bond acceptor, H-bond donor
Electronic	Dipole-mag, HOMO, LUMO, Sr
Spatial	RadOfGyration, Jurs_SASA, Jurs_PPSA_1, Jurs_PNSA_1, Jurs_DPSA_1, Jurs_PPSA_2, Jurs_PNSA_2, Jurs_DPSA_2, Jurs_PPSA_3, Jurs_PNSA_3, Jurs_DPSA_3, Jurs_FPSA_1, Jurs_FNSA_1, Jurs_FPSA_2, Jurs_FNSA_2, Jurs_FPSA_3, Jurs_FNSA_3, Jurs_WPSA_1, Jurs_WNSA_1, Jurs_WPSA_2, Jurs_WNSA_2, Jurs_WPSA_3, Jurs_WNSA_3, Jurs_RPCG, Jurs_RNCG, Jurs_RPCS, Jurs_RNCS, Jurs_TPSA, Jurs_TASA, Jurs_RPSA, Jurs_RASA, Shadow_XY, Shadow_XZ, Shadow_YZ, Shadow_XYfrac, Shadow_XZfrac, Shadow_YZfrac, Shadow_nu, Shadow_Xlength, Shadow_Ylength, Shadow_Zlength, Area, Vm, Density, PMI_mag
Information indices	IC (mean information content index), SIC (structural information content index), CIC (complementary information index), BIC (bonding information content index).

came into force from 1st June 2007. This regulation aims, among other things, at identifying, evaluating and regulating "*Persistent, Bioaccumulating and Toxic substances*" effectively [3]. REACH aims to provide toxicity information for about 30,000 out of the more than 1,00,000 chemicals listed on the European Inventory of Existing Commercial Chemical Substances (EINECS), for which there is insufficient toxicological information on their hazardous properties. Within REACH, there are requirements to use sufficiently validated computational prediction models based on QSAR to fill in the toxicity data gaps, and thus save time, money and help to reduce the numbers of animals used for experimental testing purposes [3]. Guidelines for QSAR model development and validation proposed by the Organization for Economic Cooperation and Development (OECD) are expected to help increase the acceptability of QSAR models for regulatory purposes [7].

The goal of any ecotoxicological QSAR is to determine the efficiency of the developed QSAR model for the toxicity of chemicals which cover a large structural diversity spanning a variety of mechanisms of toxic action, including narcoses and electrophilic mechanisms. These chemicals are capable of causing a wide range of adverse effects including general toxicity, allergenic reactions, mutagenicity, and carcinogenicity [8]. Dermal, oral, or respiratory exposures include gastrointestinal, neurological, and reproductive disorders; liver cirrhosis; hepatitis; cataracts; respiratory and skin irritation; nephrotoxicity; and hematological defects [9–11].

Daphnia magna, an important freshwater invertebrate species in aquatic food webs, has been used world-wide for many years as a representative test species for ecotoxicological evaluation of industrial chemicals [12–14]. For the development of quantitative toxicity models also, *D. magna* has been largely used in recent time. Deneer [15] developed QSAR models to accurately predict the joint acute toxicity of 50 organic chemicals to *D. magna*. Tao et al. developed a fragment constant QSAR model for evaluating the EC_{50} values of 217 organic chemicals to *D. magna* [16]. Zvinavashe et al. developed QSAR model for the toxicity of a series of organothiophosphate pesticides to *D. magna* [17].

Von der ohe et al. [18] developed QSAR for the toxicity of diverse organic chemicals to *D. magna* using the octanol–water partition coefficient ($\log K_{o/w}$) as a predictor variable. However, their statistical analysis was confined to only internal validation without true external validation. Also, they never used the total number of chemicals available for model development. They developed various models with 36, 33,193, 91 and 17 chemicals in different runs. In the present paper, we have used the same dataset of toxicity of 300 organic chemicals to *D. magna* and developed predictive global QSAR models on the dataset using 2D, 3D and combination of both types of descriptors along with $\log K_{o/w}$ and structural parameters. Sufficient validation strategies (internal and external validations, model randomization) have been applied to check the predictability of the developed models.

2. Materials and methods

2.1. Dataset

Three hundred diverse organic chemicals with 48 h *D. magna* toxicity in terms of $\log(LC_{50})$ reported by von der Ohe et al. [18] were used as the model dataset. Toxicity values were multiplied with −1 and thus $\log(LC_{50})$ values were converted to $\log(1/LC_{50})$ which was used as the response variables. The data set covers a $\log K_{o/w}$ (octanol/water partition coefficient) range from −2 to 8 and a toxicity (daphnia) range of 0.46–10.09. In regard to the chemical domain, the data set includes hydrocarbons, aliphatic alcohols, phenols, ethers, and esters; anilines, amines, nitriles, nitroaromatics, amides, and carbamates; urea and thiourea derivatives; iso-thiocyanates; thiols; phosphorothionate and phosphate esters; and halogenated derivatives. The list of compounds along with their daphnia toxicity values are shown in Table S1 in Supplementary Materials.

Three chemicals were excluded from the modeling exercise due to their atypical nature [diquat (containing quaternary nitrogen), mancozeb (metallic compound), dithiocarbamate (undefined structure in the reference paper)]. So, our modeling work was carried out with 297 organic chemicals.

2.2. Descriptor calculation

We have performed QSAR studies on 297 chemicals reported by von der Ohe et al. with two-dimensional (topological and information) and three-dimensional (spatial and electronic) descriptors along with $\log K_{o/w}$ and a few structural descriptors. The categorical list [19] of descriptors used in the development of QSAR models is reported in Table 1. The listed descriptors (Table 1) have been selected for the present study for their wide spread use and easy interpretability in terms of mechanism of action and/or physical meaning.

For the calculation of 3D descriptors, multiple conformations of each molecule were generated using "optimal search" as the conformational search method using Cerius2 version 4.10 software [19] followed by an energy minimization using smart minimizer under open force field (OFF) to generate the lowest energy conformation for each structure. The charges were calculated according to the Gasteiger method.

