*Available on line* www.jocpr.com

***Journal of Chemical and Pharmaceutical Research***

*J. Chem. Pharm. Res., 2011, 3(4): 565-575*

**ISSN No: 0975-7384**
**CODEN(USA): JCPRC5**

# Computational approaches to the predication of the octanol-water partition coefficient ($\mathbf{LogP_{o/w}}$)

**S. Vahdani and Z. Bayat***

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

## ABSTRACT

*A quantitative structure–activity relationship (QSAR) study was performed to develop models those relate the structures of 41 anti-cancer drugs compounds to their n-octanol–water partition coefficients (log Po/w). Among the different constitutional, topological, geometrical, electrostatic and quantum-chemical descriptors that were considered as inputs to the model. The models were constructed using 33 molecules as training set, and predictive ability tested using 11 compounds. Modeling of log Po/w of these compounds as a function of the theoretically derived descriptors was established by multiple linear regression (MLR). The usefulness of the quantum chemical descriptors, calculated at the level of the HF theories using 6-31G* basis set for QSAR study of anti-cancer drugs was examined. A multi-parametric equation containing maximumeight descriptors at HF/6-31G* method with good statistical qualities ($R^2_{train}$=0.893, $F_{train}$=24.93, $Q^2_{LOO}$=0.816,$R^2_{adj}$=0.857,$Q^2_{LGO=0.730}$) was obtained by Multiple Linear Regression using stepwise method.The accuracy of the proposed MLR model was illustrated using the following evaluation techniques: cross-validation, validation through an external test set, and Y-randomisation. The predictive ability of the model was found to be satisfactory and could be used for designing a similar group of compounds.*

**Keywords:** *n*-Octanol–water partition coefficients, Quantitative structure–activity relationship (QSAR), Multiple linear regression (MLR), Hartreefock (HF).

## INTRODUCTION

Doxorubicin is widely used anthracyclines anti-cancer agent. 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 suppressioncan now be reduced by the use of hematopoietic growth factors[1].The *n*-octanol/water partition coefficient is the ratio of the concentration of a chemical in *n*-octanol to that in water in a two-phase system at equilibrium. The logarithm of this coefficient, log $P_{o/w}$, has been shown to be one of the key parameters in quantitative structureactivity/property relationship (QSAR/QSPR) studies. The octanol–water partitioncoefficient is a measure of the hydrophobicity andhydrophilicity of a substance. Hydrophobic "bonding" is actuallynot bond formation at all, but rather the tendency of hydrophobic molecules or hydrophobic parts of moleculesto avoid water because they are not readily accommodatedin the highly ordered hydrogen bonded structure of water[2]. Hydrophobic interaction is favored thermodynamically because of increased entropy of the water molecules thataccompanies the association of non-polar molecules, which squeeze out water. There are some reports about the applications of MLR [3–6] and artificial neural network [7–10] modeling to predict the *n*-octanol/water partition coefficient of anti-cancer drugs. In our previous papers, we reported on the application of QSAR techniques in the development of a new, simplified approach to prediction of compounds properties [11–17]. Experimental determination of log *P*o/w is often complex and time-consuming and can be done only for already synthesized compounds. For this reason, a number of computational methods for the prediction of this parameter have been proposed. In this work a QSAR study is performed, to develop models that relate the structures of a heterogeneous group of 41 drug compounds to their *n*-octanol–water partition coefficients. However, using *in vivo* methods to measure the logarithmic values of partition coefficient drug concentration ratios (log P) in humans is not possible, and to do so in animal models is expensive and time consuming. Finally, the accuracy of the proposed model was illustrated using the following: leave one out, bootstrapping and external test set, cross-validations and Y-randomisation techniques.

## 2. Data and methods

The QSAR model for the estimation of the log $P_{o/w}$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–log $P_{o/w}$ model is generated by the MLR, and statistical analysis.

### 2.1. Data

All log*P*o/w data for all 41 compounds was taken from the literature.The data set was split into a training set (33compounds) and a prediction set (8 compounds). The log $P_{o/w}$ of these compounds are deposited in Journal log as supporting material (see Tables 2). Chemical structure of drugs that illustrated in this study is shown in Table 2.

