Journal of Environmental Chemistry and Ecotoxicology Vol. 2(4), pp. 47-51, May 2010
Available online at http://www.academicjournals.org/jece
ISSN-2141-226X © 2010 Academic Journals

Full Length Research Paper

Prediction of inhibition effect of some aliphatic and aromatic organic compounds using QSAR method

Nasser Goudarzi[1]* Mohammad Goodarzi [2, 3] and M. Arab Chamjangali[1]

[1]Faculty of Chemistry, Shahrood University of Technology, P. O. Box 316, Shahrood, Iran.
[2]Department of Chemistry, Faculty of Sciences, Azad University, Arak, Iran.
[3]Young Researchers Club, Azad University, Arak, Iran.

Accepted 4 December, 2009

A quantitative structure-activity relationship (QSAR) model was developed for prediction of log IC_{50} values of aliphatic and aromatic alcohols based on their molecular descriptors. In this study, we have attempted to develop a simple and fast MLR model with high accuracy and precision. The molecular descriptors, which cover different information of molecular structures, were calculated by Dragon software. The most feasible descriptors were selected using forward selection. The QSAR model was validated by external set compounds without any contribution in model development step. The root means square error of prediction (RMSEP) and determination coefficient (R^2) for training and test sets were 0.0938, 0.1819, 0.9909 and 0.9714, respectively. Results obtained show the validation of the proposed model in the modeling of the Log IC50 of aliphatic and aromatic alcohols.

Key words: Median inhibition concentration, IC_{50}, quantitative structure–activity relationship, MLR.

INTRODUCTION

The ability of a compound to penetrate various biological membranes, tissues and barriers is a primary factor in controlling the interaction of these compounds with biological systems. IC_{50} values were used to measure biological activity, which is defined as median inhibition concentration (concentration that reduces the effect by 50%) (Banarjee et al., 1980). Aliphatic and aromatic alcohols are amphiprotic compounds which have both polar and nonpolar parts in their structure. These compounds are interested with respect to the noncreative toxic effects on the microorganism *Pseudomonas putida*. Alcohol toxicity to bacteria since this group is an important component of the ecosystem, to use the bacteria as early indicators of environmental problems and to establish structure-activity relationships (Miller et al., 1985; kemoto et al., 1992).

Aliphatic and aromatic alcohols act as nonspecific toxicants which have inhibitory effects on bacterial cells corresponding to chemical concentration. There are several experimental methods (*in vivo* or *in vitro*) for testing the toxicity and inhibition effect of chemicals and can provide the most reliable quantitative and qualitative data about the interaction of a given compound with a biological system (Veith et al., 1979; Kamlet et al., 1986; Stephenson and Stuart, 1986; Stephenson et al., 1984). Nevertheless, it wastes time and consumes too much material as well as being expensive, so it is not suitable for the screening of large data sets of compounds. In recent years, researches in the field of theoretical prediction of toxicity and inhibitor activity based on quantitative structure–activity relationships (QSAR) have become very attractive (Isnard and Lambert, 1988).

In quantitative structure activity relationship (QSAR) models in which physicochemical parameters of drugs and the other compounds are correlated with biological activities, lipophilicity (partition coefficients, chromatographic parameters) has a major role. Other important parameters are polarizability, electronic and steric parameters, molecular weight, geometry, conformational entropies *etc*. Recently, many molecular modeling methods based on widely spread quantitative structure–property/activity relationships (QSPR/QSAR) techniques found their place as an important tool for the chemical engineers, chemists and especially for different aims (Goodarzi and Freitas, 2008; Fatemi and Goudarzi, 2005; Goudarzi and Goodarzi, 2008; Goodarzi and Freitas, 2008; Goodarzi and Freitas,

*Corresponding author. E-mail: goudarzi@shahroodut.ac.ir, goudarzi10@yahoo.com. Tel: +982733335441. Fax: +982733335441.

2008). The aim of the present work is to propose a validate model for QSAR study of IC_{50} based on the large space of theoretically calculated molecular descriptors.

DESCRIPTOR GENERATION AND DATASET

In this study, the experimental data set was taken from Gul and Ozturk (1998). The 2D structures of the molecules were drawn by the Hyperchem 7 software (HyperChem Release 7, HyperCube, Inc., http: //www.hyper.com.).

