Electronic Journal of Biotechnology ISSN: 0717-3458
Vol.11 No.3, Issue of July 15, 2008
© 2008 by Pontificia Universidad Católica de Valparaíso -- Chile
Received August 6, 2007 / Accepted January 4, 2008
DOI: 10.2225/vol11-issue3-fulltext-9
RESEARCH ARTICLE

Toxicity caused by para-substituted phenols on *Tetrahymena pyriformis*: The structure-activity relationships

Lorentz Jäntschi*
Department of Chemistry
Technical University of Cluj-Napoca
400641 Cluj-Napoca, Romania
Tel: 4 0264 401775
Fax: 4 0264 415054
E-mail: lori@chimie.utcluj.ro

Violeta Popescu
Department of Chemistry
Technical University of Cluj-Napoca
400641 Cluj-Napoca, Romania
Tel: 4 0264 401775
Fax: 4 0264 415054
E-mail: violeta@chimie.utcluj.ro

Sorana D. Bolboacă
Department of Medical Informatics and Biostatistics
"Iuliu Haţieganu" University of Medicine and Pharmacy
400349 Cluj-Napoca, Romania
Tel: 4 0264 431697
Fax: 4 0264 593847
Email: sbolboaca@umfcluj.ro

Financial support: The research supported by the UEFISCSU Romania through research grants ID_458 & ID_1051.

Keywords: para-substituted phenol derivatives, structure-activity relationships, *Tetrahymena pyriformis*, toxicity**.**

Abbreviations: MDF: Molecular Descriptors Family
MLR: Multiple Linear Regression
NN(s): Neural Network(s)
qSARs: quantitative Structure-Activity Relationships
SAR(s): structure-activity relationship(s)

The toxicity of thirty para-substituted phenols on *Tetrahymena pyriformis* was modelled using an original methodology that uses the complex structural information of the compounds. Two models were built. The methodology allows atomic properties to be assigned to toxicity based on the selection of pairs of descriptors from the entire family, which is called Molecular Descriptors Family (MDF). One model has two independent structural descriptors and the other has four. The model with four descriptors proved to have high estimated and predictive abilities (over 97% of toxicity could be explained by structural information). The partial charge distribution by bonds (molecular topology) and space (molecular geometry) interaction proved to be related with the toxicity of para-substituted phenols on *Tetrahymena pyriformis*. The predictive ability of the model was tested by using the following methods: the *cross-validation leave-one-out* and the *training versus test experiments*. The comparisons among the models were performed using the *correlated correlations* method. The embedding of the complex information from the structure using MDF methodology can lead to further investigations of the mechanism of chemicals toxicity on *Tetrahymena pyriformis*.

The development of information and computing technologies have led to the development of structure-activity/property relationships (qSARs) methods with focus on informatics and modelling (Diudea et al. 2001). The qSARs methods are used for the quantitative characterization of the relationships between the structure of compounds and their activity or property in many fields such as: drug design (Duch et al. 2007; Prathipati et al. 2007), environmental sciences (Li and Xi, 2007; Knauer et al. 2007; Jager et al. 2007), biotechnology (Li et al. 2007), and all the fields of chemistry (Niu et al. 2007; Malík et al. 2007; Scotti et al. 2007; Lubbers et al. 2007).

The toxicity of para-substituted phenols on *Tetrahymena Pyriformis* (a non-pathogenic unicellular protozoan) was studied by many researchers. The toxicity has been

*Corresponding author

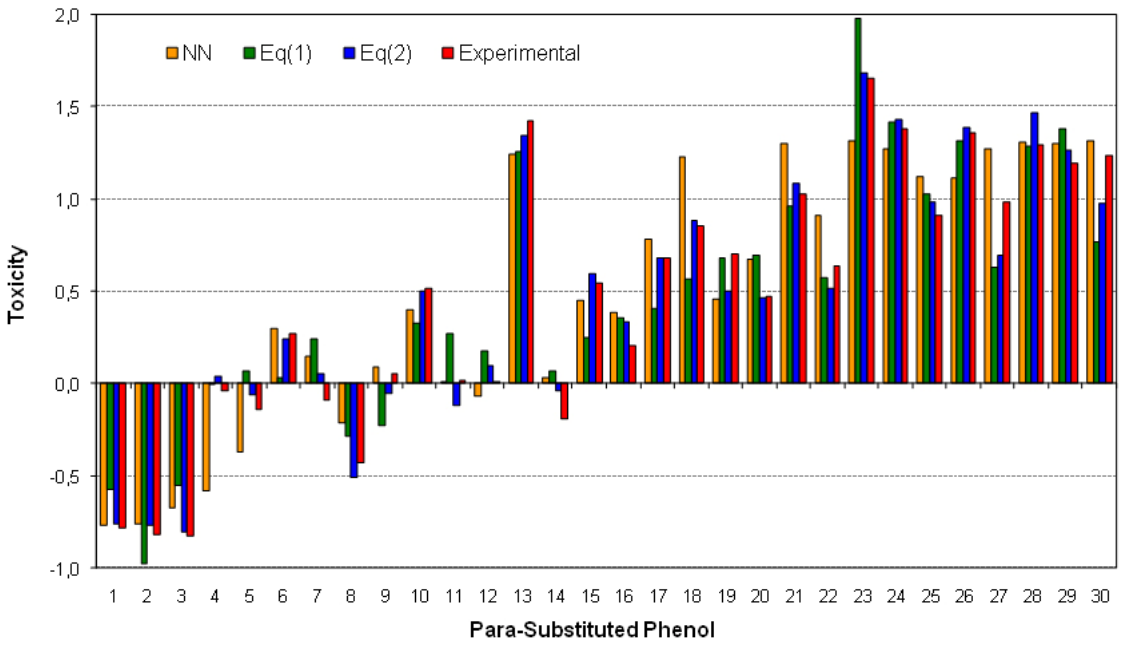

Figure 1. Experimental toxicity (Schultz, 1987b), toxicity estimated by neural-network (Ivanciuc, 1998), toxicity estimated by Eq(1) and Eq(2).

