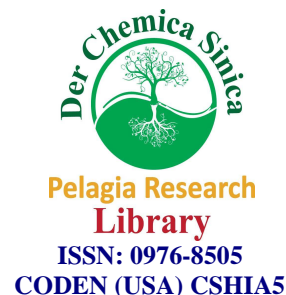




## Pelagia Research Library

Der Chemica Sinica, 2011, 2 (4):235-243



### A Quantitative Structure-Activity Relationship (QSAR) Study of Anti-cancer Drugs

S. Vahdani and Z. Bayat\*

Department of Chemistry, Islamic Azad University, Quchan Branch, Iran

#### ABSTRACT

A very simple, strong, descriptive and interpretable model, based on a quantitative structure-activity relationship (QSAR), is developed using multiple linear regression approach and quantum chemical descriptors derived from HF theories using 6-31G\* basis set for determination of the inhibit 50% of sensitive cell growth (pLD50) of some anti-cancer drugs. By molecular modeling and calculation of descriptors, two significant descriptors related to the pLD50 values of the anti-cancer drugs, were identified. A multiple linear regression (MLR) model based on 13 molecules as a training set has been developed for the prediction of the pLD50 of some anti-cancer drugs using these quantum chemical descriptors. The effects of these theoretical descriptors on the biological activity are discussed. A model with low prediction error and high correlation coefficient was obtained. This model was used for the prediction of the pLD50 values of some anti-cancer drugs. A multi-parametric equation containing maximum two descriptors at HF/6-31G\* method with good statistical qualities ( $R^2_{train}=0.915$ ,  $F_{train}=54.43$ ,  $Q^2_{LOO}=0.891$ ,  $R^2_{adj}=0.899$ ,  $Q^2_{LGO}=0.879$ ) was obtained by Multiple Linear Regression using stepwise method.

**Keywords:** MLR, HF, Anthracyclines, LD50, Ab initio.

#### INTRODUCTION

Doxorubicin is widely used anthracyclines anti-cancer agent [1]. Its clinical use is hampered by the common side-effects observed with the use of the majority of anticancer agents: bone marrow suppression, alopecia, nausea, and vomiting. Doxorubicin-induced bone marrow suppression can now be reduced by the use of hematopoietic growth factors [2,3]. The experimental measurement of the inhibition activity of chemicals is difficult, expensive and time-consuming, thus a great deal of effort has been put into attempting the estimation of activity through statistical modeling. One of the most successful approaches to the prediction of chemical

properties starting only with molecular structural information is modeling of quantitative structure– activity/property relationships (QSAR/QSPR). The concept that there exists a close relationship between bulk properties of compounds and their molecular structure allows one to provide a clear connection between the macroscopic and the microscopic properties of matter. Quantitative structure–activity relationships are mathematical equations relating chemical structure to a wide variety of physical, chemical, biological, and technological properties. QSPR models, once established can be used to predict properties of compounds as yet unmeasured or even unknown [4-7]. A major step in constructing the QSAR models is finding a set of molecular descriptors that represent variation in the structural activity of the molecules. A wide variety of descriptors such as steric, electronic and Distance based topological descriptors have been reported for use in QSAR analysis [8–14]. In most cases, it is more convenient to consider a linear relationship between activity/property and descriptors. Multiple linear regression (MLR), principal component regression (PCR), partial least squares (PLS) regression and artificial neural networks (ANN) are the most commonly used modeling methods in QSAR [15-18]. There are many reports of QSAR approaches to predict the pLD50 of drugs [19-23].

In this study, a continuation of our earlier studies to develop quantitative structure activity relationships (QSAR), a new QSAR model is developed from the descriptors derived from HF theories using 6-31G\* basis set quantum chemical calculations for predicting the pLD50 values of some of anthracyclines. Our goal here is to develop an accurate, simple, fast, and less expensive method for calculation of pLD50 values. The MLR method was applied in QSAR for modeling the relationship between inhibit 50% of sensitive cell growth (pLD50) of 13 anthracyclines. The correlation coefficient ( $R^2$ ) for the estimated versus observed pLD50 values is 0.9882 for anthracyclines.

### Data And Methods

The QSAR model for the estimation of the pLD50 of various anti-cancer drugs is established in the following six steps: the molecular structure input and generation of the files containing the chemical structures is stored in a computer readable format; quantum mechanics geometry is optimized with a abinito method; structural descriptors are computed; structural descriptors are selected; and the structure pLD50 model is generated by the MLR, and statistical analysis.