The pre-optimization was conducted using the molecular mechanics force field (MM^+) procedure included in Hyperchem and then the molecular structures were finally optimized by the semi-empirical method PM3 (parametric method-3) using the Polak–Ribiere algorithm (Helguera et al.,2006) until the root mean square gradient was 0.001 Kcal mol^{-1}. The resultant geometry was transferred into the Dragon software package, which was developed by Milano chemometrics and QSPR group (Todeschini, Milano Chemometrics and QSPR Group, http://www.disat.unimib.it/vhml.), to calculate the descriptors in constitutional, topological, geometrical, charge, GETAWAY (Geometry, Topology and Atoms-Weighted Assembly), WHIM (Weighted Holistic Invariant Molecular descriptors), 3D-MoRSE (3D-Molecular Representation of Structure based on Electron diffraction), molecular walk count, BCUT, 2D-autocorrelation, aromaticity index, randic molecular profile, radial distribution function, functional group and atom-centered fragment classes. The 1457 descriptors were first analyzed for the existence of constant or near constant variables. The detected ones were then removed and 618 descriptors were remained. Secondly, correlation between descriptors and activity of the molecules was examined and collinear descriptors (that is, correlation coefficient between descriptors is greater than 0.9) were detected. Among the collinear descriptors, the one presenting the highest correlation with the activity to be predicted was retained and others were removed from the data matrix. At the end, 281 descriptors were remained.

Finally, eight descriptors were selected by stepwise regression for construction of MLR model. The MLR modeling is the simple model for predicting of physiochemical properties or activities for a series of molecules. These descriptors are: 3D-MoRSE-signal 29/weighted by atomic polarizabilities (Mor29p), R autocorrelation of lag3/ weighted by atomic Sanderson electro negativities (R3e), 3D-MoRSE-signal 20/ weighted by atomic Sanderson electro negativities (Mor20e), Radial Distribution function-4.5 / weighted by atomic Sanderson electro negativities (RDF045e), Moran autocorrelation-lag 5 / weighted by atomic polarizabilities (MATS5p), folding degree index (FDI) , 3D-MoRSE-signal 04/ weighted by atomic masses (Mor04m) and 3D-MoRSE-signal 23/ weighted by atomic masses (Mor23m).

The general purpose of multiple linear regressions (MLR) is to model the relationship between two or more independent variables and a dependent variable by fitting a linear equation to observed data. Every value of the independent variable X is associated with a value of the dependent variable Y. Formally, the model for multiple linear regression, given n observations, is:

$$y = b_1x_1 + b_2x_2 + b_3x_3 + \ldots + b_mx_m + \varepsilon \quad (1)$$

Where m is the number of independent variables, $b_1, \ldots, b_m$ the regression coefficients and y is the dependent variable. Also ε is a constant. Regression coefficients represent the independent contributions of each calculated molecular descriptor. The algebraic MLR model is defined in Eq. (2) and in matrix notation:

$$y = Xb + e \quad (2)$$

When X is of full rank, the least-squares solution is:

$$\hat{b} = (X^TX)^{-1}X^Ty \quad (3)$$

Here, $\hat{b}$ is the estimator for the regression coefficients in $\hat{b}$. The MLR model was built using a training set and validation using an external prediction set. Multiple linear regression (MLR) techniques based on least-squares procedures are very often used for estimating the coefficients involved in the model equation [3].

RESULTS AND DISCUSSION

As a matter of fact the quantitative structure-activity or property relationship (QSAR/QSPR) plays important role in design of compounds and help us to reduce consuming of time and money. Therefore researchers paid more attention to like these studies, altogether ability of prediction of QSAR/QSPR studies affected by two parameters, which one is descriptors that could very carry enough information of molecular structure for interpretation of the activity or property and the other is the modeling method employed.

However one of the most important points in this work is that we have used multiple linear regressions as a simple, fast and precise method with high accuracy to predict activity of the mentioned compounds. In the first step, the 34 compounds data set was separated based on activity range into a training set of 27 compounds between 2.01-5.88, which is including 80% of whole dataset and a test set of 10 compounds between 2.18 - 5.25 that is including 20% of whole dataset. Then, we made use of forward selection as a common feature selection on training dataset. The forward selection (FA) method adds variables to the model one at a time. The first variable included in the model is the one which has the highest correlation with the independent variable *y*. The variable that enters the model as the second variable is one which has the highest correlation with *y*, after *y* has been adjusted for the effect the first variable. This process terminated when the last variable which entered in the model, has insignificant regression coefficient or all the variables are included in the model.