### 2.2. Molecular descriptor generation

All of the molecules were drawn into the Hyper Chem. The Gaussian 03 and Dragon packages were used for calculating the molecular descriptors(Table 1). 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.DFT method at 6-31G* were applied for optimization of anti-cancer drugs and calculation of many of the descriptors. software hyper Chem and some of the descriptors such as partition coefficient, surface area, hydration energy, and refractivity were calculated through it. The rest of the descriptors were obtained of Gaussian calculations.

A large number of descriptors were calculated by Gaussian packageand Hyperchem software. One way to avoid data redundancy is to exclude descriptors that are highly intercorrelated with each other before performing statistical analysis. The molecular structures were saved by the HIN extension and entered on the DRAGON software for the calculation of the 18 different types of theoretical descriptors for each molecule. They included (a) 0D-constitutional (atom and group counts); (b) 1D-functional groups, 1D-atom centered fragments; (c) 2D-topological, 2DBCUTs, 2D-walk and path counts, 2D-autocorrelations, 2D-connectivity indices, 2D-information indices, 2D-topological charge indices, and 2D-eigenvalue-based indices; and (d) 3D-Randic molecular profiles from the geometry matrix, 3D-geometrical, 3D-WHIM, and 3D-GETAWAY descriptors. A stepwise technique was employed that only one parameter at a time was added to a model and always in the order of most significant to least significant in terms of F-test values. Statistical parameters were calculated subsequently for each step in the process, so the significance of the added parameter could be verified. The goodness of the correlation is tested by the regression coefficient ($R^2$), the F-test and the standard error of the estimate (SEE). The test and the level of significance, as well as the confidence limits of the regression coefficient, are also reported. The squared correlation coefficient, $R^2$, is a measure of the fit of the regression model. Correspondingly, it represents the part of the variation in the observed (experimental) data that is explained by the model.

**Table1. 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 | $E_{GAP}$ |
| | 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 ($\omega=\chi^2/2\ \eta$ ) | Ω |
| | Lowest Unoccupied Molecular Orbital | LUMO | | MullikenlChargeg | MC |
| Chemical properties | Partition Coefficient | Log P | Chemical properties | 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 log *P*o/w to the molecular structure, is an important step to construct predictive models. The genetic algorithm was applied to the input set of 53 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.

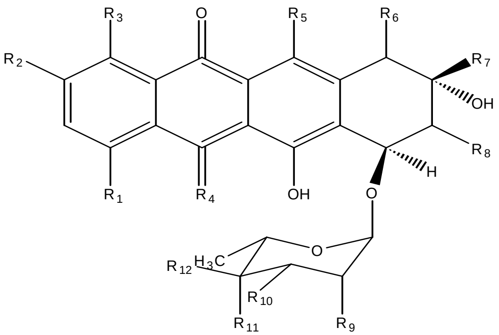
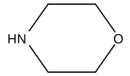
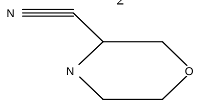