#### 2.1.Data

All pLD50 data for all 13 compounds were taken from the literature [24, 25]. The pLD50 of these compounds are deposited in Journal log as supporting material (see Tables 1). LD50 values were calculated in  $\mu\text{g/gm}$  body weight of the insect [26].

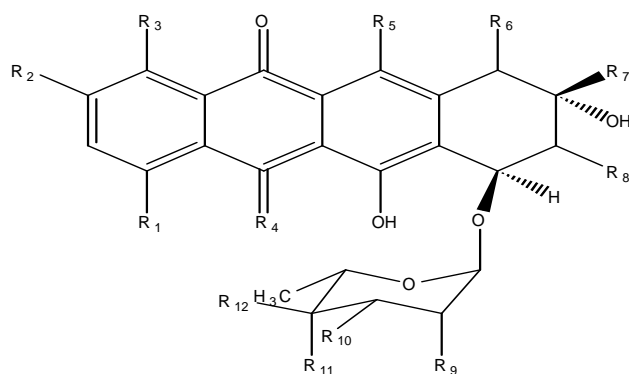
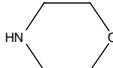


Table 1. Chemical structures and the corresponding observed and predicted pLD50 values by the MLR method.

N	R <sub>1</sub>	R <sub>2</sub>	R <sub>3</sub>	R <sub>4</sub>	R <sub>5</sub>	R <sub>6</sub>	R <sub>7</sub>	R <sub>8</sub>	R <sub>9</sub>	R <sub>10</sub>	R <sub>11</sub>	R <sub>12</sub>	Exp.	Pread.	Ref.
1	OCH <sub>3</sub>	H	H	O	OH	H	COCH <sub>2</sub> OH	H	H	NH <sub>2</sub>	OH <sub>indo</sub>	H	21.8	18.7	21
2	OCH <sub>3</sub>	H	H	O	OH	H	COCH <sub>3</sub>	H	H	NH <sub>2</sub>	OH	H	20	19.3	21
3	H	H	H	O	OH	H	COCH <sub>3</sub>	H	H	OH	OH	H	16.2	17.9	21
4	OCH <sub>3</sub>	H	H	O	OH	H	COCH <sub>2</sub> OCH <sub>3</sub>	H	H	NH <sub>2</sub>	OH	H	14.2	14.7	21
5	OCH <sub>3</sub>	H	H	O	OH	H	COCH <sub>2</sub> OH	H	H	NH <sub>2</sub>	H	H	14.1	14.9	21
6	OCH <sub>3</sub>	H	H	O	OH	H	COCH <sub>3</sub>	H	H	NH <sub>2</sub>	H	H	17.9	19	21
7	OCH <sub>3</sub>	H	H	O	OH	H	COCH <sub>2</sub> OH	H	H		OH	H	17	17.5	20
8	OH	H	H	O	OH	COOCH <sub>3</sub>	CH <sub>2</sub> CH <sub>3</sub>	H	H	N(CH <sub>3</sub> ) <sub>2</sub>	OH	H	36	36.1	20
9	OCH <sub>3</sub>	H	H	O	OH	H	COCH <sub>2</sub> OCO(CH <sub>2</sub> ) <sub>3</sub> CH <sub>3</sub>	H	H	NHCOCF <sub>3</sub>	OH	H	13.9	13.7	20
10	OH	H	H	O	OH	H	COCH <sub>3</sub>	H	H	NH <sub>2</sub>	H	H	13.5	15.9	20
11	OH	H	HO	O	H	H	COCH <sub>2</sub> OH	H	H	NH <sub>2</sub>	N(CH <sub>3</sub> ) <sub>2</sub>	H	18.7	15.3	21
12	H	H	OH	O	H	CO <sub>2</sub> CH <sub>3</sub>	CH <sub>2</sub> CH <sub>3</sub>	H	H	N(CH <sub>3</sub> ) <sub>2</sub>	OH	H	16.5	17.5	20
13	CH <sub>3</sub>	H	H	O	OH	H	COCH <sub>3</sub>	H	H	NH <sub>2</sub>	H	H	18.7	17.9	20

## 2.2. Molecular descriptor generation

All of the molecules were drawn into the Hyper Chem. The Gaussian 03 package was used for calculating the molecular descriptors. Some of the descriptors are obtained directly from the chemical structure, e. g. constitutional, geometrical, and topological descriptors. Other chemical and physicochemical properties were determined by the chemical structure (lipophilicity, hydrophilicity descriptors, electronic descriptors, energies of interaction). In this work, we used Gaussian 03 for ab initio calculations. HF method at 6-31G\* were applied for optimization of anti-cancer drugs and calculation of many of the descriptors. A large number of descriptors were calculated by Gaussian package and Hyperchem software (Table2 ). One way to avoid data redundancy is to exclude descriptors that are highly inter correlated with each other before performing statistical analysis.