2.3. Training set selection

In our present work, the total data set ($n = 297$) was divided into training set ($n = 222$) and test (external evaluation) set ($n = 75$)

(75% and 25% respectively of the total number of compounds) based on clusters obtained from *k*-means clustering [20–22] applied on topological, information and structural indices descriptor matrix. The details of the clustering method are given in Supplementary Materials section. The whole data set was clustered into five subgroups from each of which 25% of compounds were selected as members of the test set. Identification numbers of compounds under different clusters are shown in Table S2 in Supplementary Materials section.

2.4. Chemometric tools

Statistical techniques like stepwise regression [23], partial least squares (PLS) [24,25], genetic function approximation (GFA-MLR) [26] and genetic partial least squares (G/PLS) [19,25,27,28] were applied to identify the structural and physicochemical features contributing to the toxicity of chemicals. The details of the chemometric tools are discussed in Supplementary Materials.

2.5. Software

MINITAB [29] was used for stepwise regression and partial least squares methods. Cerius2 version 4.10 [19] was used for GFA and G/PLS analyses and descriptor calculation. SPSS [30] was used for *k*-means cluster analysis and preparation of intercorrelation matrix of the descriptors. STATISTICA [31] was used to determine the LOO predicted values of training set compounds. Final PLS model development based on variables selected from stepwise regression model was performed using SIMCA-P [32] (*vide infra*).

2.6. Validation methods

The robustness of the models was verified by using different types of validation criteria. For validation of QSAR models, three strategies [33] were adopted: (1) leave-one-out (LOO) internal validation or cross-validation, (2) validation by dividing the data set into training and test compounds, (3) data/model randomization or *Y*-scrambling.

The main target of any QSAR modeling is that the developed model should be robust enough to be capable of making accurate and reliable predictions of biological activities of new compounds [34–36]. So, QSAR models that are developed from a training set should be validated using new chemical entities for checking the predictive capacity of the developed models. The validation strategies check the reliability of the developed models for their possible application on a new set of data, and confidence of prediction can thus be judged [36].

For all the developed models we have reported the coefficient of variation (R^2), leave-one-out cross-validation R^2 (Q^2) and $r^2_{m(\mathrm{LOO})}$ for the training set, the R^2_{pred} and $r^2_{m(\mathrm{test})}$ values for the test set and $r^2_{m(\mathrm{overall})}$ for the total set [37–39]. The details of the validation tools are discussed in Supplementary Materials section. For calculation of R^2_{pred}, training set mean has been used as usual. However, additionally, test set mean has been used for calculation of R^2_{pred} (corrected) as suggested by Schürmann et al. [40].

The final model was also subjected to a randomisation test. In this test, the toxicity data (Y) are randomly permuted keeping the descriptor matrix intact, followed by a PLS run. Each randomisation and subsequent PLS analysis generates a new set of R^2 and Q^2 values, which are plotted against the correlation coefficient between the original Y values and the permuted Y values. The intercepts for the R^2 and Q^2 lines in this plot are a measure of the overfit. A model is considered [41] valid if $R^2_{\mathrm{int}} < 0.4$ and $Q^2_{\mathrm{int}} < 0.05$.

Table 2
Statistical quality of different models.

Type of descriptors apart from $\log K_{o/w}$ and structural indices	Statistical methods	Model no.	No. of descriptors	LVs	R^2	$Q^2_{(\mathrm{LOO})}$	$r^2_{m(\mathrm{LOO})}$	R^2_{pred}	$r^2_{m(\mathrm{test})}$	$r^2_{m(\mathrm{overall})}$
2D (topological + information)	Stepwise regression	1	11	–	0.736	0.684	0.534	0.643	0.610	0.545
	PLS	2	8	2	0.694	0.650	0.627	0.674	0.664	0.643
	GFA (linear)	3	7	–	0.700	0.671	0.523	0.691	0.692	0.546
	GFA (spline)	4	6	–	0.703	0.673	0.530	0.632	0.616	0.543
	G/PLS (linear)	5	6	4	0.675	0.627	0.601	0.652	0.650	0.612
	G/PLS (spline)	6	4	3	0.654	0.637	0.631	0.631	0.634	0.630
3D (spatial + electronic)	Stepwise regression	7	9	–	0.691	0.656	0.512	0.593	0.563	0.518
	PLS	8	8	3	0.663	0.630	0.620	0.627	0.612	0.628
	GFA (linear)	9	7	–	0.663	0.635	0.500	0.658	0.635	0.522
	GFA (spline)	10	7	–	0.707	0.680	0.532	0.644	0.650	0.552
	G/PLS (linear)	11	5	4	0.631	0.610	0.601	0.613	0.600	0.611
	G/PLS (spline)	12	9	3	0.660	0.632	0.621	0.610	0.597	0.625
2D + 3D (topological + information + spatial + electronic)	Stepwise regression	13 [Eq. (1)]	12	–	0.738	0.703	0.550	0.721	0.718	0.574
	PLS[a]	14	7	3	0.691	0.666	0.656	0.700	0.663	0.673
	GFA (linear)	15	5	–	0.682	0.657	0.514	0.660	0.648	0.530
	GFA (spline)	16	5	–	0.709	0.692	0.479	0.669	0.670	0.507
	G/PLS (linear)	17	6	4	0.673	0.650	0.641	0.656	0.665	0.645
	G/PLS (spline)	18	8	4	0.650	0.612	0.600	0.626	0.615	0.611
	PLS[b,*]	19 [Eq. (2)]	7	3	0.695	0.677	0.670	0.741	0.707	0.688

[a] Variables selected based on standardized coefficients.
[b] Variables selected based on VIP values.
[*] Based on variables selected from stepwise regression using combined set of descriptors.

