Journal of Environmental Science and Engineering A 1 (2012) 663-671
Formerly part of Journal of Environmental Science and Engineering, ISSN 1934-8932

Inhibition of Microbial Growth by Anilines: A QSAR Study

Ahmed Bouaoune, Leila Lourici, Hamza Haddag and Djelloul Messadi
Environmental and Food Safety Laboratory, Badji Mokhtar University, Annaba 23000, Algeria

Received: November 3, 2011 / Accepted: January 9, 2012 / Published: May 20, 2012.

Abstract: The relative toxicity of 48 anilines using the *Tetrahymena pyriformis* population growth characteristics IGC_{50} (concentration causing 50% growth inhibition), available in the literature, was studied. At first, the entire data set was randomly split into a training set (31 chemicals) used to establish the QSAR model, and a test set (17 chemicals) for statistical external validation. A biparametric model was developed using, as independent variables, 3D theoretical descriptors derived from DRAGON software. The GA-MLR (genetic algorithm variable subset selection) procedure was performed on the trainingset by the software mobydigs using the OLS (ordinary least squares) regression method, and GA(genetic algorithm)-VSS(variable subset selection) by maximising the cross-validated explained variance (Q^2_{LOO}). The obtained model was examined for robustness (Q^2_{LOO} cross-validation, Y-scrambling) and predictive ability through both internal (Q^2_{LMO} , bootstrap) and external validation (Q^2_{ext}) methods. Descriptors included in the QSAR model indicated that $\log IGC_{50}^{-1}$ value was related to molecular size and shape, and interaction of molecule with its surrounding medium or its target. Moreover, the applicability domain of the model was discussed.

Key words: Toxic agents, growth of microbial species, QSAR hybrid model, statistical external validation, applicability domain.

1. Introduction

Recently computational methods have been used to solve complex problems in many aspects of science. One particularly useful method—the development of QSARs (quantitative structure-activity relationships) has found application in environmental chemistry and ecotoxicology [1-5].

QSAR approach systematization which has to be associated to the work of Hansch and Fujita in 1964 [6] is based on the assumption that the structure of a molecule must contain the features responsible for its physical, chemical and biological properties and on the possibility of representing a molecule by numerical descriptors.

The underlying hypothesis for QSAR models is that all molecules interact with the receptor in same or similar mode of action [7].

Corresponding author: Leila Lourici, Ph.D., main research field: environmental chemistry. E-mail: leilalourici@yahoo.fr.

The descriptors most used in the early QSAR analyses are the octanol/water partition coefficient (log*P*), the Hammett δ constant [8, 9] acting as an electronic effect descriptor and the lipophilicity parameter π, which is defined by analogy to the electronic descriptor. Together with these empirical descriptors, the classical models employ other physical-chemical properties as parameters; some of them derived from quantum chemical calculations, namely: partial charges, HOMO/LUMO energies, etc..

An important topic in environmental chemistry and ecotoxicology consists of the effect of toxic agents on the growth of microbial species. The population growth of protozoa, in particuliar of ciliates, in varied concentrations of toxic substances has been assessed by comparing a number of specific experimental values, including population density [10], growth rates [11], growth curves [12] and number of generations [13].

In all cases, the most tested species has been *Tetrahymena pyriformis*, a common freshwater hymenostome ciliate, which approximatively measures 50 μm in length and 30 μm in width [14]. Modern electronic equipment allows the easy determination of the population growth inhibition, providing a large collection of data for toxical research.

Schultz et al. [15] evaluated the relative toxicity of 48 selected anilines using the *Tetrahymena pyriformis* population growth characteristics IGC_{50} (concentration causing 50% growth inhibition) as an endpoint. The authors shown that simple $pIGC_{50}^{-1}$ (= $\log IGC_{50}^{-1}$) versus log*P* correlation can model environmental toxicity. The predictability of this log*P* dependent QSAR can be improved with the addition of $\sum\delta$ (the summation of the substituent electronic parameter δ), as a second and orthogonal descriptor. The statistical parameters reported by the mentioned authors are only related to the fitting performances.

