Available online at www.pelagiaresearchlibrary.com

Pelagia Research Library

Der Chemica Sinica, 2011, 2(6):331-340

Pelagia Research Library
ISSN: 0976-8505
CODEN (USA) CSHIA5

Prediction of the Partition coefficient of Some anti-HIV drugs by Density Functional Theory

Bayat. Z* and Emadiyan. M

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

---

ABSTRACT

*A quantitative structure activity relationship (QSAR) study was performed to develop a model that relates the structures of 26 drug organic compounds to their partition coefficient (log P). Molecular descriptors derived solely from 3D structure were used to represent molecular structures. The compounds are represented by chemical descriptors calculated from their constitutional, geometrical and topological structure, and quantum mechanical wave function. A subset of the calculated descriptors selected using stepwise regression that used in the QSAR model development. Multiple linear regression (MLR) is utilized to construct the linear QSAR model. Stepwise regression was employed to develop a regression equation based on 21 training compounds, and predictive ability was tested on 5 compounds reserved for that purpose. The usefulness of the quantum chemical descriptors, calculated at the level of the DFT theories using 6-31+G** basis set for QSAR study of anti-HIV drugs was examined. The prediction results are in good agreement with the experimental values. A multi-parametric equation containing maximum four descriptors at B3LYP/6-31+G** method with good statistical qualities (R^2_{train}=0.9242, F_{train}=48.768, Q^2_{LOO}=0.8815, R^2_{adj}=0.9052, Q^2_{LGO}=0.8391) was obtained by Multiple Linear Regression using stepwise method.*

Key words : octanol /water partition coefficient; QSAR; MLR,HIV, DFT.

INTRODUCTION

The human immunodeficiency virus (HIV), which has been identified as the causative agent of acquired immune-deficiency syndrome (AIDS), infects many people each day and millions of people have died from the disease. The need for potent, safe, and inexpensive chemotherapeutics is clear, and the therapies must also be effective against mutant strains of HIV which arise from the circumvention of existing anti-HIV treatments. Reverse transcriptase (RT) is a key enzyme of HIV, catalyzing the RNA-depending and DNA-dependent synthesis of double strand viral DNA.

HIV-1 reverse transcriptase (HIV-1 RT) is an attractive target for the drug therapy of AIDS, because it is essential for HIV replication and it is not required for normal host cell replication. One class of RT inhibitors is the nucleoside analogues like 3'-azido-3'deoxythymidine (AZT) and 2',3'-dideoxy-inosine (ddI).Another class of HIV-RT inhibitors is non-nucleoside inhibitors (NNRTIs), which like the nucleoside analogues block reverse transcriptase but have a different mode of inhibition of viral replication [1]. The logP is an important physicochemical parameter for drugs. LogP is probably the most commonly used descriptor of lipophilicity, and is usually interpreted in biological terms as a measure of the ability of the solute to cross lipid membranes, molecules with high LogP values are trapped in the membrane. Therefore, only molecules with intermediate LogP values (e.g.between about 0 and 4) can readily cross membranes (by passive diffusion).

Partition coefficient(p) is defined as the ratio of the solute concentration in the n-octanol phase to the non-ionised solute concentration in the water phase, at equilibrium:

$$P=C_{octanol}/C_{water}$$

Where $C_{octanol}$ is the equilibrium concentration of the solute in the n-octanol phase
C_{water} is the equilibrium concentration of the solute in the water.

P describes the distribution of a compound between two phases-n-octanol and water. It is generally used in its logarithmic form (LogP). A LogP of zero indicates that the solute is equally soluble in the two phases, a negative LogP means that the solute is more soluble in water, and a positive value indicates a greater solubility in the octanol phase [2].

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 structure activity/property relationship (QSAR/QSPR) studies. The octanol–water partition coefficient is a measure of the hydrophobicity and hydrophilicity of a substance. Hydrophobic “bonding” is actually not bond formation at all, but rather the tendency of hydrophobic molecules or hydrophobic parts of molecules to avoid water because they are not readily accommodated in the highly ordered hydrogen bonded structure of water [3]. Hydrophobic interaction is favored thermodynamically because of increased entropy of the water molecules that accompanies the association of non-polar molecules, which squeeze out water. There are some reports about the applications of MLR [4–7] and artificial neural network [8–11] modeling to predict the *n*-octanol/water partition coefficient of anti-HIV drugs. 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 group of 26 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.

Data set and methods

The QSAR model for the estimation of the log $P_{o/w}$ of various anti-HIV 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 ab initio 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.