**Table 2.Chemical structures and the corresponding observed and predicted LogPo/w values by the MLR method.**

| N | $R_1$ | $R_2$ | $R_3$ | $R_4$ | $R_5$ | $R_6$ | $R_7$ | $R_8$ | $R_9$ | $R_{10}$ | $R_{11}$ | $R_{12}$ | exp | Pread | Ref |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | $OCH_3$ | H | H | O | OH | H | $COCH_2OH$ | H | H | $NH_2$ | $OH_{ndo}$ | H | 1.27 | 1.39 | 17 |
| 2 | $OCH_3$ | H | H | O | OH | H | $COCH_3$ | H | H | $NH_2$ | OH | H | 1.83 | 1.79 | 17 |
| 3 | H | H | H | O | OH | H | $COCH_3$ | H | H | OH | OH | H | 0.9 | 1.3 | 17 |
| 4 | $OCH_3$ | H | H | O | OH | H | $COCH_2OCH_3$ | H | H | $NH_2$ | OH | H | 1.37 | 0.18 | 18 |
| 5[a] | $OCH_3$ | H | H | O | OH | H | $COCH_3$ | H | H | $N(CH_3)_2$ | OH | H | 1.405 | 1.49 | 18 |
| 6 | $OCH_3$ | H | H | O | OH | H | $COCH_2OH$ | H | H | $NH_2$ | $OCH_3$ | H | 0.94 | 0.96 | 18 |
| 7 | $OCH_3$ | H | H | O | OH | H | $COCH_2OH$ | H | H | $NH_2$ | H | H | 1.83 | 1.41 | 17 |
| 8 | $OCH_3$ | H | H | O | OH | H | $COCH_2OH$ | H | H | $NH_2$ | H | H | 0.68 | 1.69 | 19 |
| 9 | $OCH_3$ | H | H | O | OH | H | $COCH_3$ | H | H | OH | OH | H | 1.34 | 1.41 | 18 |
| 10 | $OCH_3$ | H | H | O | OH | H | $COCH_2OH$ | H | H | $N(CH_3)_2$ | OH | H | 1.56 | 1.65 | 18 |
| 11 | $OC_6H_5$ | H | H | O | OH | H | $COCH_3$ | H | H | $NH_2$ | OH | H | 2.02 | 1.77 | 19 |
| 12 | $OCH_3$ | H | H | O | OH | H | $CHCH_3OH$ | H | H | $NH_2$ | OH | H | 0.62 | 0.49 | 19 |
| 13[a] | $OCH_3$ | H | H | O | OH | H | $COCH_2OH$ | H | H | OH | OH | H | 1.75 | 1.61 | 20 |
| 14 | $OCH_3$ | H | H | NH | OH | H | $COCH_3$ | H | H | $NH_2$ | OH | H | 0.8 | 1.20 | 22 |
| 15 | $OCH_3$ | H | H | O | OH | H | $COCH_2OH$ | H | H | $NH(CH(CN)(CH_2OCH_3))$ | OH | H | 0.92 | 0.60 | 21 |
| 16 | $OCH_3$ | H | H | O | OH | H | $COCH_2OH$ | H | H |  | OH | H | 0.42 | 0.27 | 23 |
| 17 | $OCH_3$ | H | H | O | OH | H | $COCH_2F$ | H | H | $NH_2$ | OH | H | 0.72 | 0.83 | 23 |
| 18 | $OCH_3$ | H | H | O | OH | H | $C(NOH)(CH_3)$ | H | H | $NH_2$ | OH | H | 0.479 | 0.58 | 22 |
| 19[a] | $OCH_3$ | H | H | O | OH | H | $COCH_2OH$ | H | H |  |  |  | 0.286 | 0.95 | 23 |
| 20 | H | H | H | O | OH | H | $COCH_2Br$ | H | F | OH | OH | H | 2.5007 | 2.50 | 19 |
| 21[a] | OH | H | H | O | OH | $COOCH_3$ | $CH_2CH_3$ | H | H | $N(CH_3)_2$ | OH | H | 2.234 | 1.37 | 18 |
| 22[a] | $OCH_3$ | H | H | O | OH | H | $C(NNHCOC_6H_5)(CH_3)$ | H | H | $NH_2$ | OH | H | 1.13 | 1.20 | 17 |
| 23 | $OCH_3$ | H | H | O | OH | H | $COCH_2OCO(CH_2)_3CH_3$ | H | H | $NHCOCF_3$ | OH | H | 2.2 | 2.1 | 17 |