In 1988, QSAR techniques suffered a great transformation due to the introduction of the so-called three dimensional (3D) molecular parameters, which accounted for the influence of different conformers, stereoisomers or enantiomers.

Several principles for assessing the validity of QSARs were proposed in 2002, as the "Setubal Principles" [16]; these were then modified in 2004 as the OECD (Organisation for Economic Co-operation and Development) Principles for QSAR validation [17]. To facilitate the consideration of a QSAR model for regulatory purpose, it should be associated with the following information: (a) a defined endpoint; (b) an unambiguous algorithm; (c) a defined applicability domain (AD); (d) appropriate measures of goodness of fit, robustness and predictivity; and (e) a mechanistic interpretation, if possible. Thus, further QSAR development on anilines should follow these guidelines.

In this study a biparametric model for the toxicity of aniline derivatives was developed using, as independent variables, 3D theoretical descriptors calculated from the chemical structure alone (Geometrical and GETAWAY descriptors). The available data set (taken from Schultz et al. [15]) was randomly split into training set (31 objects), used to develop the QSAR model, and a validation set (17 objects), used only for statistical external validation.

The model was examined for robustness and predictive ability through both internal and external validation methods. Finally, the QSAR applicability domain was discussed by the Williams plot of standardized residuals versus leverage values [18, 19].

2. Methodology

2.1 Descriptors Generation

The structures of the molecules were drawn using Hyperchem 6.03 software [20]. The final geometries were obtained with the semi empirical method AM1. All calculation were carried out at the RHF (restricted Hartree–Fock) level with non configuration interaction. The molecular structures were optimized using the algorithm Polak-Ribiere and a gradient norm limit of 0.001 kcal/Å. The resulted geometry was transferred into the software Dragon version 5.3 [21] to calculate 271 descriptors of the type Geometrical and GETAWAY (Geometry, Topology and Atoms Weighted AssemblY). Descriptors with constant or near constant values inside each group were discarded. For each pair of correlated descriptors (with correlation coefficient $r \geq 0.95$), the one showing the highest pair correlation with the other descriptors was excluded.

The GA (Genetic Algorithm) [22] has been considered superior to other methods of variable selection techniques. So, variable selection was performed on the training set, using GA in the MobyDigs version of Todeschini [23] by maximizing the cross-validated explained variance Q_{LOO}^2.

2.2 Model Development and Validation

Multiple linear regression analysis and variable selection were performed by package MobyDigs for windows/PC [23], using OLS (ordinary least squares regression) method and, as previously indicated, GA-VSS (GA for variable subset selection).

The goodness of fit of the calculated model was assessed by means of the multiple determination coefficient, R^2 and the *SDEC* (standard deviation error in calculation) defined as :

$$SDEC = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{n}} \qquad (1)$$

Cross validation techniques allow the assement of internal predectivity (Q^2_{LMO} cross-validation, bootstrap) in addition to the robustness of the model (Q^2_{LOO} cross-validation, Y-scrambling).

Cross validation by the LOO (leave-one-out) procedure employs *n* training sets of *n*-1 objects in and predicting each excluded object in the test set. The cross validated explained Q^2_{LOO} is defined as:

$$Q^2_{LOO} = 1 - \frac{PRESS}{TSS} = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_{i/i}\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2} \qquad (2)$$

where y_i and $\bar{y}$ are, respectively, the measured, and averaged (over the entire data set) values of the dependent variable; $\hat{y}_{i/i}$ denotes the response of the *i-th* object estimated by using a model obtained without using the *i-th* object; the summations run over all compounds in the training set.

The *PRESS* (predictive residual sum of squares) measures the dispersion of the predicted values. It is used to define Q^2, and the *SDEP* (standard deviation error in prediction) .

$$SDEP = \sqrt{\frac{PRESS}{n}} \qquad (3)$$

A value Q^2 > 0.5 is generally regarded as a good result and Q^2 > 0.9 as excellent [18, 19].