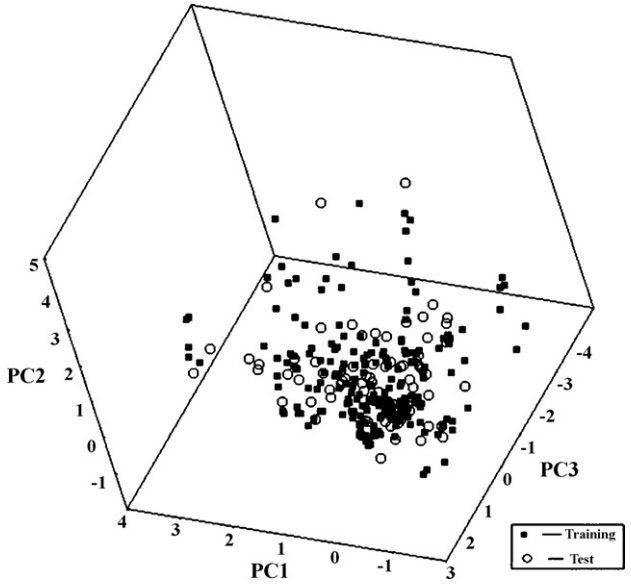

Fig. 1. PCA score plot of first three components for the standardized topological, information and structural descriptor matrix.

2.7. Applicability domain

The applicability domain (AD) of a QSAR is the physico-chemical, structural or biological space, knowledge or information on which the training set of the model has been developed, and for which it is applicable to make predictions for new compounds [42]. The purpose of AD is to state whether the model's assumptions are met. In general, this is the case for interpolation rather than extrapolation. To investigate the AD of a training set one can directly analyse properties of the multivariate descriptor space of the training compounds or, more indirectly, analyse distance (or similarity) metrics. This can be achieved by different means of feature selection and successive principle components analysis. When a compound is highly dissimilar to all compounds of the modeling set, reliable prediction of its activity is unlikely. The concept of AD [43] was used to avoid such an unjustified extrapolation of activity predictions.

The residuals of Y and X are of diagnostic value for the quality of the model [44]. Since there are many X-residuals one needs a summary for each observation (compound). This is accomplished by the residual standard deviation (SD) of the X-residuals of the corresponding row of the residual matrix E. Because this SD is proportional to the distance between the data point and the model plane in X-space, it is also often called DModX (distance to the model in X-space). Here, X is the matrix of predictor variables, of size $(N \times K)$, Y is the matrix of response variables, of size $(N \times M)$ and E is the $(N \times K)$ matrix of X-residuals, N is number of objects (cases, observations), k is the index of X-variables ($k = 1, 2, \ldots, K$) and m is the index of Y-variables ($m = 1, 2, \ldots, M$). A DModX larger than around 2.5 times the overall SD of the X-residuals (corresponding to an F-value of 6.25) indicates that the observation is outside the applicability domain of the model [44].

3. Results and discussion

The principal component analysis (PCA) score plot of the first three components of the standardized topological, information and structural descriptor matrix shows distribution of the training and test set compounds in 3D space. It may be noted that the distribution of the whole dataset into training and test sets has been done by k-means clustering and not using the PCA score plot. However, in Fig. 1, the plot shows that each test set compound is located in the close vicinity of at least one training set compound in the 3D space.

We developed three sets of 6 models, one for each with 2D (73 descriptors), 3D (54 descriptors) and combination of both kinds of descriptors. Note that $\log K_{o/w}$ and structural parameters were used in all trials of model development process. As a result total 18 models are developed with the mentioned chemometric tools. In Table 2 statistical quality of developed models are presented. Based on external prediction criteria, stepwise regression (model 13) derived model obtained from combination of 2D and 3D descriptors evolved as the best model among the 18 models. The developed equation is as follows:

$$\begin{aligned}\log\left(\frac{1}{LC_{50}}\right) &= 4.851 - 0.313(LUMO) + 27.4(Jurs - FNSA - 3) \\ &\quad - 0.224(Jurs - PNSA - 3) + 0.664(\log K_{o/w}) \\ &\quad + 57.01(Jurs - FPSA - 3) - 0.168(Jurs - DPSA - 3) \\ &\quad + 0.132(S\_ssO) + 0.55(Hbonddonor) \\ &\quad + 0.181(Dipole{-}mag) + 0.35(S\_tsC) + 0.149(S\_dsN) \\ &\quad - 2.501(Shadow\text{-}YZfrac) \\ n_{training} &= 222, \\ R^2 = 0.738,\ \ Q^2 = 0.703,\ \ R^2_{adj} &= 0.723,\ \ r^2_{m(LOO)} = 0.55, \\ n_{test}=75,\ \ R^2_{pred}=0.721,\ \ r^2_{m(test)}&=0.718,\ \ r^2_{m(overall)}=0.574 \qquad (1)\end{aligned}$$

Eq. (1) could explain 72.3% of the variance (adjusted coefficient of variation) and could predict 70.3% of the variance (leave-one-out predicted variance). External predicted variance for Eq. (1) is 72.1% which is a fairly high value for such a large number of test set compounds ($n_{test} = 75$). The corrected R^2_{pred} value, calculated using the test set mean [40], for Eq. (1) is 0.719.

To check the intercorrelation among the 12 descriptors in Eq. (1), we checked Pearson correlation matrix by SPSS software [30]. Table S3 in Supplementary Materials section gives the intercorrelation (r) data for all the descriptors used in the stepwise regression equation. Analyzing the matrix we found that Jurs-DPSA-3 and Jurs-FNSA-3 ($r = 0.889$), Jurs-DPSA-3 and Jurs-PNSA-3 ($r = 0.971$), Jurs-PNSA-3 and Jurs-FNSA-3 ($r = 0.948$) descriptors are highly correlated. Based on the intercorrelation values Jurs-DPSA-3 and Jurs-FNSA-3 were omitted. The remaining 10 descriptors did not exhibit significant intercorrelation among themselves. We attempted a PLS run using the selected 10 descriptors.

The program SIMCA [32] was used for the partial least squares (PLS) analysis. PLS is a generalization of regression, which can handle data with numerous independent variables, possibly strongly correlated and/or noisy [25]. The linear PLS model finds 'new variables' (latent variables (LVs)) that are linear combinations of the original variables. To avoid overfitting, a strict test for the significance of each consecutive LV is necessary in which no new LVs are added when they become non-significant [25].