Descriptors that have been selected using the FS method are shown in Table 1. Table 2 shows all the information about the descriptors. As it can be seen from the correlation matrix (Table 2), there is no significant correlation (≥ 0.9) between the selected descriptors. It should be noted that we constructed linear models with different number of descriptors, as it can be seen in the Figure 1. When we add the descriptors into the model, the correlation coefficient is improved and also the standard error is decreased, so this figure shows that if we use of all descriptors that have been selected by FS, the model results for prediction of log IC_{50} is better.

In order to build linear and test model, the 27 compounds data set was used as training to build model and 10 compounds as a test set for an external set

Table 1. Descriptors are presented in the models.

No.	Compounds	Mor29p	R3e	Mor20e	RDF045e	MATS5p	FDI	Mor04m	Mor23m
1	Methanol	0.027	0.644	0.138	0	0	0.84	-0.297	-0.013
2	Ethanol	-0.024	1.722	0.087	0	0	0.906	-0.353	0.044
3	1-Propanol	-0.039	1.935	0.171	1.976	0.369	0.927	-0.522	0.043
4	2-Propanol	-0.037	1.966	0.173	0.05	0	0.92	-0.241	0.038
5	1,2-Propanediol	-0.01	2.101	0.176	2.08	0.365	0.928	-0.303	0.037
6	1,2,3-Propanetriol	0.003	2.191	0.262	1.901	0.317	0.929	-0.223	-0.004
7	1-Butanol	-0.057	2.016	0.248	1.577	0.23	0.942	-0.572	0.051
8	2-Butanol	-0.064	1.99	0.193	0.94	0.393	0.932	-0.287	0.062
9	2- Methyl-1-propanol	-0.032	1.889	0.308	2.399	0.393	0.929	-0.392	0.04
10	2-Methyl-2-propanol	-0.029	2.009	0.388	0.134	0	0.924	0.005	0.037
11	1-Pentanol	-0.05	2.105	0.54	1.826	-0.037	0.964	-0.732	0.048
12	1-Hexanol	-0.05	2.105	0.54	1.826	-0.037	0.964	-0.732	0.048
13	Cyclohexanol	-0.163	2.764	0.093	3.059	0.377	0.958	-0.48	0.069
14	1-Heptanol	-0.062	2.143	0.614	2.123	-0.024	0.973	-0.778	0.063
15	1-Octanol	-0.064	2.198	0.681	2.255	-0.017	0.981	-0.87	0.061
16	Phenol	0.074	1.11	0.534	3.182	0.555	1	-0.495	-0.228
17	2-Methylphenol	0.068	1.107	0.43	4.246	0.298	0.99	-0.448	-0.183
18	3-Methylphenol	0.077	1.064	0.637	3.713	0.023	0.996	-0.476	-0.182
19	4-Methylphenol	0.072	1.16	0.664	4.715	0.226	0.997	-0.616	-0.234
20	2-Aminophenol	0.092	1.022	0.624	4.437	0.358	0.992	-0.487	-0.261
21	4-Aminophenol	0.066	1.061	0.685	5.941	0.294	0.996	-0.665	-0.263
22	3-Nitrophenol	0.108	1.013	0.522	1.956	0.201	1	-0.313	-0.086
23	4-Nitrophenol	0.11	1.085	0.459	3.276	0.324	1	-0.436	-0.066
24	2,4-Dinitrophenol	0.117	0.992	0.44	1.872	0.139	1	0.122	0.024
25	1,2-Dihydroxybenzene	0.098	1.105	0.499	5.258	0.427	0.996	-0.48	-0.306
26	1,3-Dihydroxybenzene	0.097	1.098	0.501	3.609	0.326	0.999	-0.557	-0.28
27	1,4-Dihydroxbenzene	0.101	1.145	0.452	5.651	0.386	0.998	-0.762	-0.249
28	1,2,3-Trihydroxybenzene	0.121	1.096	0.488	5.324	0.317	0.994	-0.355	-0.406
29	2-Chlorophenol	0.045	1.153	0.512	3.619	0.333	1	-0.242	-0.318
30	3-Chlorophenol	0.021	1.132	0.534	3.136	-0.149	1	-0.281	-0.254
31	4-Chlorophenol	0.016	1.15	0.492	4.356	0.436	1	-0.572	-0.23
32	4-Ethylphenol	0.057	1.628	0.774	8.447	0.003	0.982	-0.599	-0.237
33	Benzyl alcohol	0.094	1.337	0.837	2.089	0.119	0.998	-0.359	-0.172
34	2- Naphthol	-0.038	1.597	0.897	3.286	-0.073	1	-0.548	-0.404