analyzed by using the octanol/water partition coefficient (Schultz, 1987a), the hydrophobicity/ionization surface (Schultz, 1987b; Schultz et al., 1996), electrophilicity (Roy et al. 2006). Different approaches have been used: quantitative neighbourhoods of atoms (Lagunin et al. 2007), core electron binding energy (Takahata et al. 2007), quantum topological molecular similarity (Loader et al. 2007), neural networks (Ivanciuc, 1998) or back propagation artificial neural networks (Yang et al. 2006).

The main objective of the present study was to characterize the toxicity caused by para-substituted phenols on *Tetrahymena pyriformis* by using the molecular descriptors family on the structure-activity relationships approach. This approach proved its estimated and predictive abilities on different classes of chemical compounds, both on properties and activities (Jäntschi and Bolboacă, 2007).

MATERIALS AND METHODS

A sample of thirty para-substituted phenols ($HO-C_6H_4-R$) was included into the study. The experimental toxicities on *Tetrahymena pyriformis* (Tox_{exp}), expressed as the logarithm of the inverse of the IGC (inhibitory growth concentration) value in mmol/l, were taken from a previously reported research (Schultz, 1987b).

The MDF on the SARs (Jäntschi, 2004; Jäntschi, 2005) was applied. This method included the following six steps:

Step 1: The topological (2D) and geometrical (3D) model of investigated para-substituted phenols was obtained using the HyperChem software. The geometry of the compounds was optimized by applying the semi-empirical Extended Hückel model (Hoffmann, 1963) and the quantum mechanics model (Cornell et al. 1995). The output files were stored as *.hin files.

Step 2: The experimental data were collected and were stored into a *.txt file.

Step 3: It includes the construction, generation, calculation and filtration of the molecular descriptors family. The *.hin files, which contain information about the topology, geometry and charges distribution for each para-substituted phenol, represented the primary data file required to construct, generate, and calculate the molecular descriptors family. A set of five PHP programs generated the MDF for para-substituted phenols: ▪ *0_mdf_prepare.php* creates the structure of tables for the investigated compounds; ▪ *1_mdf_generate.php* generates the MDF of the para-substituted phenols and stores them into a table; ▪ *2_mdf_linearize.php* applies the linearizing operator and stores valid records into tables; ▪ *3_mdf_bias.php* sorts the descriptors by squared correlation coefficient and deletes identical entries; ▪ *4_mdf_order.php* orders the descriptors from highest to lowest by the squared correlation coefficient again and creates a new table. The results are stored on a FreeBSD server from IntraNet [IP 172.27.211.5] using a MySQL database server.

Each molecular descriptor has a name consisting of seven-letters that describes the modality of its construction. The description of each possible character is presented in Table 1.

Step 4: It searches and identifies the most significant SAR models. The following criteria were used (Bolboacă and Jäntschi, 2007): the squared correlation coefficient (value closed to 1 indicates a good model), the standard error of estimated (value closed to 0 indicates a good model) and statistical parameters associated with the model (the Fisher parameter, which has a less than 5% probability of type I error, confidence intervals for the intercept and slope, standard error of intercept and slope, student parameter and its probability of type I error).

Step 5: The models were validated in order to characterize their estimated and predictive abilities. The leave-one-out cross-validation analysis (Baumann, 2003) was conducted (Leave-one-out Analysis, 2005). The obtained score (abbreviated as $r^2_{loo\text{-}cv}$), the standard error of predictive and the Fisher parameter were obtained and interpreted.

Step 6: The analysis of the models was performed by assessing the following: ▪ model stability (the model is considered more stable if the difference between the squared correlation coefficient and the cross-validation leave-one-out score is closer to 0) ▪ predictive ability of the model with the higher squared correlation coefficient was assessed in training and test experiments (Training vs. Test Experiment, 2005), ▪ comparison with previously reported models (where appropriate) through a correlated correlation analysis (Steiger, 1980). A difference between the squared correlation coefficient (r^2) and the leave-one-out cross-validation score ($r^2_{loo\text{-}cv}$) lower than 0.3 indicates the absence of an over fitted model, irrelevant independent variables, and/or outliers (Bolboacă and Jäntschi, 2007). Moreover, in order to identify the outliers in the investigated compounds, the graphical representation methods were used (Bolboacă and Jäntschi, 2007).

Note that the MDF SAR approach uses a genetic algorithm for selection of descriptors from descriptor's pool (Jäntschi et al. 2007).