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 24[a] | OH | H | H | O | OH | H | $COCH_3$ | H | H | $NH_2$ | OH | H | 0.74 | 1.02 | 23 |
| 25 | OH | $OCH_3$ | H | O | H | H | $CH_3$ | H | $OCH_3$ | OH | $OCH_3$ | H | 0.871 | 0.81 | 19 |
| 26 | H | H | H | O | OH | H | COCH3 | H | F | OH | OH | H | 0.917 | 0.95 | 19 |
| 27 | H | H | H | O | OH | H | COCH3 | H | F | $OCOCH_3$ | $OCOCH_3$ | H | 1.675 | 1.65 | 22 |
| 28[a] | OH | H | OH | O | OH | $COOCH_3$ | $CH_2CH_3$ | H | H | $NH_2$ | OH | H | 1.916 | 1.78 | 22 |
| 29 | OH | H | H | O | OH | $COOCH_3$ | $CH_2CH_3$ | H | H | $NH_2$ | OH | H | 0.78 | 0.95 | 22 |
| 30 | $OCH_3$ | H | H | O | OH | H | $COCH_3$ | H | OH | $NH_2$ | OH | H | 1.27 | 1.08 | 17 |
| 31 | $OCH_3$ | H | H | O | OH | H | $COCH_3$ | H | H | $NH_2$ | H | H | 0.87 | 0.92 | 18 |
| 32 | H | H | H | O | OH | H | $COCH_2OH$ | F | H | $NH_2$ | OH | H | 1.34 | 1.43 | 22 |
| 33 | $OCH_3$ | H | H | O | OH | H | $COOCO(CH_2)_2NH_2$ | H | F | OH | OH | H | 1.89 | 2.16 | 17 |
| 34 | $OCH_3$ | H | H | O | OH | H | $COCH_2OH$ | H | H | $NH_2$ | OH(exo) | H | 1.85 | 1.71 | 17 |
| 35 | $OCH_3$ | H | H | O | OH | H | $COCH_2OH$ | H | H | $NHCOCF_3$ | OH | H | 1.92 | 1.68 | 17 |
| 36 | $OCH_3$ | H | H | O | OH | H | $COCH_3$ | H | H | OH | $NH_2$ | H | 1.112 | 1.28 | 24 |
| 37 | $OCH_3$ | H | H | O | OH | H | $COCH_3$ | H | H | $N(CH_{3)2}$ | OH | H | 1.23 | 1.63 | 24 |
| 38 | $OCH_3$ | H | H | O | OH | H | $COCH_2OH$ | H | Cl | $NH_2$ | OH | H | 1.45 | 1.07 | 24 |
| 39[a] | H | H | H | O | OH | H | $COCH_2OH$ | H | H | $NH_2$ | OH | H | 0.9 | 1.08 | 24 |
| 40 | $OCH_3$ | H | H | O | OH | H | $COCH_3$ | H | H | $NH_2$ | F | H | 1.53 | 1.59 | 24 |
| 41 | $OCH_3$ | H | H | O | OH | H | $COCH_3$ | H | Br | $NH_2$ | OH | H | 0.87 | 0.91 | 24 |

[a] *test set*

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 log *P*o/w, were used to find subsets of molecular descriptors that provide a good relationship to the log *P*o/w. Given an X-matrix of descriptors data and a log *P*o/w 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 ofcross-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 log *P*o/w[21].

## 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 (PC1and PC2) described 24.39% of the overall variables, as follows: PC1 = 16.79% and PC2 =7.6.%. 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:

$$\mathrm{VIF}=\frac{1}{1-r2} \qquad (1)$$

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 [30]. 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 hat 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:

$$\mathrm{MF}_j=\frac{\beta\sum_{i=1}^{i=n} dij}{\sum_j^m \beta j \sum_i^n \beta ij} \qquad (2)$$

Where *MFj*represents the mean effect for the considered descriptor *j, βj* 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 eight parameters selected by the GA-MLR method**

| Descriptor | Chemical meaning | MFa | VIFb |
|---|---|---|---|
| Constant | Intercept | 0 | 0 |
| $EP_{26}$ | Electrostatic potential 26 | 1.260265 | 1.148737 |
| $NPA_{13}$ | Natural population analysis 13 | -0.15876 | 1.182888 |
| $SAPAC_{22}$ | Surface area approx atomic charg22 | 0.00414 | 1.105926 |
| PW3 | Path/walk3-randic shape index | -0.07902 | 1.284402 |
| Mor16m | 3D-MoRSE-signal16/weighted by atomic masses | 0.005628 | 1.321815 |
| Mor18m | 3D-MoRSE-signal18/weighted by atomic masses | 0.002912 | 1.105745 |
| Mor24m | 3D-MoRSE-signal24/weighted by atomic masses | -0.00102 | 1.226363 |
| G2u | 1st component symmetry directional WHIM index/unweighted | -0.03414 | 1.099806 |

[a]*Mean effect*
[b]*Variation inflation factors*

All descriptors were calculated for the neutral species. The log $P_{o/w}$ is assumed to be highly dependent upon the $EP_{26}$, $NPA_{13}$, $SAPAC_{22}$,PW3 ,Mor16m,Mor18m,Mor24m and G2u. In the present study, the QSAR model was generated using a training set of 33 molecules (Table 2). The test set of 8 molecules (Table 2) with regularly distributed log *P*o/w values was used to assess the predictive ability of the QSARmodels produced in the regression.