Though PLS model was constructed with all 10 selected descriptors, but subsequently, descriptors with smaller Variable Importance for the Projection (VIP) values were gradually deleted until a model with the best leave-one-seventh-out cross-validation correlation coefficient, $Q^2_{(1/7)}$, was obtained. Then final PLS model (model 19) was also run by the program MINITAB [29], which calculates the leave-one-out correlation coefficient, LOO-Q^2. External predictivity was also judged with the test set of compounds by developed equation. Model 19 [Eq. (2)] is shown below.

$$\begin{aligned}\log\left(\frac{1}{LC_{50}}\right) &= 2.919 + 0.641(\log K_{o/w}) \\ &\quad +0.008(Jurs\text{-}PNSA - 3)+6.22(Jurs\text{-}FPSA - 3)\end{aligned}$$

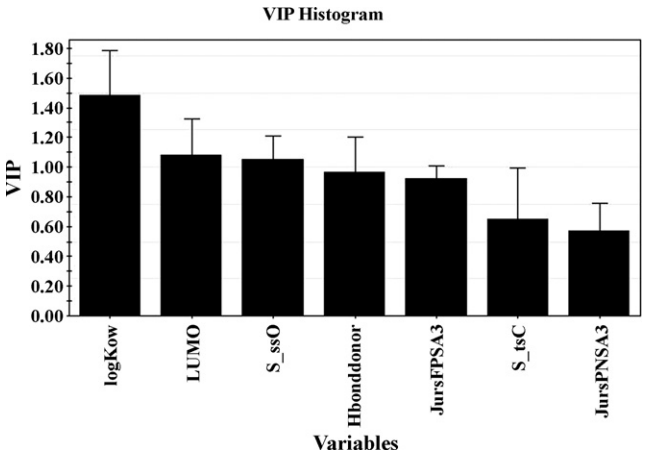

Fig. 2. Histogram of VIPs of the descriptors used in the final PLS model (model 19).

$$-0.281(\text{LUMO}) + 0.41\,\text{Hbonddonor} + 0.473(\text{S\_tsC}) + 0.118(\text{S\_ssO})$$

$$n_{\text{training}} = 222,$$

$$R^2 = 0.695,\quad Q^2_{\text{LOO}} = 0.678,\quad Q^2_{(1/7)} = 0.634,$$

$$r^2_{m(\text{LOO})} = 0.67,\quad n_{\text{test}} = 75,\quad R^2_{\text{pred}} = 0.741,$$

$$r^2_{m(\text{test})} = 0.707,\quad r^2_{m(\text{overall})} = 0.688 \qquad (2)$$

Eq. (2) involving only 7 descriptors and three latent variables (LVs) could predict 67.8% of the variance (leave-one-out predicted variance). The predicted R^2 (R^2_{pred}) value of 0.741 and $r^2_{m(\text{test})}$ value of 0.707 of model 19 outperformed the previous best model (model 18). The corrected R^2_{pred} value, calculated using the test set mean [40], for model 19 is 0.735. Also, based on $r^2_{m(\text{overall})}$ value of 0.688, model 19 outperformed the model 18. So, based on external and overall predictivity model 19 is more superior to model 18. Statistical quality of model 19 is presented in Table 2. The values of the descriptors appearing in Eq. (2) are given in Table S4 in Supplementary Materials section.

The calculated toxicity of all the chemicals obtained from the reported Eq. (2) is given in Table S1 in Supplementary Materials section, which shows that the calculated toxicity values are quite close to the observed ones.

The VIPs and coefficients of the original descriptors are presented as histogram in Figs. 2 and 3, respectively. Hydrophobicity expressed by $\log K_{o/w}$ and electrophilicity depicted by LUMO proved to be important and favourable features with positive and negative contribution, respectively. These 2 descriptors exerted significant contributions to the model with VIP values of 1.49 and 1.09, respectively. All of the remaining descriptors, i.e., H-bond donor (structural descriptor), S_ssO, S_tsC (topological descriptor) and Jurs-PNSA-3 and Jurs-FPSA-3 (spatial descriptor) have positive coefficients. According to the VIP values of the descriptors used in Eq. (2), the descriptors show the following order of importance:

$\log K_{o/w}$ > LUMO > S_ssO > H-bond donor > Jurs-FPSA-3 > S_tsC > Jurs-PNSA-3

Here, we explain the importance of each descriptor with suitable examples:

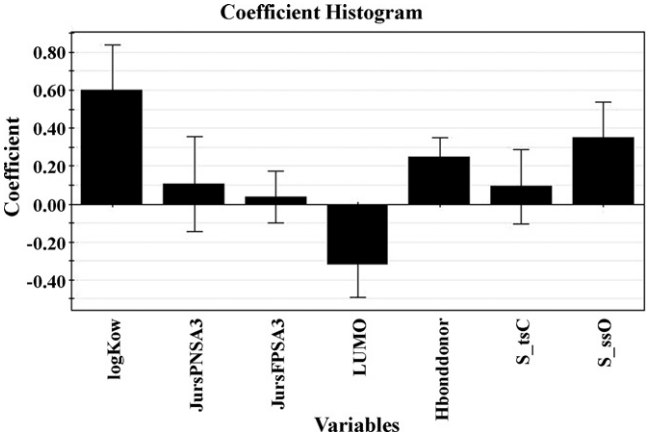

Fig. 3. Histogram of coefficients of the original descriptors used in the final PLS model (model 19).

(1) According to the VIP values, $\log K_{o/w}$ appeared as the most significant descriptor for the best model. Thus we can infer that the partition coefficient ($\log K_{o/w}$) is the most important descriptor for the toxicity. The developed model suggests that higher lipophilicity value influences the toxicity. Compounds like 2,2′,4,4′,5,5′-hexachloro-1,1′-biphenyl (**248**), 2,4,5,2′,5′-PCB (**252**) and 2,2′,3,3′,4,4′-PCB (**253**) showed toxicity values in higher range (8.44, 7.51 and 8.78 respectively) just due to the high values of $\log K_{o/w}$ (7.62, 6.98 and 7.62 respectively). On the other hand, compounds like *N*,*N*-dimethylformamide (**19**) and triethylene glycol (**295**) have $\log K_{o/w}$ values in the lower range (−0.93 and −1.75 respectively) and as a result, the corresponding toxicity values are very low (0.7 and 0.46 respectively).