However, studies have indicated that while Q^2 is a necessary condition for high predictive power in a model, its alone is not sufficient.

To avoid overestimating the predictive power of the model the leave-more out (LMO up to 50% of perturbation: LMO/50) procedure (repeated 8000 times in this study) was also performed. In a typical LMO validation, n objects of the data set are divided in G cancellation groups of equal size, $m_i\,(= n/G)$. Based on the value of n , G is generally selected between 2 and 10. A large number of models are developed with each of the $n - m_i$ objects in the training set and m_i objects in the validation set. For each corresponding model m_i objects are predicted and Q^2_{LMO} computed (as average value of the number of validation runs).

In order to evidence the existence of fortuitous correlations, the randomization test (Y-scrambling) [24] was adopted. This test consists of building a property vector whose components are the components of the actual property vector, but randomly permuted in their positions. This new activity vector is used as if it was really an experimental one, and a QSAR model is computed in the usual way. This process was repeated 300 times, in order to test the capacity factor of the model to extract actual structure/activity relationships.

By bootstrap validation technique, the original size of the data set (n) is preserved for the training set, by the selection of n objects with repetition; in this way the training set usually consists of repeated objects and the evaluation set of the object left out [25]. The model is calculated on the training set and responses are predicted on the evaluation set. All the squared differences between the true response and the predictive response of the objects of evaluation set are collected in *PRESS*. This procedure of building training sets and evaluation sets is repeated 5,000 times in this study, *PRESS* are summed and the average predictive power is calculated.

By using the selected model the values of the response for the test objects are calculated and the quality of these predictions is defined in terms of Q^2_{ext}, which is defined as:

$$Q^2_{ext} = 1 - \frac{\sum_{i=1}^{n_{ext}} \left(\hat{y}_{i/i} - y_i\right)^2 / n_{ext}}{\sum_{i=1}^{n_{tr}} \left(y_i - \bar{y}\right)^2 / n_{tr}} = 1 - \frac{PRESS/n_{ext}}{TSS/n_{tr}} \quad (4)$$

where n_{ext} and n_{tr} are the number of objects in the external set (or left out by bootstrap), and the number of training set objects, respectively.

Other useful parameters are R^2, calculated for the validation chemicals by applying the model developped on the training set, and external standard deviation error of prediction ($SDEP_{ext}$), defined as:

$$SDEP_{ext} = \sqrt{\frac{1}{n_{ext}} \sum_{i=1}^{n_{ext}} \left(y_i - \hat{y}_i\right)^2} \quad (5)$$

where the sum runs over the test set objects (n_{ext}).

2.3 QSAR AD (Applicability Domain)

The AD was discussed by the Williams plot [18, 19] of jacknified residuals versus leverages (hat diagonal) values (h_i). The jacknifed residuals (or Studentized residuals) are the standardized cross-validated residuals. Each residuals is divided by its standard deviation, which is calculated without the *i*-th observation. The leverage (h_i) value of a chemical in the original variable space is defined as :

$$h_i = \boldsymbol{x_i}\left(\boldsymbol{X^T X}\right)^{-1}\boldsymbol{x_i^T} \quad (i = 1,\ldots,n) \quad (6)$$

where x_i is the descriptor row-vector of the query compound, and $\boldsymbol{X}$ is the $n(p+1)$ matrix of p model parameter values for n training set compounds. The superscript *T* refers to the transpose of the matrix/vector.

The warning leverage value (h^*) is defined as $3(p+1)/n$. When h value of a compound is lower than h^*, the probability of accordance between predicted and actual values is as high as that for the compounds in the training set. A chemical with $h_i > h^*$ will reinforce the model if the chemical is in the training set. But such a chemical in the validation set and its predicted data may be unreliable. However, this chemical may not appear to be an outlier because its residual may be low. Thus the leverage and the jacknified residual should be combined for the characterization of the AD.