RESULTS

By integrating the complex knowledge extracted from the structure of the studied para-substituted phenols, two SAR models were identified, one with two and the other with four descriptors:

$$\hat{Y}_{2v} = -2.261 + 0.037 \cdot ASMmVQt - 0.216 \cdot lfDdOQg \quad [1]$$

$$\hat{Y}_{4v} = -3.295 + 0.035 \cdot ASMmVQt - 0.326 \cdot lfDdOQg + 0.079 \cdot InMrLQg - 0.346 \cdot LsDMpQg \quad [2]$$

where: $\hat{Y}_{2v}$ = toxicity estimated by Eq(1); $\hat{Y}_{4v}$ = toxicity estimated by Eq(2); *ASMmVQt*, *lfDdOQg*, *InMrLQg*, and *LsDMpQg* = molecular descriptors.

The values of the experimental determinations (Tox_{exp}), of the calculated descriptors and of the toxicity estimated by Eq(1) and Eq(2) are presented in Table 2.

The values of the squared correlation coefficients between each descriptor and the experimental toxicity (Tox_{exp}) as well as between pairs of descriptors were as follows:

SAR model with two descriptors - Eq(1):

$$r^2(ASMmVQt, Tox_{exp}) = 0.2661$$

$$r^2(lfDdOQg, Tox_{exp}) = 0.3599 \quad [3]$$

$$r^2(ASMmVQt, lfDdOQg) = 0.12152$$

SAR model with four descriptors - Eq(2):

$r^2(ASMmVQt, Tox_{exp}) = 0.2661$; $r^2(lfDdOQg, Tox_{exp}) = 0.3599$

$r^2(InMrLQg, Tox_{exp}) = 0.4329$; $r^2(LsDMpQg, Tox_{exp}) = 0.0747$

$r^2(ASMmVQt, lfDdOQg) = 0.1215$; $r^2(ASMmVQt, InMrLQg) = 0.0057$ [4]

$r^2(ASMmVQt, LsDMpQg) = 0.1136$; $r^2(lfDdOQg, InMrLQg) = 0.1769$

$r^2(lfDdOQg, LsDMpQg) = 0.3159$; $r^2(InMrLQg, LsDMpQg) = 0.10368$

The statistics associated with the models with two - Eq(1) - and four - Eq(2) molecular descriptors are presented in Table 3.

The graphical representation of the relation among the estimated toxicity of para-substituted phenols on *Tetrahymena Pyriformis* by Eq(1), Eq(2), neural network (Ivanciuc, 1998) and experimental toxicity (Schultz, 1987b) is presented in Figure 1.

The statistics on the similarity of the activity estimated by Eq(1) ($\hat{Y}_{2v\text{-}Eq(1)}$) and by Eq(2) ($\hat{Y}_{4v\text{-}Eq(1)}$) as well as the experimental toxicity (Tox_{exp}) of para-substituted phenols are presented in Table 4. In Table 4 the best estimation values, expressed as the lowest value of the difference between experimental and estimated toxicity, are shaded in gray.

The validation results of the model with four descriptors in training versus test experiments (for the sample size that varied from 18 to 22 in training) are presented in Table 5.

The comparison between the SAR model with four descriptors and the previously reported MLR (Multiple Linear Regression, (Ivanciuc, 1998)) and Neural Network (NN, (Ivanciuc, 1998)) models is presented in Table 6.

DISCUSSION

The integration of the structural information obtained from the para-substituted phenol compounds allows the estimation and prediction of toxicity on *Tetrahymena pyriformis*. Two models proved to have good estimated and predictive abilities (one model with two (Eq(1)) and the other with four descriptors (Eq(2)).

The analysis of the results presented in Table 2 reveals the influence of the substituent on the toxicity of para-substituted phenols. Thus, the phenyl group determined a higher toxicity of para-substituted phenols (between 1.01237 for 4-hydroxybenzophenone - compound no. 21, Table 2, and 1.6547 for 4- hydroxybenzene - compound no. 23, Table 2). A high toxicity is also determined by the nitro group, as in the case of the 4-nitrophenol (1.4257, Table 2).

Both SAR models were statistically significant, the significance level being lower than 0.0001 (Table 3). In toxicity modelling, three descriptors refer to molecular geometry (*lfDdOQ**g***, *InMrLQ**g*** and *LsDMpQ**g***) and one refers to molecular topology (*ASMmVQ**t***). All descriptors consider the partial electric change as the atomic property (*ASMmV**Q**t*, *lfDdO**Q**g*, *InMrL**Q**g*, *LsDMp**Q**g*).

The values of the correlation coefficient obtained by the model with two descriptors (r = 0.9472, Table 3) sustain the role of these two descriptors in the estimation of toxicity. Almost ninety percent of the toxicity variation of the studied para-substituted phenols can be explained by its linear relationship with the *ASMmVQt* and the *lfDdOQg* descriptors. The prediction ability of the model with two variables is sustained by the results obtained in leave-one-out cross-validation analysis: leave-one-out cross-validation score ($r^2_{loo\text{-}cv}$ = 0.8745, Table 3), standard error of predicted (s_{loo} = 0.2613, Table 3), Fisher parameter and associated significance ($p_{pred} = 7.58\cdot10^{-13}$, Table 3). The analysis of the model with two variables showed that molecular descriptors are not able to provide individually relevant models (Eq(3)). Note also that there is no collinearity between the descriptors used by the model with two descriptors (r^2(*ASMmVQt*, *lfDdOQg*) = 0.12152). The model with two variables reveals that the toxicity of the studied para-substituted phenols on *Tetrahymena pyriformis* is of geometrical and topological nature and it is also dependent on partial electric changes.