## 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 trainingset shown in Table 2.A small number of molecular descriptors ($EP_{26}$,$NPA_{13}$,$SAPAC_{22}$,PW3 ,Mor16m,Mor18m ,Mor24m and G2u) proposed were used to establish a QSAR model. Additional validation was performed on an external data set consisting of 8 drug compounds.

Multiple linear regression analysis provided a useful equation that can be used to predict the log *P*o/w of drug based upon these parameters. The best equation obtained for the Lipophilicity of the drug compounds is

$$\mathrm{LogP}=150.269(\pm 37.396)-12.787(\pm 2.570)EP_{26}+3.882(\pm 0.762)NPA_{13}-0.097(\pm 0.025)SAPAC_{22}+30.446(\pm 9.409)PW31.056(\pm 0.236)Mor16m+0.445(\pm 0.168)Mor18m-1.418(\pm 0.258)Mor24m+34.976(\pm 7.513)G2u$$

$N=41 \quad N_{train}=33 \quad N_{test}=8 \quad R^2_{train}=0.893 \quad F_{train}=24.934 \quad R^2_{test}=0.541$

$F_{test}=-0.045 \quad R^2_{adj}=0.857 \quad Q^2_{LOO}=0.816 \quad Q^2_{LGO}=0.730$

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

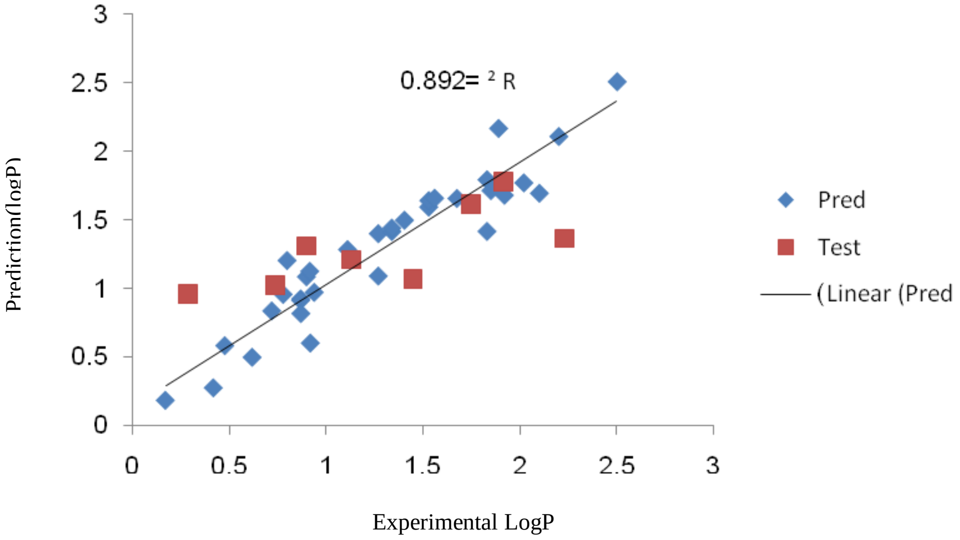