Table2. The calculated descriptors used in this study.

Descriptors	Symbol	Abbreviation	Descriptors	Symbol	Abbreviation
Quantum chemical descriptors	Molecular Dipole Moment	MDP	Quantum chemical descriptors	difference between LUMO and HOMO	E <sub>GAP</sub>
	Molecular Polarizability	MP		Hardness [ $\eta=1/2 (HOMO+LUMO)$ ]	H
	Natural Population Analysis	NPA		Softness ( $S=1/\eta$ )	S
	Electrostatic Potentialc	EP		Electro negativity [ $\chi=-1/2 (HOMO-LUMO)$ ]	X
	Highest Occupied Molecular Orbital	HOMO		El Electro philicity ( $\omega=\chi^2/2\eta$ )	$\Omega$
Chemical properties	Lowest Unoccupied Molecular Orbital	LUMO	Chemical properties	Mullikenl Chargeg	MC
	Partition Coefficient	Log P		Molecule surface area	SA
	Mass	M		Hydration Energy	HE
	Molecule volume	V		Refractivity	REF

## 2.3. Genetic algorithm for descriptor selection

Genetic algorithm variable selection is a technique that helps identify a subset of the measured variables that are, for a given problem, the most useful for a precise and accurate regression model. The selection of relevant descriptors, which relate the pLD50 to the molecular structure, is an important step to construct predictive models. The genetic algorithm was applied to the input set of 13 molecular descriptors for each chemical of the studied data sets and the related response, in order to extract the best set of molecular descriptors, which are, in combination, the most relevant variables in modeling the response of the training set chemicals.

Genetic algorithm (GA), included in the PLS Toolbox version 2.0, was used for variables selection (based on the training set). Using GA-based MLR variable selection procedures, the dependent variables, i.e., the pLD50, were used to find subsets of molecular descriptors that provide a good relationship to the pLD50. Given an X-matrix of descriptors data and a pLD50 of values to be predicted, one can choose a random subset of variables from **X** and, through the use of cross-validation and MLR regression method, determine the root-mean-square error of cross-validation (RMSECV) obtained when using only that subset of variables in a regression model. Genetic algorithms use this approach iteratively to locate the variable subset (or subsets) which gives the lowest RMSECV. The first step of the GA is to generate a large number (e.g., 32, 64, 128) of random selections of the variables and calculate the RMSECV for each of the given subsets. Each subset of variables is called an individual (or chromosome) and the yes/no flags indicating which variables are used by that individual is the gene for that individual. The pool of all tested individuals is the population. The RMSECV values, described as the fitness of the individual, indicate how predictive each individual's selection of variables is for the pLD50 [27].

## RESULTS AND DISCUSSION

The diversity of the training set and the test set was analyzed using the principal component analysis (PCA) method. The PCA was performed with the calculated structure descriptors for the whole data set to detect the homogeneities in the data set, and also to show the spatial location of the samples to assist the separation of the data into the training and test sets. The PCA results showed that three principal components (PC1 and PC2) described 24.39% of the overall variables, as follows: PC1 = 64.03% and PC2 = 35.97%. Since almost all the variables can be accounted for by the first three PCs, their score plot is a reliable representation of the spatial distribution of the points for the data set. The multi-collinearity between the above seven descriptors were detected by calculating their variation inflation factors (VIF), which can be calculated as follows:

$$\text{VIF} = \frac{1}{1-r^2} \quad (1)$$

where *r* is the correlation coefficient of the multiple regression between the variables in the model. If VIF equals to 1, then no inter-correlation exists for each variable; if VIF falls into the range of 1–5, the related model is acceptable; and if VIF is larger than 10, the related model is unstable and a recheck is necessary [28]. The corresponding VIF values of the seven descriptors are shown in Table 2. As can be seen from this table, most of the variables had VIF values of less than 5, indicating that the obtained model has statistic significance. To examine the relative importance as well as the contribution of each descriptor in the model, the value of the mean effect (MF) was calculated for each descriptor. This calculation was performed with the equation below:

$$\text{MF}_j = \frac{\beta \sum_{i=1}^n dij}{\sum_j^m \beta_j \sum_i^n \beta_{ij}} \quad (2)$$