3. Results and Discussion

3.1 Development and Validation of QSAR Models

Application of the GA-VSS led to several good models for the prediction of $pIGC_{50}^{-1}$ based on different sets of molecular descriptors. The best two dimensional model was constructed using the radius of gyration ($RGyr$) and R maximal autocorrelation of lag 3 weighted by van der Waals atomic volume v ($R3v+$). All data concerning value of $RGyr$, $R3v+$ and biological activity are summarized in Table 1.

The equation of the optimal model can be written as:

$$\begin{aligned} pIGC_{50}^{-1} &= -3.602\,(\pm 0.174) + 1.439\,(\pm 0.069)\,RGyr \\ &+ 16.342\,(\pm 1.416)\,R3v+ \end{aligned} \quad (7)$$

All relevant statistical parameters are reported in Table 2.

Values of R^2 and R^2_{adj} attest the good fitting performances of the model which, moreover, is very highly significant (great value of the Fisher parameter *F*).

The model is robust, the difference between R^2 and Q^2 is small (1%). Fig. 1 shows a plot contrasting experimental and cross-validated $pIGC_{50}^{-1}$. The point dispersion is small, although in this case there are two points a little bit far away from the rest.

The model demonstrates a very good stability in internal validation (difference between Q^2_{LOO} and

Table 1 Values of *RGyr*, *R3v+* and inhibition of growth concentration $pIGC_{50}^{-1}$ for a set of 48 anilines. The first 17 chemicals are the test set.

Chemical	$pIGC_{50}^{-1}$	*RGyr*	*R3v+*	Chemical	$pIGC_{50}^{-1}$	*RGyr*	*R3v+*
4-hexylaniline	2.04	3.434	0.023	3,4-dimethylaniline	-0.29	2.132	0.029
2,3-chloro	1.02	2.169	0.087	3-ethyl	-0.12	2.204	0.021
4-methyl	-0.02	2.008	0.021	2-chloro	-0.09	1.982	0.041
2,4,6-trichloro	1.01	2.588	0.039	2,4-dimethyl	-0.30	2.133	0.022
2-bromo	0.46	2.013	0.056	2-ethyl	-0.25	2.107	0.023
4-butyl	1.05	2.998	0.022	3-fluoro	0.04	1.932	0.025
2-chloro-6-methyl	0.12	2.156	0.037	2-propyl	0.06	2.414	0.023
2-phenyl	0.86	2.680	0.019	3-chloro	0.09	2.111	0.042
3-iodo	0.61	2.158	0.071	2-isopropyl	0.10	2.190	0.024
3,4,5-trichloro	1.51	2.451	0.088	4-isopropyl	0.21	2.403	0.024
4-ethyl	0.04	2.286	0.021	2-chloro-5-methyl	0.20	2.245	0.037
3-chloro-4-methyl	0.45	2.208	0.049	4-octyl	2.34	3.914	0.022
5-chloro-2-methyl	0.51	2.300	0.03	2-iodo	0.35	1.981	0.070
2,6-dichloro	0.33	2.291	0.04	4-chloro-2-methyl	0.35	2.298	0.033
3-phenyl	0.78	2.815	0.021	3-chloro-2-methyl	0.45	2.153	0.044
2,5-dichloro	0.58	2.448	0.039	2,4-dichloro	0.56	2.390	0.040
3,5-dichloro	0.71	2.423	0.038	4-propyl	0.49	2.611	0.024
2-methyl	-0.55	1.892	0.024	3-bromo	0.52	2.179	0.058
3-methyl	-0.43	1.968	0.026	2,6-dichloro-3-methyl	0.69	2.407	0.041
2,6-dimethyl	-0.43	2.047	0.024	4-phenyl	0.95	2.904	0.022
3,5-dimethyl	-0.37	2.158	0.017	3,4-dichloro	1.14	2.281	0.089
2,5-dimethyl	-0.35	2.134	0.023	2,4,5-trichloro	1.30	2.577	0.086
2-fluoro	-0.31	1.846	0.025	2,3,4-trichloro	1.35	2.405	0.087
2-chloro-4-methyl	0.24	2.211	0.039	4-pentyl	1.67	3.246	0.022