**Figure 1.The predicted versus the experimental LogP by MLR.**

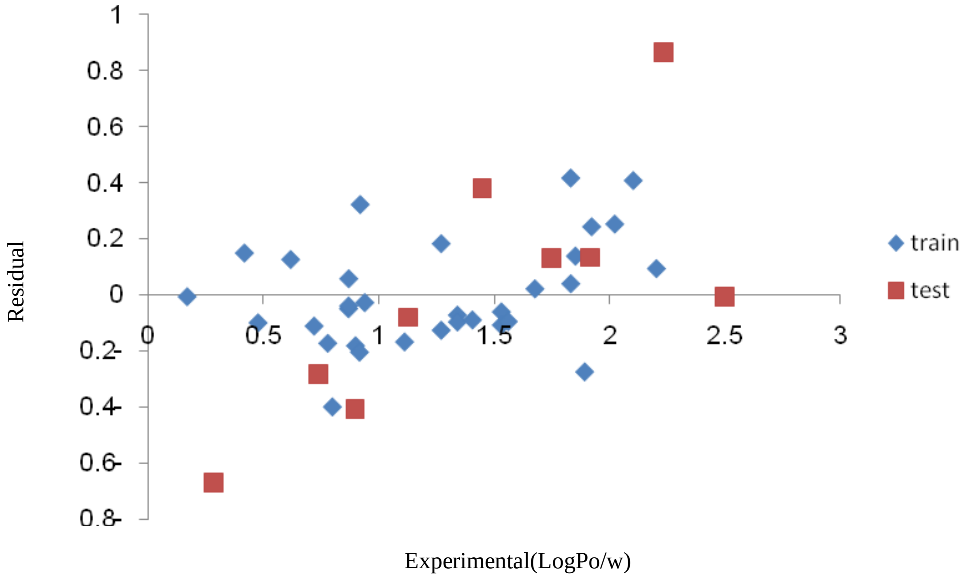


**Figure 2.The residual versus the experimental LogP by GA-MLR**

Also, in order to assess the robustness of the model, the Y-randomisation test was applied in this study [25–28]. The dependent variable vector (LogP) was randomly shuffled and The new QSAR models (after several repetitions) would be expected to have low $R^2$ and $Q^2_{LOO}$ 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 $R^2_{train}$ and $Q^2_{LOO}$ values after several Y-randomisation tests**

| No | $Q^2$ | $R^2$ |
|---|---|---|
| 1 | 0.113284 | 0.472045 |
| 2 | 0.048896 | 0.230775 |
| 3 | 0.003785 | 0.234683 |
| 4 | 0.012186 | 0.31958 |
| 5 | 0.042953 | 0.180091 |
| 6 | 0.042723 | 0.320828 |
| 7 | 0.019219 | 0.21774 |
| 8 | 0.083071 | 0.279033 |
| 9 | 0.005137 | 0.320529 |
| 10 | 0.059051 | 0.166103 |

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

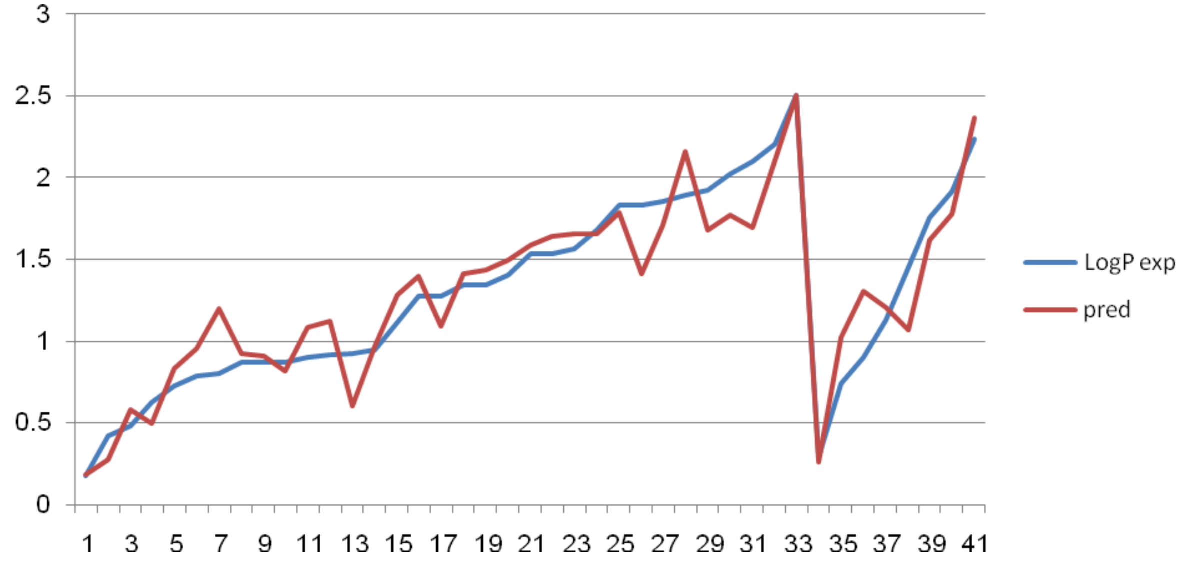
**Table 5. The correlation coefficient existing between the variables used in different MLR and equations with HF/6-31G* method**