Where *MF<sub>j</sub>* represents the mean effect for the considered descriptor *j*, *β<sub>j</sub>* is the coefficient of the descriptor *j*, *dij* stands for the value of the target descriptors for each molecule and, eventually, *m* is the descriptors number for the model. The MF value indicates the relative importance of a

descriptor, compared with the other descriptors in the model. Its sign indicates the variation direction in the values of the activities as a result of the increase (or reduction) of the descriptor values. The mean effect values are shown in Table 3.

**Table 3. The linear model based on the tree parameters selected by the GA-MLR method.**

Descriptor	Chemical meaning	MF <sup>a</sup>	VIF <sup>b</sup>
Constant	Intercept	0	0
V	Volume	0.124187	1.085467
$\sigma_6$	Isotropic parameter 6	0.875813	1.085467

<sup>a</sup> Mean effect

<sup>b</sup> Variation inflation factors

All descriptors were calculated for the neutral species. The pLD50 is assumed to be highly dependent upon the V and  $\sigma_6$ . In the present study, the QSAR model was generated using a training set of 13 molecules (Table 2).

### 3.1. MLR analysis

The software package used for conducting MLR analysis was Spss 16. Multiple linear regression (MLR) analysis has been carried out to derive the best QSAR model. The MLR technique was performed on the molecules of the training set shown in Table 1. A small number of molecular descriptors (V and  $\sigma_6$ ) proposed were used to establish a QSAR model. Multiple linear regression analysis provided a useful equation that can be used to predict the pLD50 of drug based upon these parameters. The best equation obtained for the toxicity of the drug compounds is:

$$\text{pLD50} = 157.49(\pm 14.25) - 0.013V(\pm 0.004) - 1.03\sigma_6(\pm 0.098) \quad (3)$$

$$N=13 \quad R^2=0.915 \quad F=54.428 \quad R^2_{\text{adj}}=0.899 \quad Q^2_{\text{LOO}}=0.891 \quad Q^2_{\text{LGO}}=0.879$$

In this equation, N is the number of compounds,  $R^2$  is the squared correlation coefficient,  $Q^2_{\text{LOO}}$ ,  $Q^2_{\text{LGO}}$  are the squared cross-validation coefficients for leave one out, F is the Fisher F statistic. The figures in parentheses are the standard deviations. As can be seen from Table 1, the calculated values for the pLD50 are in good agreement with those of the experimental values. The predicted values for pLD50 for the compounds in the training set using equation 3 were plotted against the experimental pLD50 values in Figure 1. A plot of the residual for the predicted values of pLD50 for both the training and test sets against the experimental pLD50 values are shown in Figure 2.

The real usefulness of QSAR models is not just their ability to reproduce known data, verified by their fitting power ( $R^2$ ), but is mainly their potential for predictive application. For this reason the model calculations were performed by maximizing the explained variance in prediction, verified by the leave-one-out cross- the possibility of overestimating the model predictivity by using  $Q^2_{\text{LOO}}$  procedure, as is strongly recommended for QSAR modeling. The  $Q^2_{\text{LOO}}$  and  $Q^2_{\text{LGO}}$  for the MLR model are shown in Equation 3. This indicates that the obtained regression model has a good internal and external predictive power.