Table 2 Statistical parameters of the developed model.

n_{tr}	n_{ext}	Q^2_{LOO}	R^2	$Q^2_{LMO/50}$	Q^2_{BOOT}	R^2_{adj}	Q^2_{ext}
31	17	93.85	94.99	92.34	92.48	94.64	92.13
SDEC		*SDEP*	*SDEP*$_{ext}$		*s*	*F*	
0.151		0.168	0.184		0.159	265.64	

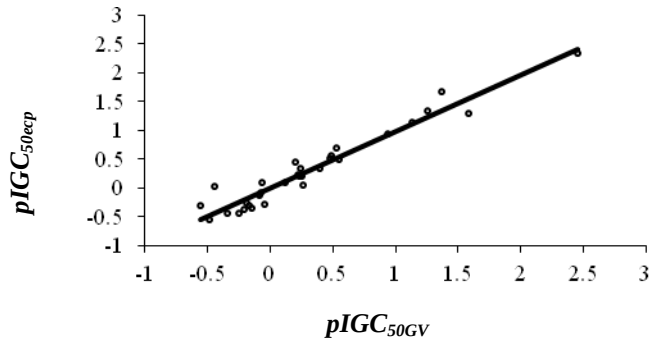

Fig. 1 Experimental ($pIGC_{50\,exp}$) versus cross-validated ($pIGC_{50\,cv}$) activity for the training set objects.

$Q^2_{LMO/50}$ is about 1%), while bootstrapping confirms the internal predictivity and stability of the model. $SDEP_{ext}$ is a little bit different from $SDEP$; the model works slightly worse in external prediction than in internal prediction. The model was also verified by Y-scrambling. Fig. 2 clearly ensures the existence of a linear relationship between $pIGC_{50}^{-1}$ and the descriptors $RGyr$ and $R3v+$. As can be observed the permuted responses yield poor predictive models, all having $Q^2 < 0.2$. On the other hand, the correctly ordered $pIGC_{50}^{-1}$ yield good statistical parameters, and therefore it is located isolated in the plot.

Using the same training set as before, a model was calculated by us on the molecular descriptors selected by Schultz et al. [15]. It follows the expression :

$$pIGC_{50}^{-1} = -1.404(\pm 0.113) - 0.4848(\pm 0.123)\sum \delta + 0.727(\pm 0.047)\log K_{OW} \quad (8)$$

The corresponding fitting and prediction parameters reported in Table 3 show that the model presented in this paper is slightly better than the one based on the Schultz et al. [15] approach.

3.2 Mechanistic Interpretation

By interpreting the descriptors in the proposed model, it is possible to gain some insight into factors that are likely related to inhibition of microbial growth. Of the two descriptors, one is GETAWAY ($R3v+$) and one is Geometrical ($RGyr$).

R-GETAWAY descriptors which are represented by $Rk(w)$ were calculated as follows. The molecular influence matrix was denoted by $\boldsymbol{H}$ and resembled the leverage (or influence) matrix defined in regression diagnostics [26].