(2) LUMO is the energy of lowest unoccupied molecular orbital. This represents the electrophilicity of a molecule. LUMO has unfavorable contribution towards the toxicity value as evidenced by the negative regression coefficient. This characteristic is important in governing the chemical reactivity and properties. Soft electrophiles are associated with relatively low, or a negative, LUMO energy [45]. The negative coefficient of LUMO energy in Eq. (2) suggests that soft electrophilc compounds are more toxic. Compounds like dichlorvos (**17**), carbon disulfide (**25**) and 2,4,6-trinitrotoluene (**125**) have the low values of LUMO (−0.70, −0.98 and −1.01 respectively; hence, these compounds are highly electrophilic) resulting in considerable toxicity values (9.1, 4.56 and 4.39 respectively) though the values of other descriptors for these compounds are moderate.

(3) Among remaining 5 descriptors, two spatial descriptors combine the shape and electronic information characterizing the molecules and thus encode features responsible for polar interactions for the toxicity.

(i) Jurs-PNSA-3 is the sum of the products of atomic solvent-accessible surface areas and partial charges $\bar{q}_a$ over all negatively charged atoms, i.e.,

$$\text{PNSA}_3 = \sum\nolimits_{a-} q_a^- SA_a^-$$

(ii) Jurs-FPSA-3 descriptor indicates the fractional charged partial positive surface area and is defined as the atomic charge weighted positive surface area (PPSA_3) divided by the total molecular solvent-accessible surface area (SASA), i.e.,

$$\text{FPSA}_3 = \frac{\text{PPSA}_3}{\text{SASA}}$$

Though both the descriptors have positive coefficients but the values of Jurs-PNSA-3 and Jurs-FPSA-3 are negative and positive respectively. So, eventually Jurs-PNSA-3 exerts negative impact and Jurs-FPSA-3 exerts positive impact on the model quality. As a result we can conclude that more is the negatively charged surface area, less is the toxicity. Compounds like 4-methoxybenzenamine (**89**), *m*-toluidine (**108**) and carbendazim (**234**) having comparatively less

Table 3
Comparison of quality of different models on *Daphnia* toxicity.

Toxicity endpoint	Equation statistics	Reference
D. magna 48-h; EC_{50} (mol/L); (non-polar narcosis)	$\log EC_{50} = -0.95(\log K_{o/w}) - 1.32$; $n = 49$, $R^2 = 0.95$, $q^2 = 0.94$, $s = 0.34$	[46]
D. magna 48-h; EC_{50} (mol/L); (polar narcotics)	$\log(1/EC_{50}) = 0.78(\log P) + 1.37$; $n = 21$, $R^2 = 0.896$, $s = 0.26$	[47]
D. magna 48-h; EC_{50} (mol/L); (immobilization)	$\log EC_{50} = -0.321(N_{OH})^2 - 0.869(\log P_{o/w}) - 0.494$; $n = 19$, $R^2 = 0.919$, $s = 0.270$, $F = 104$	[14]
D. magna 96-h; LC_{50} (mmol/L)	$\log(1/C) = -226.250 + 227.538[{}^{O}X_{CW}(a_{k,}({}^{O}EC_kP2_k),CC]$; $n = 220$, $R^2 = 0.782$, $s = 0.849$, $F = 783$ (training set), $n = 42$, $R^2 = 0.7388$, $s = 0.941$, $F = 113$ (test set)	[48]
D. magna 48-h; LC_{50} (mol/L)	$\log LC_{50} = -0.748(\pm 0.030)\log K_{o/w} - 2.393(\pm 0.101)$; $n = 193$, $R^2 = 0.76$, SE = 0.76, $F = 614$	[18]
D. magna 48-h; LC_{50} (mol/L)	Eq. (2) of this paper.; $n_{training} = 222$, $R^2 = 0.695$, $Q^2 = 0.678$ (training set); $n_{test} = 75$, $R^2_{pred} = 0.741$ (test set)	Present study

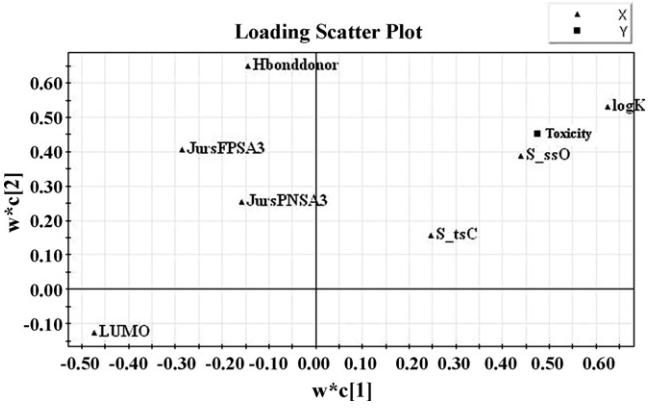

Fig. 4. The loading plot of the first two principal components (model 19).

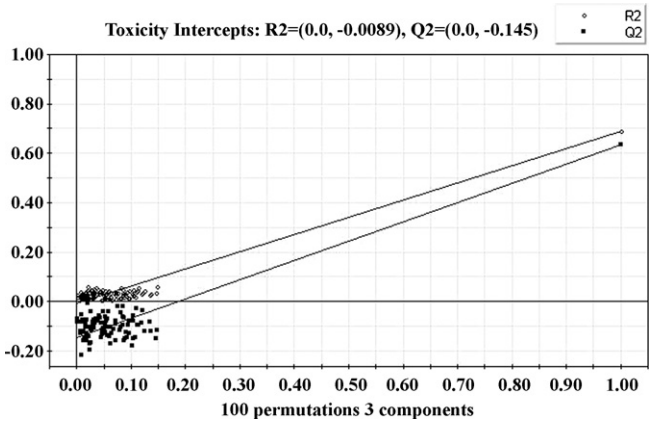

Fig. 5. Validation of the final PLS model (model 19). Validation plot based on 100 randomization cycles. R_Y refers to the correlation coefficient of the Y vector (toxicity) itself.

negative values of Jurs-PNSA-3 and comparatively higher positive values of Jurs-FPSA-3 show high toxicity values (5.57, 5.17 and 5.54 respectively).