Table 2. Correlation matrix for the eight selected descriptors by forward selection and details of names of descriptors.

	Mor29p	R3e	Mor20e	RDF045e	MATS5p	FDI	Mor04m	Mor23m
Mor29p	1							
R3e	0.849	1						
Mor20e	0.169	0.202	1					
RDF045e	0.230	0.232	0.256	1				
MATS5p	0.101	0.110	0.078	0.110	1			
FDI	0.408	0.513	0.565	0.405	0.036	1		
Mor04m	0.039	0.008	0.139	0.185	0.0002	0.087	1	
Mor23m	0.393	0.504	0.384	0.552	0.068	0.494	0.039	1

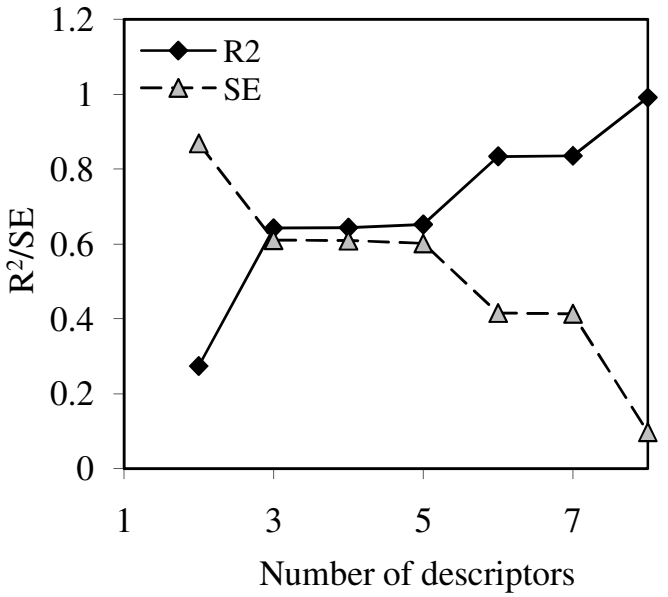

Figure 1. Plot of correlation coefficient and standard error versus the number of descriptors that have used in different MLR models.

compounds, which did not have contribution in model development steps that was used to test the built model. Finally with the selected eight descriptors, we built the linear model using the training set data and the following equation was obtained:

$LogIC_{50}$ = 24.47+13.152×Mor29p+2.0441×R3e-1.6888×Mor20 e-0.23626 ×RDF045e+1.2104 × MATS5p -24.62 × FDI -0.82693 × Mor04m -5.3278 × Mor23m.

The constructed model was used to predict the test sets data. The prediction results were given in Table 3. Figure 2 shows experimental values versus the Log IC_{50} predicted by multiple linear regressions (MLR). Also, the residuals of the MLR calculated values of $logIC_{50}$ are plotted against the experimental values in Figure 3. The propagation of residuals at both sides of the zero line indicates that no systematic error exists in the development of MLR model. Obviously for evaluating the prediction ability of a multivariate calibration model we can use of several statistical test such as the *F* statistical, *t* test, determination coefficient (R^2), root mean square error of prediction (RMSEP) and relative standard error of prediction (RSEP) and mean absolute error (MAE) values (Goodarzi et al., 2007).

The statistical results in Table 4 show that FS-MLR was achieved in this study and reliable in predicting of $logIC_{50}$ of this class compounds. The some of these descriptors encode the size, shape, electronegativities, polarizabilities and 3-D structure of a molecule in respect to some molecular properties. The appearance of these descriptors in the model reveals the role of these properties on the IC_{50} activity.