The value of $\boldsymbol{H}$ was calculated from the molecular matrix $\boldsymbol{M}$ ($\boldsymbol{M}$ has A rows corresponding to the Cartesian coordinates x, y, z of each atom in optimized molecular structure) as follows:

$$\boldsymbol{H} = \boldsymbol{M}\left(\boldsymbol{M}^{\boldsymbol{T}}\boldsymbol{M}\right)^{-1}\boldsymbol{M}^{\boldsymbol{T}} \quad (9)$$

where the superscript T refers to the transposed matrix. The maximal contributed to the autocorrelation at each lag represented by $Rk(w)+$ can be defined as:

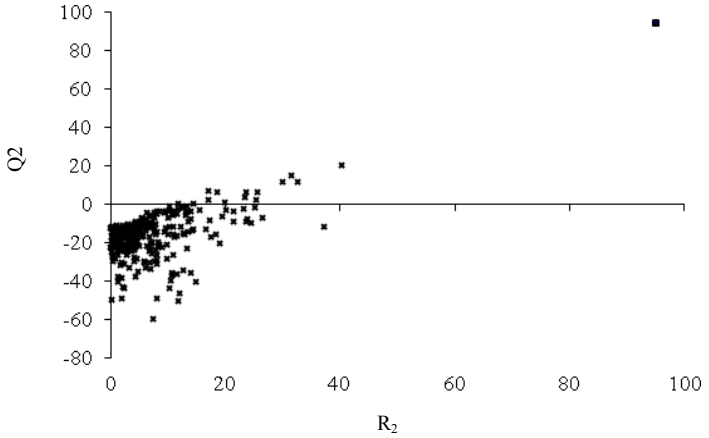

Fig. 2 Randomization test associated to the previous QSAR model. Crosses represent the randomly ordered activities, and the square corresponds to the real activities.

Table 3 Statistical parameters of the Schultz et al. [15] approach model.

n_{tr}	n_{ext}	Q^2_{LOO}	R^2	$Q^2_{LMO/50}$	Q^2_{BOOT}	R^2_{adj}	Q^2_{ext}
31	17	89.24	91.62	85.38	86.06	91.03	81.28
SDEC		*SDEP*	*SDEP*$_{ext}$		*s*	*F*	
0.196		0.222	0.293		0.206	153.1523	

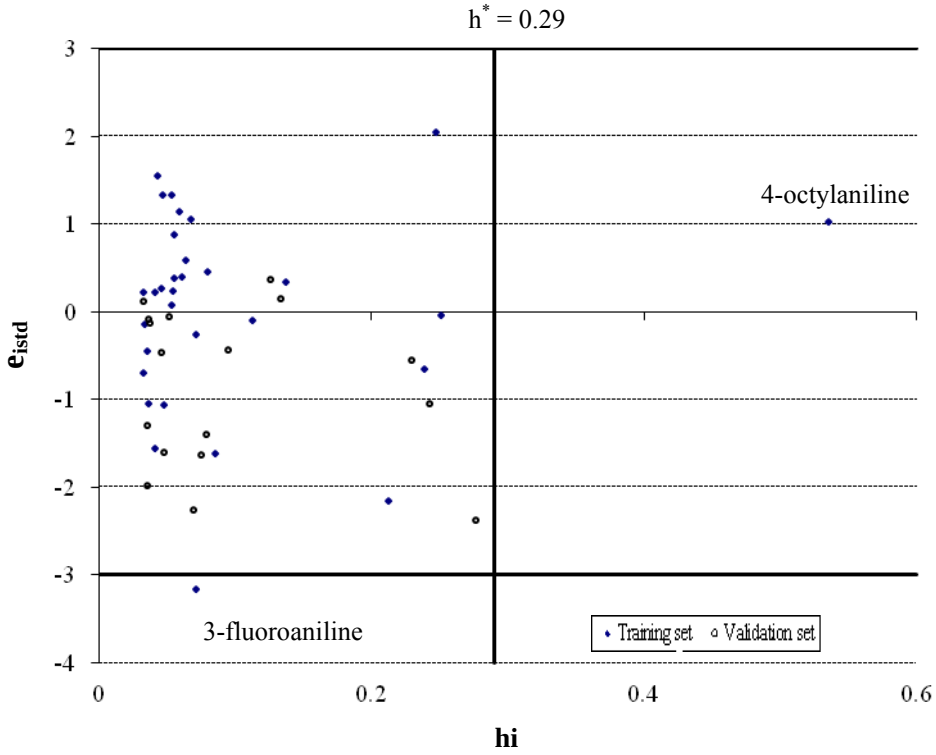

Fig. 3 Williams plot of the current QSAR model.