Both descriptors used by the model with two descriptors are found again in the model with four descriptors (Eq(2)). Ninety-seven percent of toxicity variation of the para-substituted phenols could be explained by its linear relationship with the molecular descriptors used by this model. The value of the multiple correlation coefficient (r = 0.9868, Table 3) supports the estimated ability of the SAR model. The predictive ability of the model with four descriptors is supported by the following: the value of the leave-one-out cross-validation score ($r^2_{loo\text{-}cv}$ = 0.9650, Table 3), the type I error of the Fisher parameter ($p_{pred} = 1.50\cdot10^{-21}$, Table 3), the standard error of predicted (s_{loo} = 0.1429, Table 3) and the stability of the model ($r^2 - r^2_{loo\text{-}cv}$ = 0.0086, Table 3). No significant correlation was identified neither between the descriptor and the experimental toxicity nor between the pairs of descriptors (Eq(4)). The toxicity of the para-substituted phenols on *Tetrahymena pyriformis* is of geometrical and topological nature. It is also dependent on the partial electric charge of the compounds.

The analysis of the results presented in Table 4 indicates that the best proximity of the estimated and experimental toxicity was obtained by the SAR model with four variable (on twenty-one out of thirty compounds the estimated value was in the proximity of the experimental value), followed by the model with two variables (five compounds out of thirty obtained the best proximity) and the neural network (Ivanciuc, 1998) (four compounds out of thirty obtained the best proximity).

The predictive ability of the model with four descriptors was studied on training and test sets. With one exception, all investigated sample sizes obtained statistically significant models at a significance level of 1% (Table 5). The exception was observed in the experiment with twenty-one compounds in the training set and nine compounds in the test set. For this model the type I error was of $1.4\cdot10^{-2}$ and $1.6\cdot10^{-14}$, respectively. The average of the squared correlation coefficient obtained in training sets was almost identical with the average of the squared correlation coefficient in the test sets (0.971 vs. 0.972, Table 5). The dispersion of the correlation coefficients in both sets was low (see Table 5). The above mentioned results support the validity of the SAR model with four descriptors as well as its power of predicting the toxicity of para-substituted phenols. The molecular descriptors of a new para-substituted phenol could be calculated using the online DC Demo Calculator (DC Demo Calculator, 2005). Therefore the 2D and 3D structure of the compound has to be constructed using the HyperCem software. As result, the calculate values of the molecular descriptors are displayed. Moreover, the 2D and 3D structure of a new para-substituted phenol could be used in order to predict its activity (MDF SAR Predictor, 2005). The following steps must be followed: ▪ selecting the name of learning set (RRC443_ for the para-substituted phenols set); ▪ selecting the predictor equation (the model with two or four molecular descriptors); and ▪ browsing and submitting the *.hin file of the new compound proposed for investigation. Consequently, the equation used for prediction, the calculated values of the molecular descriptors family on the structure-activity relationships for the new compound as well as the activity predicted by the model are displayed.

The comparison between the SAR model with four descriptors and the previously reported models (Ivanciuc, 1998) (Table 6) showed that the probability of coincidence between the SAR model and the MLR model is of $1.14 \cdot 10^{-2}$, while that between the SAR model and the NN model is of $4.51 \cdot 10^{-2}$. It can be concluded that the correlation coefficient obtained by the SAR model with four descriptors is significantly higher compared with the correlation coefficients obtained by the previously reported models (Ivanciuc, 1998).

Many approaches have been developed in order to translate the chemical information of a compound into a useful numerical value (Todeschini and Consonni, 2000). The radial basis functions (Hemmer et al. 1999), GATEWAY (Consonni et al. 2002), 3-MoRSE electron diffraction (Todeschini and Consonni, 2000) and other descriptors represent similar approaches. These approaches are useful for further investigations if their application leads to significant statistical models. The difference between models in terms of structure-activity relationships could then be investigated using the correlated correlation analysis (Steiger, 1980).

The above-mentioned results support the estimated and predictive abilities of the SAR model with four descriptors to characterize the toxicity of para-substituted phenols on *Tetrahymena pyriformis*. In conclusion, the toxicity of the studied para-substituted phenols on *Tetrahymena pyriformis* is of both geometrical and topological nature and depends on the partial electric charges of the compounds. Furthermore, the application of the SAR method in the modelling of the para-substituted phenols toxicity on *Tetrahymena pyriformis* could be the first step in discovering and characterizing new compounds. Such further investigations could lead to the discovery of compounds with higher activity at lower costs.

REFERENCES

BAUMANN, K. Cross-validation as the objective function for variable-selection techniques. *Trends in Analytical Chemistry*, 2003, vol. 22, no. 6, p. 395-406.

BOLBOACĂ, S.D. and JÄNTSCHI, L. Modelling the property of compounds from structure: Statistical methods for models validation. *Environmental Chemistry Letters*, October 2007.