(4) Ether linkage expressed by E-state index of fragment –O– (S_ssO) positively influences the toxicity of chemicals. Compounds like malathion (**132**), diazinon (**157**) and TEDP (**226**) with moderate lipophilicity values (2.29, 3.86 and 3.98 respectively) and low to moderate values of other descriptors show higher toxicity values (7.36, 8.45 and 9.15 respectively) just due to higher values of the S_ssO descriptor.

(5) Acetylenic carbon expressed by E-state index of fragment –C≡ (S_tsC) is responsible for greater toxicity of chemicals. Compounds like thiocyanic acid methylene ester (**230**) and iodopropynyl butylcarbamate (**258**) with very low values of other descriptors have high toxicity values (6.25 and 6.85 respectively) due to the corresponding high values of the S_tsC descriptor.

(6) H-bond donor descriptor signifies the number of hydrogen bond donor groups. Chemicals containing larger number of hydrogen bond donor groups show greater toxicity. Compounds like thiourea (**16**), methylthiourea (**173**) and dithiothreitol (**224**) have very low lipophilicity values, mod-

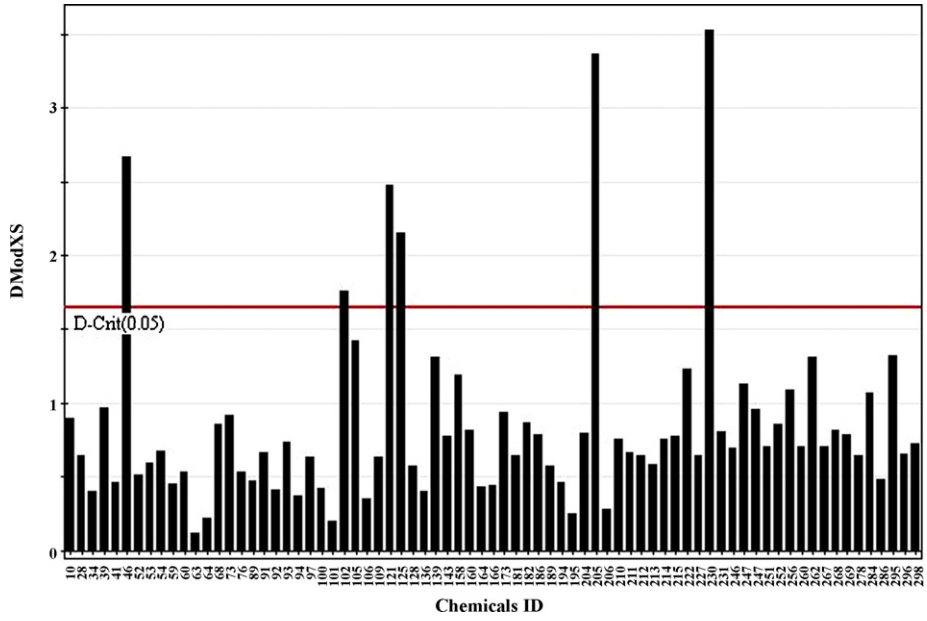

Fig. 6. DModX values of the 75 test set compounds at 95% level for model 19. The thick horizontal line signifies the critical DModX value (1.646) at the 95% confidence level.

erate electrophilicity values, less negatively charged surface areas, no ether linkage and no acetylenic carbon. However, due to the corresponding high number of hydrogen bond donor features (4, 3 and 4 respectively), these compounds show moderate toxicity values.

The PLS loading plot for the response variable (toxicity) and the descriptors included in the final model are shown in Fig. 4. The toxicity is explained to an extent of 46% by the first PLS component and 44% by the second component. LUMO and H-bond donor majorly contribute to the features of the first component and second component respectively. Log $k_{o/w}$, S_ssO, S_tsC, Jurs-PNSA-3 and Jurs-FPSA-3 share the features of both components.

Analysing Fig. 4 we can infer that electronic characteristics along with lipophilicity are playing dominant role for the toxicity of chemicals against *D. magna*. As 2D and 3D descriptors increase predictability of the models when used along with log $K_{o/w}$, the electronic, spatial, structural and topological factors are also found to be important for the toxicity along with hydrophobicity. It is interesting to note that model 19 shows better internal, external and overall validation characteristics than those of the models developed by von der Ohe et al. [18]. Table 3 shows a comparison of the statistical quality of Eq. (2) with that of other reported models on daphnia toxicity [46,47,14,48].

Model 19 was validated using a randomization test through randomly reordering (100 permutations) response data (default [32] is 20) using SIMCA 10.0 [30]. In this test, the toxicity data (Y) are randomly permuted keeping the descriptor matrix intact, followed by a PLS run. Each randomization and subsequent PLS analysis generates a new set of R^2 and Q^2 values, which are plotted against the correlation coefficient between the original Y values and the permuted Y values. R^2 and Q^2 were plotted against the correlation coefficient of the Y vector itself (R_Y) yielding intercepts close and below zero, respectively, indicating robustness of the model (Fig. 5). A model is considered valid if $R^2_{int} < 0.4$ and $Q^2_{int} < 0.05$. Model 19 is well above the permissible limit indicating that the model is not obtained by chance. Toxicity intercepts values are $R^2 = 0.0, -0.0089$, $Q^2 = 0.0, -0.145$.

Fig. 6 represents the residual SD of X-residuals (DModX) of test set compounds for model 19 (Eq. (2)). At 95% confidence level, DModX values of six test compounds are above the critical value of 1.646. These compounds are 2,4,6-trinitrophenol, endosulfan, 2,4,6-trinitrotoluene, chlorothalonil, thiocyanic acid methylene ester and acrylonitrile. So, these six test compounds are outside of the AD of model 19 and their predictions are less reliable.