Table 3. Experimental values observed and predicted values of the log IC50.

No.	Compounds	Exp. Log IC50	Predicted
1*	Methanol	5.250	5.542
2	Ethanol	5.360	5.279
3	1-Propanol	4.880	4.983
4	2-Propanol	5.000	5.044
5	1,2-Propanediol	5.440	5.492
6	1,2,3-Propanetriol	5.880	5.814
7	1-Butanol	4.350	4.337
8*	2-Butanol	4.570	4.585
9	2- Methyl-1-propanol	4.490	4.538
10	2-Methyl-2-propanol	4.600	4.558
11*	1-Pentanol	3.590	3.343
12	1-Hexanol	3.340	3.343
13	Cyclohexanol	4.050	3.996
14	1-Heptanol	2.720	2.820
15	1-Octanol	2.680	2.660
16	Phenol	3.800	3.734
17	2-Methylphenol	3.390	3.230
18*	3-Methylphenol	2.360	2.574
19	4-Methylphenol	3.210	3.036
20	2-Aminophenol	3.440	3.470
21	4-Aminophenol	2.600	2.732
22	3-Nitrophenol	3.050	2.958
23	4-Nitrophenol	3.080	3.069
24	2,4-Dinitrophenol	2.010	2.171
25	1,2-Dihydroxybenzene	3.790	3.955
26*	1,3-Dihydroxybenzene	4.240	4.043
27	1,4-Dihydroxbenzene	3.780	3.893
28	1,2,3-Trihydroxybenzene	4.590	4.588
29*	2-Chlorophenol	3.380	3.376
30*	3-Chlorophenol	2.180	2.203
31	4-Chlorophenol	2.900	2.777
32	4-Ethylphenol	2.920	2.829
33	Benzyl alcohol	3.390	3.319
34	2- Naphthol	2.730	2.841

* Compounds were used in test set.

Conclusion

In this study, multiple linear regressions were used as a simple method to construct a quantitative relation between the $logIC_{50}$ and their calculated descriptors. We have used forward selection as a common feature selection that this technique has selected eight important descriptors. Descriptors appearing in this QSAR models were Mor29p, R3e, Mor20e, RDF045e, MATS5p, FDI, Mor04m and Mor23m that provided enough information related to different molecular properties, which can

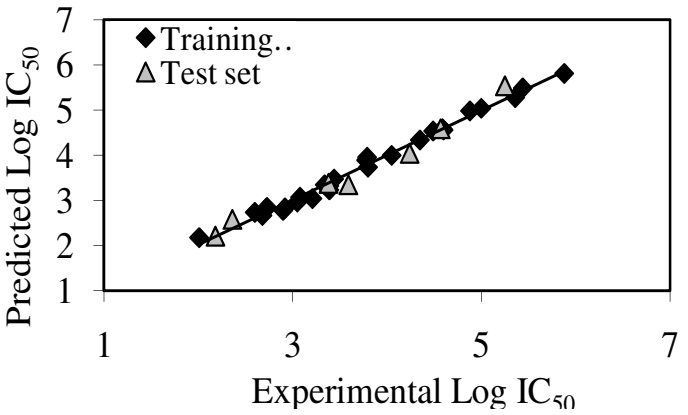

Figure 2. Plot of the calculated Log IC_{50} against the experimental values.

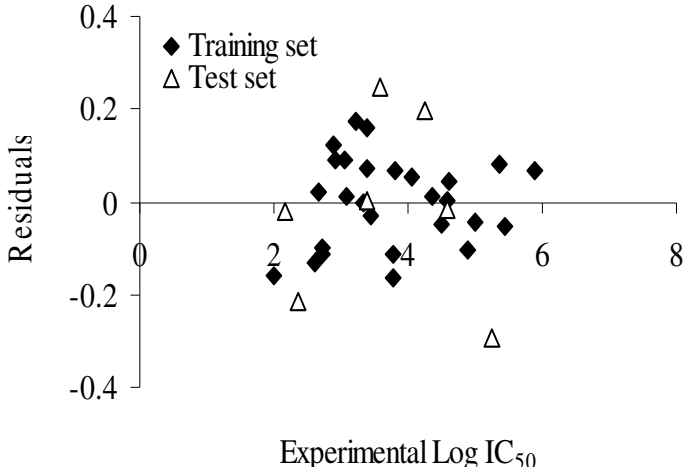

Figure 3. Plot of the residuals *versus* experimental values of log IC_{50}.

participate in the physicochemical process that affected the log IC_{50} of the compounds. The results obtained demonstrated that the MLR is a simple and fast model which has good ability for prediction of IC50.