| | EP26 | NPA13 | SAPAC22 | PW3 | Mor16m | Mor18m | Mor24m | G2u |
|---|---|---|---|---|---|---|---|---|
| EP26 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| NPA13 | 0.054065 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| SAPAC22 | -0.2344 | -0.12944 | 1 | 0 | 0 | 0 | 0 | 0 |
| PW3 | 0.012593 | 0.236561 | -0.34562 | 1 | 0 | 0 | 0 | 0 |
| Mor16m | -0.33313 | -0.24288 | -0.00361 | -0.23044 | 1 | 0 | 0 | 0 |
| Mor18m | 0.177918 | 0.081215 | -0.13201 | 0.061633 | -0.19956 | 1 | 0 | 0 |
| Mor24m | 0.157028 | 0.378326 | -0.18368 | 0.073222 | -0.35014 | 0.252322 | 1 | 0 |
| G2u | 0.047641 | 0.049852 | -0.17016 | 0.322559 | -0.04335 | 0.01788 | -0.08377 | 1 |

Figure 2 has showed that results were obtained from equation HF/6-31G* to the experimental values.

## 3.3. Interpretation of descriptors

The QSAR developed indicated that electrostatic properties (EP), natural population analysis (NPA), surface area approx atomic charge 22 (SAPAC), Path/walk3-randic shape index(PW3) 3D-MoRSE-signal(16,18,24)/weighted by atomic masses (Mor16m,Mor18m, Mor24m), 1stcomponent symmetry directional WHIM index/unweighted (G2u)drug *n*-octanol/water partition coefficients.Positive values in the regression coefficients indicate that the indicated descriptor contributes positively to the value of log *P*o/w, whereas negative values indicate that the greater the value of the descriptor the lower the value of log*P*o/w.In other words, increasing the $EP_{26}$ and Mor24mwill decrease log *P*o/w and increasing the $NPA_{13,}$$SAPAC_{22}$,PW3,Mor16m,G2u and Mor18m increases extent of log *P*o/w of the anti-cancer drugs. The standardized regression coefficient reveals the significance of an individual descriptor presentedin the regression model.



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

**Figure 3. The comparison between biological activity (log p) using experimental and prediction**

The greater the absolute value of a coefficient, the greater the weight of the variable in the model. Mor16m is the forth descriptor, appearing in the model. It is one of the 3D-molecule representations of structuresbased on electron diffraction (3D-MoRSE) descriptors. The 3D-MoRSE descriptors are derived from infrared spectral simulation using a generalised scattering function [31]. Thisdescriptor was proposed as signal (16, 24)/weighted by the atomicmasses which relates to the atomic masses of the molecule.The Mor(16,24)m displays a positive sign, which indicates that theLogP$_{o/w}$is directly related to this descriptor.The next descriptor is the path/walk 3Randic shape index (PW3), which is one of the topological descriptors. The atomic path/walk indices are defined for each atom as the ratio between the atomic path count and the atomic walk count of the same length. Whereas the number of paths in a molecule is bounded and determined by the molecule’s diameter, the number of walks is unbounded. However, being interested only in quotients, the walk count is terminated when it exceeds the maximum allowed length of the corresponding path [31]. The molecular path/walk indices are defined as the average sum of atomic path/walk indices of equal length. As the path/walk count ratio is independent of molecular size, these descriptors can be considered as shape descriptors. As is apparent from Table 2, the PW3 mean effect has a negative sign which indicates that the LogP$_{o/w}$is inversely related to this descriptor; therefore, increasing the PW3 of molecules leads to a decrease in its LogP$_{o/w}$ values.

## CONCLUSION

In this article, a QSAR study of 41 anti-cancer drugs was performed based on the theoretical molecular descriptors calculated by the DRAGON and 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]Huuskonen J., Salo M., Taskinen J., *J. Pharm. Sci.*, 86, 450—454 (**1997**).