$$Rkw+ = \max_{ij}\left[\frac{\sqrt{h_{ii}h_{jj}}}{r_{ij}} w_i w_j \delta\left(k, d_{ij}\right)\right] i \neq j \text{ and}$$

$$k = 1,2,...,8 \qquad (10)$$

where $Rk(w)+$ is the w -weighted kth order maximal R index, r_{ij} is the 3D geometric distances between each pair of atoms i and j, d_{ij} is the topological diameter, h_{ii} and h_{jj} are diagnonal terms of the $\boldsymbol{H}$ matrix and δ is a Dirac delta function defined as:

$$\delta\left(k, d_{ij}\right) = \begin{cases} 1 \text{ if } d_{ij} = k \\ 0 \text{ if } d_{ij} \neq k \end{cases} \qquad (11)$$

$R3v+$ describes the size and the shape of the molecules. It is known that size, shape and symmetry of molecules play a key role in the process of distribution of molecule between two immiscible liquid phases. At the same time this descriptor indicates the role of the volume (v) in deciding the activity.

Mean size is a simple and very significant property of a molecule [27]. Easily obtained from light scattering experiments, a common measure of mean size such as the radius of gyration ($RGyr$) provides valuable information on interaction of molecule with its surrouding medium or its target.

3.3 Applicability Domain

As shown in the Williams plot (Fig. 3), the only high leverage chemical ($h_i > h^* = 0.29$) of the training set (4-octylaniline) is perfectly predicted, as normally happens for chemicals influential in training sets. Only one outlier is observed (3-fluoroaniline) which can be judged by its standardized residual greater than three standard deviation units (3σ).

4. Conclusion

A QSAR model on inhibition of microbial growth by anilines was developed using the OECD guidelines.

The available data set was randomly split into training and validation sets.

The QSAR model proposed in this paper is stable, robust, with good fitting and predictive performance. It is predictive for the chemicals used in the model development (internal validation on training chemicals) and also for chemicals not used in the model development (statistical external validation on validation set chemicals). The AD of the QSAR model was also described. The factors governing biological activities are the molecular size and shape, and interactions of molecule with its surrounding medium or its target.

References

[1] A.D. Deweese, T.W. Schultz, Structure-activity relationships for aquatic toxicity to *Tetrahymena*: Halogensubstituted aliphatic esters, Environmental Toxicology 16 (2001) 54-60.

[2] J.D. Leblond, B.M. Applegate, F.M. Menn, T.W. Schultz, G.S. Sayler, Structure -toxicity assessment of metabolites of the aerobic bacterial transformation of substituted naphtalenes, Environmental Toxicology and Chemistry 19 (2000) 1235-1246.

[3] A. Cotescu, M.V. Diudea, QSTR study on aquatic toxicity against poecilia reticulata and tetrahymena pyriformis using topological indices, Internet Electronic Journal of Molecular Design 5 (2006) 116-134.

[4] F. Li, J. Chen, Z. Wang, J. Li, X. Qia, Determination and prediction of xenoestrogens by recombinant yeast-based assay and QSAR, Chemosphere 74 (2009) 1152-1157.

[5] G.H. Lu, C. Wang, X.L. Guo, Prediction of toxicity of phenols and anilines to algae by quantitative structure-activity relationship, Biomededical and Environmental Sciences 21 (2008) 193-196.

[6] C. Hansch, T. Fujita, ρ-σ-π analysis: A method for the correlation of biological activity and chemical structure, Journal of the American Chemical Society 86 (1964) 1616-1626.

[7] M. Nendza, A. Wenzel, Discriminating toxicant classes by mode of action-1.(Eco) toxicity profiles, Environmental Science and Pollution Research 13 (2006) 192-203.