4. Conclusion

An adequate, global and robust PLS model was established for 297 structurally diverse chemicals, providing an informative illustration of the contributing molecular, physicochemical, electronic, spatial properties and structural fragments which are responsible for the greater toxicity of the diverse organic chemicals. Above results suggest that higher lipophilicity and electrophilicity values significantly increase the toxicity. Furthermore, presence of ether linkage, hydrogen bond donor features and acetylenic carbon and less negatively charged surface areas contribute to the toxicity. As the compounds of the data set are chemically diverse, they also expectedly show diversity in their mechanisms of toxic actions (narcotics, reactive chemicals, oxidative phosphorylation uncouplers, electrophiles, proelectrophiles, etc.) as discussed in detail in Ref. [45]. The success of the present study is to develop a global PLS model (involving 7 descriptors and three latent variables) applicable for diverse classes of chemicals, and efficiency of the model in predicting daphnia toxicity of new chemicals has been adequately validated. The QSAR model described in the present paper for diverse organic chemicals may be useful for ecotoxicological hazard assessment against *D. magna* and environmental fate estimation for toxic chemicals in cases such information is not available in the existing toxicological databases.

Acknowledgement

Financial assistance from the Ministry of Human Resource Development, Govt. of India, New Delhi in the form of a scholarship to SK is thankfully acknowledged.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jhazmat.2009.12.038.

References

[1] D. Mackay, J. Hubbarde, E. Webster, The role of QSARs and fate models in chemical hazard and risk assessment, QSAR Comb. Sci. 22 (2003) 106–112.
[2] European Union (EU), 2001. White Paper: Strategy for a Future Chemicals Policy. Commission of the European Committees, Brussels, Belgium, COM (2001) 88 final, pp. 1–32.
[3] European Commission, Directive 2006/121/EC of the European Parliament and of the Council of 18 December 2006 amending Council Directive 67/548/EEC on the approximation of laws, regulations and administrative provisions relating to the classification, packaging and labelling of dangerous substances in order to adapt it to Regulation (EC) No 1907/2006 concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) and establishing a European Chemicals Agency. Off. J. Eur. Union (2006), L 396/850 of 30.12.2006, Office for Official Publications of the European Communities (OPOCE), Luxembourg.
[4] I. Lessigiarska, A.P. Worth, T.I. Netzeva, J.C. Dearden, M.T.D. Cronin, Quantitative structure–activity–activity and quantitative structure–activity investigations of human and rodent toxicity, Chemosphere 65 (2006) 1878–1887.
[5] L.G. Valerio Jr., K.B. Arvidson, R.F. Chanderbhan, J.F. Contrera, Prediction of rodent carcinogenic potential of naturally occurring chemicals in the human diet using high-throughput QSAR predictive modelling, Toxicol. Appl. Pharmacol. 222 (2007) 1–16.
[6] A.P. Worth, A. Bassan, J. DeBruijn, A. Gallegos-Saliner, G. Netzeva, G. Patlewicz, M. Pavan, I. Tsakovska, S. Eisenreich, The role of the European chemicals bureau in promoting the regulatory use of (Q)SAR methods, SAR QSAR Environ. Res. 18 (2007) 111–125.
[7] Organization for Economic Cooperation and Development (OECD), 2007. Guidance Document on the Validation of (Quantitative) Structure–activity Relationships [(Q)SAR] Models, ENV/JM/MONO 2 (2007) 1-154.
[8] Y.K. Koleva, J.C. Madden, M.T.D. Cronin, Formation of categories from structure–activity relationships to allow read-across for risk assessment: toxicity of α,β-unsaturated carbonyl compounds, Chem. Res. Toxicol. 21 (2008) 2300–2312.
[9] V.E. Kuz'min, E.N. Muratov, A.G. Artemenko, L. Gorb, M. Qasim, J. Leszczynski, The effect of nitroaromatics' composition on their toxicity in vivo: novel, efficient non-additive 1D QSAR analysis, Chemosphere 72 (2008) 1373–1380.
[10] R. Krieger (Ed.), Handbook of Pesticide Toxicology, 2nd ed., Academic Press, San Diego, CA, 2001.
[11] A.W. Hayes, Principles and Methods of Toxicology, 4th ed., Taylor and Francis, Philadelphia, PA, 2001.
[12] C. Barata, P. Alañon, S. Gutierrez-Alonso, M.C. Riva, C. Fernández, J.V. Tarazona, A *Daphnia magna* feeding bioassay as a cost effective and ecological relevant sublethal toxicity test for environmental risk assessment of toxic effluents, Sci. Total Environ. 405 (2008) 78–86.
[13] J.H. Yim, K.W. Kim, S.D. Kim, Effect of hardness on acute toxicity of metal mixtures using *Daphnia magna*: prediction of acid mine drainage toxicity, J. Hazard. Mater. B 138 (2006) 16–21.
[14] Y. Kamaya, Y. Fukaya, K. Suzuki, Acute toxicity of benzoic acids to the crustacean *Daphnia magna*, Chemosphere 59 (2005) 255–261.
[15] J.W. Deneer, Toxicity of mixtures of pesticides in aquatic systems, Pest. Manag. Sci. 56 (6) (2000) 516–520.
[16] S. Tao, X. Xi, F. Xu, B. Li, J. Cao, R. Dawson, A fragment constant QSAR model for evaluating the EC50 values of organic chemicals to *Daphnia Magna*, Environ. Pollut. 116 (2002) 57–64.
[17] E. Zvinavashe, T. Du, T. Griff, H.H.J. van den Berg, A.E.M.F. Soffers, J. Vervoort, A.J. Murk, I.M.C.M. Rietjens, Quantitative structure–activity relationship modeling of the toxicity of organothiophosphate pesticides to Daphnia magna and Cyprinus carpio, Chemosphere 75 (2009) 1531–1538.
[18] P.C. von der Ohe, R. Ku¨ hne, R. Ebert, R. Altenburger, M. Liess, G. Schürmann, Structural alerts-A new classification model to discriminate excess toxicity from narcotic effect levels of organic compounds in the acute daphnid assay, Chem. Res. Toxicol. 18 (2005) 536–555.