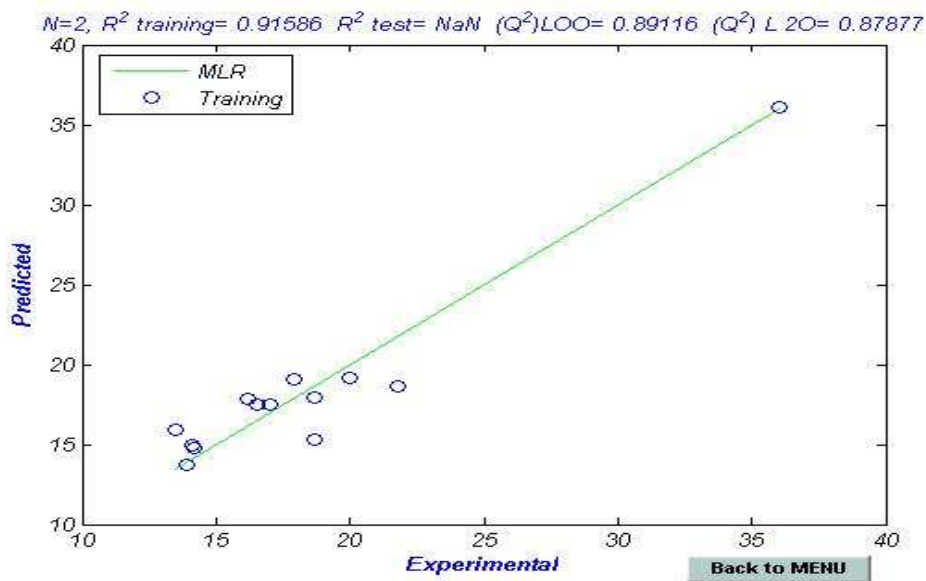


Figure 1. The predicted versus the experimental pLD50 by MLR

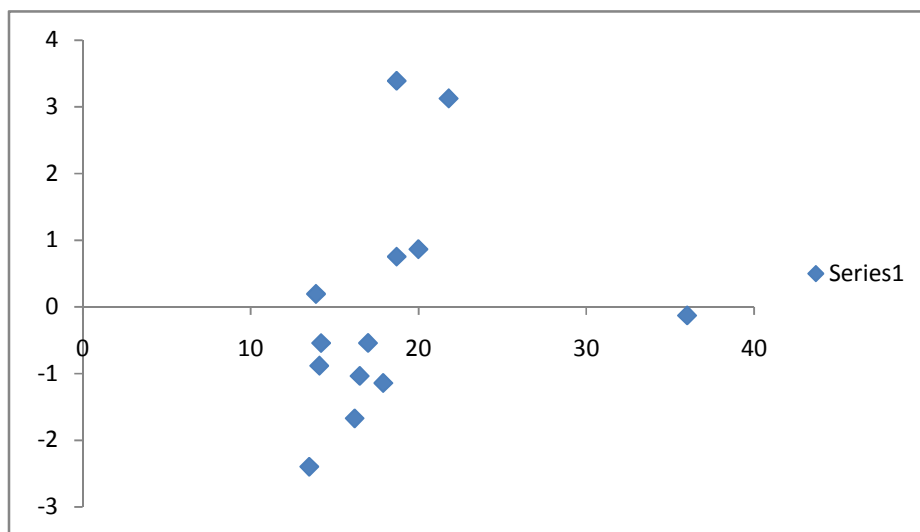


Figure 2. The residual versus the experimental pLD50 by GA-MLR.

Also, in order to assess the robustness of the model, the Y-randomization test was applied in this study [29, 30]. The dependent variable vector (pLD50) was randomly shuffled and the new QSAR models (after several repetitions) would be expected to have low  $R^2$  and  $Q^2_{\text{LOO}}$  values (Table 4). If the opposite happens then an acceptable QSAR model cannot be obtained for the specific modeling method and data.

Table 4. The  $R^2_{\text{train}}$  and  $Q^2_{\text{LOO}}$  values after several Y-randomization tests

No	$Q^2$	$R^2$
1	0.054512	0.001715
2	0.380388	0.798541
3	0.005013	0.204231
4	0.019634	0.16814
5	8.22E-05	0.048333
6	0.152036	0.017046
7	0.078982	0.073981
8	0.00268	0.073007
9	0.201614	0.035702
10	0.010152	0.096762

The MLR analysis was employed to derive the QSAR models for different the anti-cancer drugs. MLR and correlation analyses were carried out by the statistics software SPSS (Table 5).

Table 5. The correlation coefficient existing between the variables used in different MLR and equations with HF/6-31G\* method.