CONSONNI, V.; TODESCHINI, R. and PAVAN, M. Structure/Response Correlations and Similarity/Diversity Analysis by GETAWAY Descriptors. 1. Theory of the Novel 3D Molecular Descriptors. *Journal of Chemical Information and Computer Sciences*, 2002, vol. 42, no. 3, p. 682-692.

CORNELL, W.D.; CIEPLAK, P.; BAYLY C.I.; GOULD I.R.; MERZ, K.M. JR.; FERGUSON D.M.; SPELLMEYER D.C.; FOX, T.; CALDWELL J.M. and KOLLMAN, P.A. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *Journal of the American Chemical Society*, 1995, vol. 117, p. 5179-5197.

DC Demo Calculator [online]. ©2005, Virtual Library of Free Software [cited 20 November 2007]. Available from Internet: http://l.academicdirect.org/Chemistry/SARs/MDF_SARs/j_mdf_demo.php.

DIUDEA, M.; GUTMAN, I. and JÄNTSCHI L. *Molecular Topology*. Huntington, New York; Nova Science, 2001. 332 p. ISBN 1-56072-957-0.

DUCH, W.; SWAMINATHAN, K. and MELLER, J. Artificial intelligence approaches for rational drug design and discovery. *Current Pharmaceutical Design*, 2007, vol. 13, no. 14, p. 1497-1508.

HEMMER, M.C.; STEINHAUER, V. and GASTEIGER, J. Deriving the 3D structure of organic molecules from their infrared spectra. *Vibrational Spectroscopy*, 1999, vol. 19, p. 151-164.

HOFFMANN, R. An extended Hückel theory. I. Hydrocarbons. *Journal of Chemical Physics*, 1963, vol. 39, p. 1397-1412.

IVANCIUC, O. Artificial Neural Networks Applications. Part 4. Quantitative structure-activity relationships for the estimation of relative toxicity of phenols for Tetrahymena. *Revue Roumaine de Chimie*, 1998, vol. 43, no. 3, p. 255-260.

JAGER, T.; POSTHUMA, L.; de ZWART, D. and van de MEENT, D. Novel view on predicting acute toxicity: Decomposing toxicity data in species vulnerability and chemical potency. *Ecotoxicology and Environmental Safety*, 2007, vol. 67, no. 3, p. 311-322.

JÄNTSCHI, L.; KATONA, G. and DIUDEA, M. Modeling molecular properties by Cluj indices. *MATCH Communications in Mathematical and in Computer Chemistry*, 2000, vol. 41, p. 151-188.

JÄNTSCHI, L. MDF - A New QSPR/QSAR Molecular Descriptors Family. *Leonardo Journal of Sciences*, 2004, vol. 4, no. 3, p. 68-85.

JÄNTSCHI, L. Molecular Descriptors Family on Structure Activity Relationships 1. Review of the Methodology. *Leonardo Electronic Journal of Practices and Technologies*, 2005, vol. 6, no. 4, p. 76-98.

JÄNTSCHI, L. and BOLBOACĂ, S. Results from the use of molecular descriptors family on structure property/activity relationships. *International Journal of Molecular Sciences*, 2007, vol. 8, no. 3, p. 189-203.

JÄNTSCHI, L.; BOLBOACĂ, S. and DIUDEA M.V. Chromatographic retention times of polychlorinated biphenyls: from structural information to property characterization. *International Journal of Molecular Sciences*, 2007, vol. 8, no. 11, p. 1125-1157.

KNAUER, K.; LAMPERT, C. and GONZALEZ-VALERO, J. Comparison of *in vitro* and *in vivo* acute fish toxicity in relation to toxicant mode of action. *Chemosphere*, 2007, vol. 68, no. 8, p. 1435-1441.

LAGUNIN, A.A.; ZAKHAROV, A.V.; FILIMONOV, D.A. and POROIKOV, V.V. A new approach to QSAR modelling of acute toxicity. *SAR and QSAR in Environmental Research*, 2007, vol. 18, no. 3-4, p. 285-298.

Leave-one-out Analysis [online]. ©2005, Virtual Library of Free Software [cited 20 July 2007]. Available from Internet: http://l.academicdirect.org/Chemistry/SARs/MDF_SARs/loo/

LI, Y. and XI, D.-l. Quantitative structure-activity relationship study on the biodegradation of acid dyestuffs. *Journal of Environmental Sciences*, 2007, vol. 19, no. 7, p. 800-804.

LI, Z.R.; HAN, L.Y.; XUE, Y.; YAP, C.W.; LI, H.; JIANG, L. and CHEN, Y.Z. MODEL - Molecular descriptor lab: A web-based server for computing structural and physicochemical features of compounds. *Biotechnology and Bioengineering*, 2007, vol. 97, no. 2, p. 389-396.

LOADER, R.J.; SINGH, N.; O'MALLEY, P.J. and POPELIER, P.L.A. The cytotoxicity of ortho alkyl substituted 4-X-phenols: A QSAR based on theoretical bond lengths and electron densities. *Bioorganic and Medicinal Chemistry Letters*, 2007, vol. 16, no. 5, p. 1249-1254.

LUBBERS, S.; DECOURCELLE, N.; MARTINEZ, D.; GUICHARD, E. and TROMELIN, A. Effect of thickeners on aroma compound behavior in a model dairy gel. *Journal of Agricultural and Food Chemistry*, 2007, vol. 55, no. 12, p. 4835-4841.