[2] P.J. Taylor, C. Hansch, P.G. Sammes, J.B. Taylor, Comprehensive Medicinal Chemistry, Pergamon Press, Oxford, **1990**.

[3] I. Moriguchi, S. Hirono, I. Nakagome, H. Hirano, *Chem. Pharm. Bull.* 42 (**1994**) 976.

[4] W.M. Meylan, P.H. Howard, J. Pharm. Sci. 84 (**1995**) 83.

[5] V.K. Gombar, K. Enslein, *J. Chem. Inf. Comput. Sci.* 36 (**1996**) 1127.

[6] S.C. Basak, B.D. Gute, G.D. Grunwald, *J. Chem. Inf. Comput. Sci.* 36 (**1996**) 1054.

[7] J.J. Huuskonen, D.J. Livingstone, I.V. Tetko, *J. Chem. Inf. Comput. Sci.* 40 (**2000**) 947.

[8] I.V. Tetko, V.Y. Tanchuk, A.E.P. Villa, *J. Chem. Inf. Comput. Sci.* 41 (**2001**) 1407.

[9] L. Molnar, G.M. Keseru, A. Papp, Z. Gulyas, F. Darvas, *Bioorg. Med. Chem. Lett.* 14 (**2004**) 851.

[10] A.F. Duprat, T. Huynh, G. Dreyfus, *J. Chem. Inf. Comput. Sci*. 38 (**1998**) 586.

[11] J. Ghasemi, S. Shahmirani, E.V. Farahani, *Anal. Chim.*96 (**2006**) 327.

[12] J. Ghasemi, S. Saaidpour, S.D. Brown, *J. Mol. Struct.* (Theochem.)805 (**2007**) 27.

[13] J. Ghasemi, S. Saaidpour, *Chem. Pharm. Bull.*55 (**2007**) 669.

[14] J. Ghasemi, Sh. Ahmadi, *Anal. Chim.*97 (**2007**) 69.

[15] J. Ghasemi, S. Asadpour, A. Abdolmaleki, *Anal. Chim.Acta* 588 (**2007**) 200.

[16] S. Qanei Nassab, Z. Bayat, J. Movaffagh, *J. Chem. Pharm. Res.*, **2011**, 3(1):64-71

[17] Z. Bayat and S. Vahdani, *J. Chem. Pharm. Res*., **2011**, 3(1):93-102

[18] D.C. Young, Computational Chemistry, John Wiley & SonsInc., **2001**.

[19] Pomona college medicinal chemistry project, Claremont, CA91711, LogP database, (C.Hansh and A.Leo), july **1987** edition

[20] E. Frich, P.B. Jones, M.Roed, T.Skovsgaard and N.I.Nissen, *Biochem, pharmacol,* 39(11), 1721-1726(**1990**)

[21] I.Faccheti and A.Vigevani, *pharmacochem.library* 10,138-140(**1987**)

[22] C.Hansh, A.Leo and D.Hokman, American chemical society,Washington(**1995**)

[23] E.M.Action, G.L.Tong, D.L.Taylor, D.G.Streeter,J.A.Fillibi and R.L.Wolgemuth, *J.Med.Chem*, 29(10),2074-2079(**1986**)

[24] Z. Bayat and M. Fakoor Yazdan Abad, *J. Chem. Pharm. Res.*, **2011**, *3(1):48-55*

[25] Z. Bayat, S. Qanei Nassab, *J. Chem. Pharm. Res.*, **2010**, 2(6):306-315

[26] C.j.Coulson, and V.J. Smith, *J.Pharm.Sci*, 69(7),799-801(**1980**)

[27] Hoffman, H.G.Berscheid, D.Borttger, H.H.Sedlacek an H.P.Kraemer,*j.med.chem*,33(1),166-171(**1990**)

[28] E.L.Gabbay, D.Grier, R.E.Fingerle, R.Reimer, R.LevyW.D.Wilson, Biochemistry, 15(10), 2062-2070(**1976**)

[30] G. Espinosa, D. Yaffe, Y. Cohen, A. Arenas, F. Giralt, *J. Chem. Inf. Comput. Sci*. **2000**, 40,859–879.

[31]Todeschini R, Consonni V. Handbook of Molecular Descriptors. Weinheim, Germany: Wiley-VCH, **2000**;1–667.