[19] Cerius 2 Version 4.10 is a product of Accelrys Inc., San Diego,CA.
[20] B. Everitt, S. Landau, M. Leese, Cluster Analysis, Arnold, London, 2001.
[21] E.R. Dougherty, J. Barrera, M. Brun, S. Kim, R.M. Cesar, Y. Chen, M. Bittner, J.M. Trent, Inference from clustering with application to gene-expression microarrays, J. Comput. Biol. 9 (2002) 105–126.
[22] A.R. Johnson, W.D. Wichern, Applied Multivariate Statistical Analysis, 5th ed., Pearson, Delhi, 2005, pp. 668–730.
[23] R.B. Darlington, Regression and Linear Models, McGrawHill, New York, 1990.
[24] L. Eriksson, E. Johansson, N. Kettaneh-Wold, S. Wold, Multi- and Megavariate Data Analysis: Principles and Applications, Umetrics, Umea, 2001.
[25] S. Wold, PLS for multivariate linear modelling, in: H. van de Waterbeemd (Ed.), Chemometric Methods in Molecular Design, VCH, Weinheim, Germany, 1995, pp. 195–218.
[26] D. Rogers, A.J. Hopfinger, Application of genetic function approximation to quantitative structure activity relationships and quantitative structure property relationships, J. Chem. Inf. Comput Sci. 34 (1994) 854–866.
[27] Y. Fan, L.M. Shi, K.W. Kohn, Y. Pommier, J.N. Weinstein, Quantitative structure–antitumor activity relationships of camptothecin analogues: cluster analysis and genetic algorithm-based studies, J Med Chem. 44 (2001) 3254–3263.
[28] T. Kimura, K. Hasegawa, K. Funatsu, J. Chem. Inf. Comput. Sci 38 (1998) 276–282.
[29] MINITAB is a statistical software of Minitab Inc., USA, <http://www.minitab.com>.
[30] SPSS is a statistical software of SPSS Inc., Chicago, IL, <http://www.spss.com>.
[31] STATISTICA is a statistical software of STATSOFT Inc., USA, <http://www.statsoft.com/>.
[32] UMETRICS SIMCA-P 10.0, info@umetrics.com: www.umetrics.com, Umea, Sweden, 2002.
[33] P.P. Roy, J.T. Leonard, K. Roy, Exploring the impact of the size of training sets for the development of predictive QSAR models, Chemom. Intell. Lab. Sys. 90 (2008) 31–42.
[34] K. Roy, On some aspects of validation of predictive quantitative structure–activity relationship models, Expert Opin. Drug Discov. 2 (2007) 1567–1577.
[35] J.T. Leonard, K. Roy, On selection of training and test sets for the development of predictive QSAR models, QSAR Comb. Sci. 25 (2006) 235–251.
[36] K. Roy, A.S. Mandal, Development of linear and nonlinear predictive QSAR models and their external validation using molecular similarity principle for anti-HIV indolyl aryl sulfones, J. Enzyme Inhib. Med. Chem. 23 (6) (2008) 980–995.
[37] H. Kubinyi, F.A. Hamprecht, T. Mietzner, Three-dimensional quantitative similarity–activity relationships (3D QSiAR) from SEAL similarity matrices, J. Med. Chem. 41 (1998) 2553–2564.
[38] P.P. Roy, K. Roy, On some aspects of variable selection for partial least squares regression models, QSAR Comb. Sci. 27 (2008) 302–313.
[39] P.P. Roy, S. Paul, I. Mitra, K. Roy, On two novel parameters for validation of predictive QSAR models, Molecules 14 (2009) 1660–1701.
[40] R. Schürmann, J. Ebert, B. Chen, R. Wang, Kühne, External validation and prediction employing the predictive squared correlation coefficient s test set activity mean vs training set activity mean, J. Chem. Inf. Model. 48 (2008) 2140–2145.
[41] S. Wold, M. Sjostrom, L. Eriksson, Partial least squares projections to latent structures (PLS) in chemistry, in: P.v.R. Schleyer (Ed.), Encyclopedia of Comparative Chemistry, vol. 3, Wiley, Chichester, GB, 1998, p. 2006.
[42] http://en.wikipedia.org/wiki/Applicability_Domain (accessed on August 4, 2009).
[43] L. Zhang, H. Zhu, T. Oprea, A. Golbraikh, A. Tropsha, QSAR modeling of the blood–brain barrier permeability for diverse organic compounds, Pharm. Res. 25 (2008) 1902–1914.
[44] S. Wold, M. Sjostrom, L. Eriksson, PLS-regression: a basic tool of chemometrics, Chemom. Intell. Lab. Sys. 58 (2001) 109–130.
[45] T.I. Netzeva, M. Pavan, A.P. Worth, Review of (quantitative) structure–activity relationships for acute aquatic toxicity, QSAR Comb. Sci. 27 (2008) 77–90.
[46] H.J.M. Verhaar, W. Mulder, J.L.M. Hermens, 1995. Overview of Structure–activity Relationships for Environmental Endpoints. Part 1. General Outline and Procedure, in: J.L.M. Hermens (Ed.), QSARs for Ecotoxicity. Report prepared within the framework of the project "QSAR for Prediction of Fate and Effects of Chemicals in the Environment", an international project of the Environmental; Technologies RTD Programme (DGXII/D-1) of the European Commission under contract number EV5V-CT92-0211.
[47] G. Hodges, D.W. Roberts, S.J. Marshall, J.C. Dearden, The aquatic toxicity of anionic surfactants to Daphnia magna—a comparative QSAR study of linear alkylbenzene sulphonates and ester sulphonates, Chemosphere 63 (2006) 1443–1450.
[48] A.A. Toropov, E. Benfenati, QSAR models for Daphnia toxicity of pesticides based on combinations of topological parameters of molecular structures, Chemosphere 50 (2003) 403–408.