MALÍK, I.; SEDLÁROVÁ, E.; CSÖLLEI, J.; ANDRIAMAINTY, F. and ČIŽMÃIRIK, J. Relationship between physicochemical properties, lipophilicity parameters, and local anesthetic activity of dibasic esters of phenylcarbamic acid. *Chemical Papers*, 2007, vol. 61, no. 3, p. 206-213.

MDF SAR Predictor [online]. ©2005, Virtual Library of Free Software [cited 20 July 2007]. Available from Internet: http://l.academicdirect.org/Chemistry/SARs/MDF_SARs/sar/.

NIU, B.; LU, W.-C.; YANG, S.-S.; CAI, Y.-D. and LI, G.-Z. Support vector machine for SAR/QSAR of phenethylamines. *Acta Pharmacologica Sinica*, 2007, vol. 28, no. 7, p. 1075-1086.

PRATHIPATI, P.; DIXIT, A. and SAXENA, A.K. Computer-aided drug design: Integration of structure-based and ligand-based approaches in drug design. *Current Computer-Aided Drug Design*, 2007, vol. 3, no. 2, p. 133-148.

ROY, D.R.; PARTHASARATHI, R.; SUBRAMANIAN, V. and CHATTARAJ, P.K. An electrophilicity based analysis of toxicity of aromatic compounds towards *Tetrahymena pyriformis*. *QSAR and Combinatorial Science*, 2006, vol. 25, no. 2, p. 114-122.

SCHULTZ, T.W. The use of the ionization constant (pKa) in selecting models of toxicity in phenols. *Ecotoxicology and Environment Safety*, 1987a, vol. 14, no. 2, p. 178-183.

SCHULTZ, T.W. Relative toxicity of para-substituted phenols: log KOW and pKa-dependent structure-activity relationships. *Bulletin of Environment Contamination and Toxicology*, 1987b, vol. 38, no. 6, p. 994-999.

SCHULTZ, T.W.; BEARDEN A.P. and JAWORSKA, J.S. A novel QSAR approach for estimating toxicity of phenols. *SAR and QSAR in Environmental Research*, 1996, vol. 5, no. 2, p. 99-112.

SCOTTI, L.; SCOTTI, M.T.; ISHIKI, H.M.; FERREIRA, M.J.P.; EMERENCIANO, V.P.; de S. MENEZES, C.M. and FERREIRA, E.I. Quantitative elucidation of the structure-bitterness relationship of cynaropicrin and grosheimin derivatives. *Food Chemistry*, 2007, vol. 105, no. 1, p. 77-83.

STEIGER, J.H. Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, 1980, vol. 87, p. 245-251.

TAKAHATA, Y.; ARAKAWA, M.; FUNATSU, K.; COSTA, M.C.A. and SEGALA, M. Core Electron Binding Energy (CEBE) as descriptors in Quantitative Structure - Activity Relationship (QSAR) analysis of cytotoxicities of a series of simple phenols. *QSAR and Combinatorial Science*, 2007, vol. 26, no. 3, p. 378-384.

TODESCHINI, R. and CONSONNI, V. Handbook of Molecular Descriptors. Wiley-UCH, Weinheim, 2000, 688 p. ISBN: 978-3527299133.

Training vs. Test Experiment [online]. ©2005, Virtual Library of Free Software [cited 20 July 2007]. Available from Internet: http://l.academicdirect.org/Chemistry/SARs/MDF_SARs/qsar_qspr_s/.

YANG, L.; WANG, P.; JIANG, Y.-L. and XIA, B. QSAR for toxicities of phenols using improved genetic algorithm combined with BP artificial neural network. *Journal of Harbin Institute of Technology*, 2006, vol. 38, no. 2, p. 216-218.

APPENDIX
TABLES

Table 1. Characters in the name of the molecular descriptors.

Letter	Characters
1st	Operator of linearization: *I* (identity, no change), *i* (inverse of *I*), *A* (absolute *I*), *a* (inverse of *A*), *L* (logarithm of *A*), *l* (logarithm of *I*)
2nd	Superposing operator of the molecular level: Conditional group: *m* (smallest fragmental descriptor value from the array), *M* (highest), *n* (smallest absolute), *N* (highest absolute) Averagegroup: *S* (sum of descriptor values), *A* (average mean for valid fragments), *a* (average mean for all fragments), *B* (average mean by atom), *b* (average mean by bond) Geometricgroup: *P* (multiplication of descriptor values), *G* (geometric mean for valid fragments), *g* (geometric mean for all fragments), *F* (geometric mean by atom), *f* (geometric mean by bond) Harmonic group: *s* (harmonic sum of values), *H* (harmonic mean for valid fragments), *h* (harmonic mean for all fragments), *I* (harmonic mean by atom), *i* (harmonic mean by bond)
3rd	Pair-based fragmentation criteria (Jäntschi et al. 2000; Diudea et al. 2001): *m* (minimal fragments criterion), *M* (maximal fragments criterion), *D* (Szeged distance based fragments criteria), *P* (Cluj path based fragments criteria)
4th	Interaction model: *R* (rare model and resultant relative to fragment's head), *r* (rare model and resultant relative to conventional origin), *M* (medium model and resultant relative to fragment's head), *m* (medium model and resultant relative to conventional origin), *D* (dense model and resultant relative to fragment's head), *d* (dense model and resultant relative to conventional origin)
5th	Interaction descriptor: *D* (distance), *d* (inverse of distance), *O* (first atom's property), *o* (inverse of `*O*`), *P* (product of the atomic properties), *p* (inverse of `*P*`), *Q* (squared `*P*`), *q* (inverse of `*Q*`), *J* (product of first atom property and distance), *j* (inverse of `*J*`), *K* (product of atomic properties and distance), *k* (inverse of `*K*`), *L* (product of distance and squared atomic properties), *l* (inverse of `*L*`), *V* (first atom's property potential), *E* (first atom's property field), *W* (first atom property work), *w* (properties work), *F* (first atom's property force), *f* (properties force), *S* (first atom's property weak nuclear force), *s* (properties weak nuclear force), *T* (first atom's property strong nuclear force), *t* (properties strong nuclear force)
6th	Atomic property: *C* (cardinality), *H* (count of directly bonded hydrogen's), *M* (relative atomic mass), *E* (atomic electronegativity), *G* (group electronegativity), *Q* (partial charge)
7th	Distance operator: *g* (geometry), *t* (topology)