	V	$\sigma_6$
V	1	0
$\sigma_6$	-0.2806	1

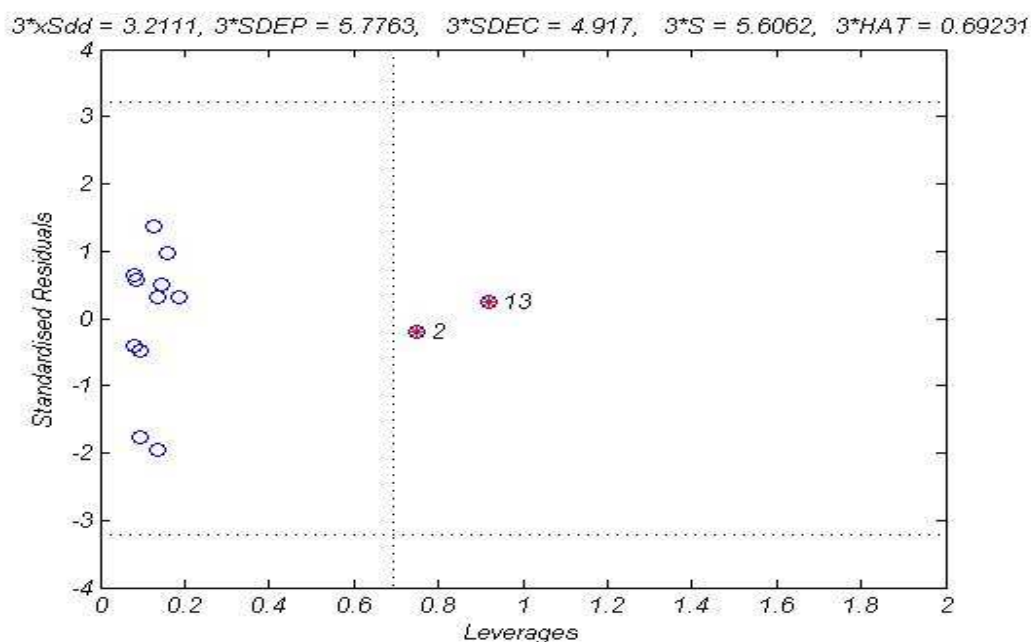
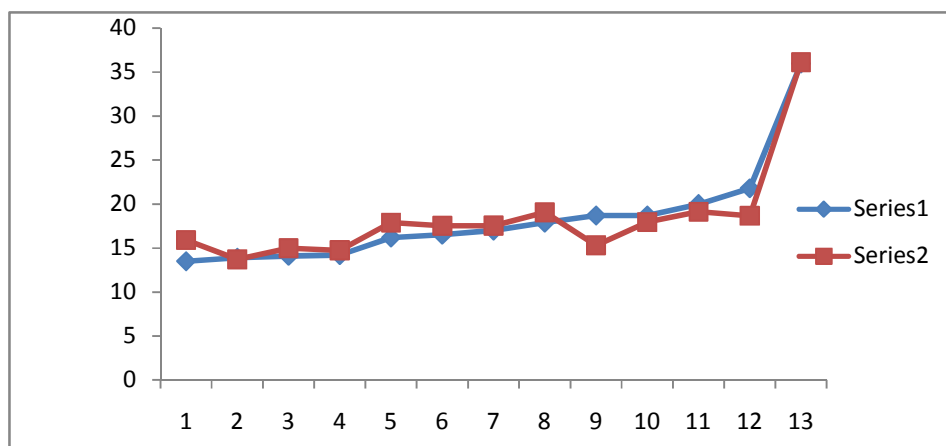


Figure 3. The William plot of the GA-MLR model.

The Williams plot (Figure 3), the plot of the standardized residuals versus the leverage, was exploited to visualize the applicability domain. The leverage indicates a compound's distance from the centroid of X. The leverage of a compound in the original variable space is defined as [31]:

Figure 4 has showed that results were obtained from equation HF/6-31G\* to the experimental values.



Series 1: the values of pLD50 were obtained by using prediction.  
 Series 2: the values of pLD50 were obtained by using Experimental methods

Figure 4. The comparison between biological activity (pLD50) using experimental and prediction

## CONCLUSION

In this article, a QSAR study of 13 anti-cancer drugs was performed based on the theoretical molecular descriptors calculated by the GAUSSIAN software and selected. The built model was assessed comprehensively (internal and external validation) and all the validations indicated that the QSAR model built was robust and satisfactory, and that the selected descriptors could account for the structural features responsible for the anti-cancer drugs activity of the compounds. The QSAR model developed in this study can provide a useful tool to predict the activity of new compounds and also to design new compounds with high activity.