Table 4. Statistical parameters obtained using the FS-MLR.

Parameters	FS-MLR	
	Training set	Test set
R^2	0.991	0.971
RMSEP	0.094	0.182
RSEP (%)	2.415	4.788
MAE (%)	5.403	14.228
F test	2712.7	169.750
T stat	52.084	13.028

REFERENCES

Banarjee S, Yalkowsky SH, Valvani SC (1980). Water solubility and octanol/water partition coefficients of organics. Limitations of the solubility-partition coefficient correlation. Environ. Sci. Technol. 14: 1227-1229.

Miller M, Wasik SP, Huang GL, Shiu WY, Mackay D (1985). Relationship between water octanol-water coefficient and aqueous solubility. Environ. Sci. Technol. 19: 522-529.

Kemoto I, Motoba K, Suzuki T, Uchida M (1992). Quantitative structure-activity relationships of nonspecific and specific toxicants in several organism species environ. Toxicol. Chem. 11: 931-939.

Veith D, Austin NN, Morris RT (1979). A rapid method for estimating log *P* for organic chemicals. Water Res. pp. 1343-1347.

Kamlet JRM, Doherty GD, Veith RW, Taft MH (1986). Abraham, Solubility properties in polymers and biological media. 7. An analysis of toxicant properties that influence inhibition of bioluminescence in Photobacterium phosphoreum (the Microtox test), Environ. Sci. Technol. 19: 690-695.

Stephenson R, Stuart J (1986). Mutual binary solubilities: water-alcohols and water-esters, J. Chem. Eng. Data. 31: 56-70.

Stephenson R, Stuart J, Tabak M (1984). Mutual solubility of water and aliphatic alcohols J. Chem. Eng. Data. 29: 287-290.

Isnard P, Lambert S (1988). Estimating bioconcentration factors from octanol-water partition coefficient and aqueous solubility, Chemosphere 17: 21-34.

Goodarzi M, Freitas MP (2008). Augmented three-mode MIA-QSAR modelling for a series of anti-HIV-1 compounds. QSAR Comb. Sci. 27: 1092-1098.

Fatemi MH, Goudarzi N (2005). Electrophoresis 26: 2968-2973.

Goudarzi N, Goodarzi M (2008). Prediction of the Logarithmic of Partition Coefficients (Log P) of some Organic Compounds by Least Square Support Vector Machine (LSSVM). Mol. Phys. 106: 2525-2535.

Goodarzi M, Freitas MP (2008). Predicting Boiling Points of Aliphatic Alcohols through Multivariate Image Analysis Applied to Quantitative Structure–Property Relationships. J. Phys. Chem. A. 112: 11263-11265.

Goodarzi M, Freitas MP (2008).On the use of PLS and N-PLS in MIA-QSAR: Azole Antifungals. Chemometr. Intell. Lab. Syst., doi:10.1016/j.chemolab.2008.11.007 in press.

Gul S, Ozturk D (1998).Determination of Structure-Toxicity Relationship of Amphiprotic Compounds by Means of the Inhibition of the Dehydrogenase Activity of *Pseudomonas putida*. Turk. J. Chem. 22: 341-349.

HyperChem Release 7, HyperCube, Inc., http://www.hyper.com.

Helguera AM, Duchowicz PR, Pérez MAC, Castro EA, Cordeiro MNDS, González MP (2006). Application of the replacement method as a novel variable selection strategy in QSAR. 1. Carcinogenic potential. Chemometr. Intell. Lab. Syst. 81: 180-187.

Todeschini R, Milano Chemometrics and QSPR Group, http://www.disat.unimib.it/vhml.

Goodarzi M, Goodarzi T, Ghasemi N (2007). Spectrophotometric Simultaneous Determination of Manganese (II) and Iron (II) in Pharmaceutical by Orthogonal Signal Correction-Partial Least Squares. Ann. Chim. 97: 303-312.