Table 2. Experimental toxicity (Tox_{exp}) of para-substituted phenols, the values of the molecular descriptors used, and the estimated toxicity.

PhNo.	Substituent	Tox_{exp}	ASMmVQt	lfDdOQg	lnMrLQg	LsDMpQg	$\hat{Y}_{2v\text{-}Eq(1)}$	$\hat{Y}_{4v\text{-}Eq(2)}$
1	$CONH_2$	-0.7802	5.8981	-6.7972	-7.2253	-1.9763	-0.5751	-0.7594
2	$NHCOCH_3$	-0.8189	10.749	-4.1048	0.8469	-2.1507	-0.9803	-0.7707
3	CH_2CH_2OH	-0.8275	8.8351	-6.3911	0.4634	-0.1705	-0.5556	-0.8074
4	CH_2CN	-0.3840	5.3722	-9.5510	-4.9193	-1.2123	0.0014	0.0383
5	OCH_3	-0.1425	6.5051	-9.6749	-1.1651	0.1627	0.0696	-0.0613
6	CHO	0.2661	10.505	-8.8226	-3.3549	-1.6029	0.0313	0.2380
7	$COCH_3$	-0.0932	12.744	-9.4022	-6.6088	-1.0333	0.2385	0.0505
8	H	-0.4310	3.4445	-8.5607	-0.5243	0.2467	-0.2833	-0.5096
9	COC_2H_5	0.0557	11.289	-7.4760	0.0779	-1.1726	-0.2313	-0.0517
10	CN	0.5161	10.441	-10.203	-3.5618	-1.1069	0.3276	0.4979
11	F	0.0169	5.9127	-10.684	-13.455	-1.5838	0.2662	-0.1193
12	OC_2H_5	0.0130	6.6356	-10.153	-2.8067	-0.2047	0.1777	0.0966
13	NO_2	1.4257	42.349	-9.1012	-11.466	-3.1947	1.2550	1.3458
14	CH_3	-0.1920	5.2303	-9.8915	-1.2311	0.1551	0.0698	-0.0376
15	Cl	0.5447	1.5458	-11.349	-1.0827	-0.6323	0.2505	0.5941
16	C_2H_5	0.2058	7.3731	-10.845	-2.1408	-0.0230	0.3544	0.3377
17	Br	0.6806	2.5962	-11.892	-0.2017	-0.0663	0.4063	0.6814
18	I	0.8544	3.5436	-12.465	-0.1225	0.0118	0.5649	0.8804
19	OC_4H_9	0.7016	9.8871	-11.919	-6.3394	-0.1820	0.6785	0.4986
20	$CH(CH_3)_2$	0.4732	9.3579	-12.076	-3.3635	0.6969	0.6932	0.4623
21	COC_6H_5	1.0237	8.4519	-13.467	-10.680	-1.5542	0.9610	1.0857
22	C_3H_7	0.6350	9.4953	-11.501	-3.7447	-0.0624	0.5738	0.5123
23	$N{=}NC_6H_5$	1.6547	17.488	-16.636	-7.9952	1.2302	1.9765	1.6816
24	C_6H_5	1.3828	18.703	-13.846	-5.8290	-0.0576	1.4174	1.4308
25	$C(CH_3)_3$	0.9126	11.508	-13.250	-6.9593	-0.3020	1.0256	0.9817
26	OC_6H_5	1.3550	63.333	-5.8276	0.9003	-1.4657	1.3136	1.3860
27	$CH_2CH(CH_3)_2$	0.9797	13.905	-11.009	0.0942	0.2733	0.6285	0.6922
28	$c\text{-}C_5H_9$	1.2916	16.560	-13.599	-4.3256	-0.2702	1.2857	1.4682
29	$CH_2C_6H_5$	1.1946	66.575	-5.5776	0.2146	-1.1809	1.3780	1.2644
30	$CH_2C(CH_3)_3$	1.2326	19.628	-10.686	0.1528	-0.2727	0.7677	0.9795

Table 3. Structure-Activity Relationships models: statistics.