## REFERENCES

- [1] Parabathina R, Muralinath E, Lakshmana S, Krishna H, Srinivasa R. *Der Pharmacia Sinica*, **2011**; 2 (2): 285-298.
- [2] Huuskonen J., Salo M., Taskinen J., *J. Pharm. Sci.*, 86, 450—454 (1997).
- [3] Ravichandiran V, Masilamani K, Senthilnathan B. Liposome- A Versatile Drug Delivery System *Der Pharmacia Sinica*, **2011**; 2 (1): 19-30.
- [4] M. Ravi, A.J. Hopfinger, R.E. Hormann, L. Dinan, *J. Chem. Inf. Comput. Sci.* 41 (2001) 1587.
- [5] B.T. Luke, *J. Mol. Struct. (Theochem)* 468 (1999) 13.
- [6] P. Bruneau, *J. Chem. Inf. Comput. Sci.* 41 (2001) 1605.
- [7] A.R. Katritzky, R. Petrukhin, D. Tatham, *J. Chem. Inf. Comput. Sci.* 41 (2001) 679.
- [8] V. Consonni, R. Todeschini, M. Pavan, P. Gramatica, *J. Chem. Inf. Comput. Sci.* 42 (2002) 693.
- [9] G. Krenkel, E.A. Castro, A.A. Toropov, *J. Mol. Struct. (Theochem)* 542 (2001) 107.
- [10] L.B. Kier, L.H. Hall, *Molecular Connectivity in Structure–Activity Analysis*, RSP-Wiley, Chetster UK, 1986.



- [11] J. Ghasemi, S. Shahmirani, E.V. Farahani *Ann. Chim.* 96 (2006) 327.
- [12] J. Ghasemi, Sh. Ahmadi, *Ann. Chim.* in press.
- [13] Subramaniam R, Rao G, Nagesh S. *Der Pharmacia Sinica*, 2011;2 (3): 146-155.
- [14] Sanmati K. J, Sarthak R, Amita J. *Der Pharmacia Sinica*, 2011; 2 (3): 20-30.
- [15] S. Wold, M. Sjostrom, L. Eriksson, *Chemomet. Intell. Lab. Syst.* 58 (2001) 109.
- [16] K. Tang, T. Li, *Chemomet. Intell. Lab. Syst.* 64 (2002) 55.
- [17] T.I. Aksyonova, V.V. Volkovich, I.V. Tetko, *Sys. Anal. Model. Simul.* 43 (2003) 1331.
- [18] B. Hemmateenejad, M.A. Safarpour, F. Taghavi, *J. Mol. Struct. (Theochem)* 635 (2003) 183.
- [19] B. Hemmateenejad, H. Sharghi, M. Akhond, M. Shamsipur, *J. Solution Chem.* 32 (2003) 215.
- [20] E. Soriano, S. Cerdan, P. Ballesteros, *J. Mol. Struct. (Theochem)* 684 (2004) 121.
- [21] L. Xing, R.C. Glen, R.D. Clark, *J. Chem. Inf. Comput. Sci.* 43 (2003) 870.
- [22] L. Xing, R.C. Glen, *J. Chem. Inf. Comput. Sci.* 42 (2002) 796.
- [23] C.M. Chang, *J. Mol. Struct. (Theochem)* 622 (2003) 249.
- [24] Monneret C. *Eur. J. Med. Chem.* 36, 483-493(2001)
- [25] <http://chem2.sis.nlm.nih.gov/chemidplus/ProxyServlet>
- [26] Upadhyay R, Yadav N, Ahmad S. *Advances in Applied Science Research*, 2011, 2 (2): 367-381
- [27] J.N. Miller, J.C. Miller, *Statistics and Chemometrics for Analytical Chemistry*, Prentice Hall, London, 2000.
- [28] Eriksson L, Jaworska J, Worth AP, Cronin MTD, McDowell RM, Gramatica P. *Environ Health Perspect* 2003; 111:1361–1375.
- [29] Waller CL, Bradley MP. *J Chem Inf Comput Sci* 1999; 39:345–355.
- [30] Aires-de-Sousa J, Hemmer MC, Casteiger J. *Anal Chem* 2002; 74:80–90.
- [31] Netzeva TI, Worth AP, Aldenberg T, Benigni R, Cronin MTD, Gramatica P, Jaworska JS, Kahn S, Klopman G, CA Marchant, Myatt G, Nikolova- Jeliaskova N, Patlewicz GY, Perkins R, Roberts DW, Schultz TW, Stanton DT, van de Sandt JJM, Tong W, Veith G, Yang C. The report and recommendations of ECVAM Workshop 52. *ATLA-Altern Lab Anim* 2005; 33:155–173.