[8] L.P. Hammett, The effect of structures upon the reactions of organic compounds, Benzene derivatives, Journal of the American Chemical Society 59 (1937) 96-103.

[9] L.P. Hammett, Physical Organic Chemistry, Mc Graw Hill, New York, 1940.

[10] J.S. Roth, Certain effects of 2-aminofluorene and α- and β-naphthylamines on *Tetrahymena pyriformis*, Cancer Research 2 (1954) 346-350.

[11] N.R. Cooley, J.M. Keltner Jr, J. Forester, Polychlorinated biphenyls, aroclors 1248 and 1260: Effect on and accumulation by *tetrahymena pyriformis*, The Journal of Protozoology 20 (1975) 443-445.

[12] S. Apostol, A bioassay of toxicity using protozoa in the study of aquatic environment pollution and its prevention, Environmental Research 6 (1973) 365-372.

[13] D. Dive, H. Leclerc, Standardized test method using protozoa for measuring water pollutant toxicity, Progress in Water Technology 7 (1975) 67-72.

[14] D.L. Hill, The Biochemistry and Physiology of Tetrahymena, Academic Press, New York, 1972.

[15] T.W. Schultz, D.T. Lin, T.S. Wicke, L.M. Arnold, Quantitative structure-activity relationships for the *Tetrahymena pyriformis* population growth endpoint: A mechanism of action approach, in: W. Karcher, J. Devillers (Eds.), Practical Applications of Quantitative Structure-Activity Relationships (QSAR) in Environmental Chemistry and Toxicology, Kluwer Academic Publishers, Dordrescht, 1990, pp. 241-262

[16] J.S. Jaworska, M. Comber, C. Auer, C.J. Van Leeuwen, Summary of a workshop on regulatory acceptance of (Q)SAR for human health and environmental endpoints, Environmental Health Perspectives 111 (2003) 1358-1360.

[17] Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models [Online], Series on Testing and Assessment 69, OCDE's Environment Directorate, OECD Environment Health and Safety Publications, Mar. 30, 2007, http://www.oecd.org/data//33/37/37849783.pdf.

[18] L. Eriksson, J. Jaworska, A. Worth, M. Cronin, R.M. Mc Dowell, P. Gramatica, Methods for reliability, uncertainty assessment, and applicability evaluations of regression based and classification QSARs, Environmental Health Perspectives 111 (2003) 1361-1375.

[19] A. Tropsha, P. Gramatica, V.K. Grombar, The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models, QSAR and Combinatorial Science 22 (2003) 69-76.

[20] Hyperchem™ Release 7, Hypercube for Windows, Molecular Modeling System, 2000.

[21] R. Todeschini, V. Consonni, M. Pavan, DRAGON Software for the Calculation of Molecular Descriptors, Release 5.4 for Windows, Milano, 2006.

[22] R. Leardi, R. Boggia, M. Terrile, Genetic algorithms as a strategy for feature selection, Journal of Chemometrics

6 (1992) 267-281.

[23] R. Todeschini, D. Ballabio, V. Consonni, A. Mauri, M. Pavan, MOBYDIGS, version 1.1, Copyright TALETE srl, 2009.

[24] S. Wold, L. Eriksson, Statistical Validation of QSAR Results, Validation Tools in Chemometrics Methods in Molecular Design, VCH Publishers, New York, 1995, pp. 309-318.

[25] B. Efron, The Jacknife, the Bootstrap and Other Resampling Planes, Society for Industrial and Applied Mathematics, Philadelphia, 1994.

[26] V. Consonni, R. Todeschini, M. Pavan, Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors, 1—Theory of the novel 3D molecular descriptors, Journal of Chemical Information and Modeling 42 (2002) 682-692.

[27] G.A. Arteca, Analysis of shape transitions using molecular size descriptors associated with inner and outer regions of a polymer chain, Journal of Molecular Structure: THEOCHEM 630 (2003) 113-123.