Characteristic (abbreviation)	SAR model	
	Eq(1)	Eq(2)
Correlation coefficient (r)	0.9472	0.9868
Squared correlation coefficient (r^2)	0.8972	0.9737
Adjusted squared correlation coefficient (r^2_{adj})	0.8896	0.9695
Standard error of estimated (s_{est})	0.2355	0.1238
Fisher parameter of estimated (F_{est})	118	231
Type I error associated with the Fisher parameter of estimated (p_{est})	$4.56 \cdot 10^{-14}$	$1.50 \cdot 10^{-21}$
Leave-one-out cross-validation score ($r^2_{loo\text{-}cv}$)	0.8745	0.9650
Fisher parameter of predicted (F_{pred})	93	172
Fisher probability of predicted (p_{pred})	$7.58 \cdot 10^{-13}$	$9.34\ 10^{-20}$
Standard error of predict (s_{loo})	0.2613	0.1429
Difference between squared correlation coefficient and leave-one-out cross-validation score (r^2 - $r^2_{loo\text{-}cv}$)	0.0227	0.0086

Table 4. Classification of the distance between measured and estimated toxicity.

PhNo	$\lvert NN - Tox_{exp}\rvert$	$\lvert \hat{Y}_{2v\text{-}Eq(1)} - Tox_{exp}\rvert$	$\lvert \hat{Y}_{4v\text{-}Eq(1)} - Tox_{exp}\rvert$
1	0.0102	0.2051	0.0208
2	0.0601	0.1605	0.0491
3	0.1516	0.2719	0.0201
4	0.5463	0.0398	0.0767
5	0.2317	0.2121	0.0812
6	0.0298	0.2348	0.0282
7	0.2428	0.3317	0.1437
8	0.2151	0.1477	0.0786
9	0.0368	0.2870	0.1074
10	0.1199	0.1885	0.0182
11	0.0106	0.2493	0.1362
12	0.0846	0.1647	0.0836
13	0.1800	0.1707	0.0799
14	0.2263	0.2618	0.1544
15	0.0936	0.2942	0.0494
16	0.1791	0.1486	0.1319
17	0.0976	0.2743	0.0008
18	0.3763	0.2895	0.0260
19	0.2478	0.0231	0.2030
20	0.2007	0.2200	0.0109
21	0.2786	0.0628	0.0620
22	0.2738	0.0612	0.1227
23	0.3402	0.3218	0.0269
24	0.1129	0.0346	0.0480
25	0.2046	0.1130	0.0691
26	0.2440	0.0414	0.0310
27	0.2911	0.3512	0.2875
28	0.0154	0.0059	0.1766
29	0.1049	0.1834	0.0698
30	0.0809	0.4649	0.2531

NN: values estimated using NN by (Ivanciuc, 1998); Tox_{exp}: experimental values (Schultz, 1987a); $\hat{Y}_{2v\text{-}Eq(1)}$: values estimated by Eq(1); $\hat{Y}_{4v\text{-}Eq(1)}$: values estimated by Eq(2).

Table 5. Training versus test sets: validation results.

SAR model					Training			Test		
Intercept	ASMmVQt	lfDdOQg	lnMrLQg	LsDMpQg	n_{tr}	r^2_{tr}	F_{tr} (p)	n_{ts}	r^2_{ts}	F_{ts} (p)
-3.432	0.035	-0.341	0.087	-0.379	18	0.968	98 ($1.4\cdot10^{-9}$)	12	0.980	74.92 ($7.9\cdot10^{-6}$)
-3.300	0.034	-0.329	0.082	-0.352	19	0.975	139 ($4.3\cdot10^{-11}$)	11	0.971	49.59 ($9.9\cdot10^{-5}$)
-3.253	0.035	-0.321	0.075	-0.326	20	0.972	132 ($1.7\cdot10^{-11}$)	10	0.974	43.28 ($4.5\cdot10^{-4}$)
-3.175	0.034	-0.310	0.065	-0.328	21	0.986	275 ($1.6\cdot10^{-14}$)	9	0.941	13.14 ($1.4\cdot10^{-2}$)
-3.206	0.034	-0.320	0.078	-0.314	22	0.954	89 ($3.7\cdot10^{-11}$)	8	0.994	114.7 ($1.3\cdot10^{-3}$)
						0.971	μ		0.972	
						0.011	StdDev		0.019	

n_{tr} = number of compounds in training set; r^2_{tr} = squared correlation coefficient of training set; F_{tr} = Fisher parameter of training set; n_{ts} = number of compounds of test set; r^2_{ts} = squared correlation coefficient of test set; F_{ts} = Fisher parameter of test set; μ = arithmetic mean; StdDev = standard deviation; p = type I error.

Table 6. SAR model compared with previously reported models (Ivanciuc, 1998).

Steiger Test	Model	
Characteristic	NN	MLR
$r(Tox_{exp}.\ \hat{Y}_{4v\text{-}Eq(2)})$	0.9815	
$r(Tox_{exp}.\ \hat{Y}$ (Ivanciuc, 1998)	0.9643	0.9551
$r(\hat{Y}_{4v\text{-}Eq(4)}.\ \hat{Y}$ (Ivanciuc, 1998)	0.9379	0.9273
Z (Steiger test parameter)	1.6939	2.2781
p_Z (type I error of Z parameter)	$4.51\cdot10^{-2}$	$1.14\cdot10^{-2}$

NN: neural network; MLR: Multiple Linear Regression.