Data

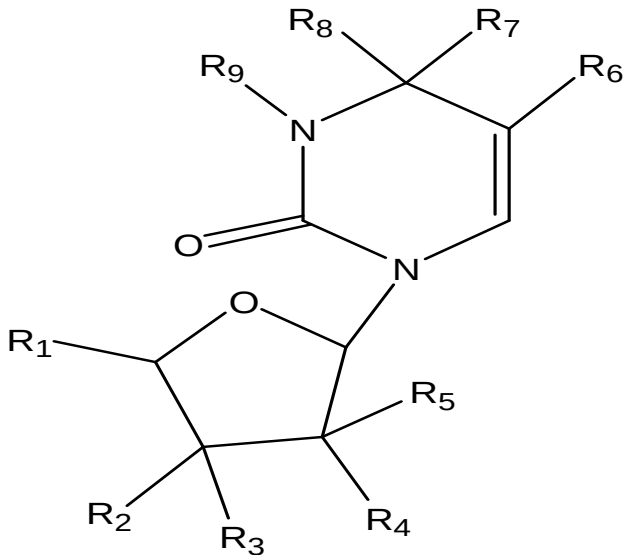
All Log $P_{o/w}$ data for all 26 compounds was taken from the literature [12-21]. The data set was split into a training set (21compounds) and a prediction set (5 compounds). The log $P_{o/w}$ of these compounds is deposited in Journal log as supporting material (see Tables 1). Chemical structure of drugs that illustrated in this study is shown in Table 1.

Molecular Modeling and Theoretical Molecular Descriptors

The derivation of theoretical molecular descriptors proceeds from the chemical structure of the compounds.

The computational chemistry software Hyperchem, Gaussian and Gauss view was used to build the molecules and perform the necessary geometry optimizations. We have chosen descriptors associated with the neutral molecules of drug in our calculations. Some of the descriptors are obtained directly from the chemical structure, e. g. constitutional, geometrical, and topological descriptors. Commonly used descriptors in the QSAR analysis are presented in Table 2. As a result, the 90 theoretical descriptors were calculated for each compound in the data sets (26 compounds). At first anti-HIV drugs were built by Hyperchem software and some of the descriptors such as surface area, hydration energy, and refractivity were calculated through it. The rest of the descriptors were obtained of Gaussian calculations. One way to avoid data redundancy is to exclude descriptors that are highly inter correlated with each other before performing statistical analysis. Reduced multi col linearity and redundancy in the data will facilitate selection of relevant variables and models for the investigated endpoint. Variable-selection for the QSAR modeling was carried out by stepwise linear regression method. 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.

Table 1.Chemical structures and the corresponding observed and predicted LogPo/w values by the MLR method



No	R^1	R^2	R^3	R^4	R^5	R^6	R^7	R^8	R^9	Exp.	Pred.	Ref.
1	OH	OH	H	H	OH	H	NH_2	H	-	-2.51	-2.45406	16
2	CH_2OH	OH	H	H	OH	H	O	-	H	-1.98	-1.65703	17
3	CH_2OH	OH	H	H	H	H	NH_2	H	-	-1.9	-1.70849	18
4	CH_2OH	OH	H	H	H	H	O	-	H	-1.62	-1.28941	18
5	CH_2OH	H	H	H	H	H	NH_2	H	-	-1.3	-1.22514	19
6	CH_2OH	OH	H	H	H	F	O	-	H	-1.16	-1.53046	20
7	CH_2OH	H	H	H	H	F	NH_2	H	-	-1.09	-0.82872	21
8	CH_2OH	H	-	-	H	H	O	-	H	-1.07	-0.88842	19
9	CH_2OH	F	H	H	H	H	NH_2	H	-	-0.89	-1.18133	21
10	CH_2OH	H	-	-	H	CH_3	O	-	H	-0.72	-0.65246	22
11	CH_2OH	N_3	H	H	H	H	NH_2	H	-	-0.7	-0.88171	21
12	CH_2OH	F	H	H	H	H	O	-	H	-0.49	-0.95754	21
13	CH_2OH	OH	H	H	H	CF_3	O	-	H	-0.46	-0.32962	20
14	CH_2OH	N_3	H	H	H	H	O	-	H	-0.32	-0.16227	21
15	CH_2OH	F	H	H	H	CH_3	O	-	H	-0.28	-0.4455	21
16	CH_2OH	H	-	-	H	H	S	-	H	-0.28	-0.52708	21
17	H	H	H	H	H	F	O	-	H	-0.27	-0.31041	20
18	CH_2OH	H	H	H	H	H	S	-	H	-0.21	-0.43117	21
19	CH_2OH	H	-	-	H	CH_3	S	-	H	0.09	0.203214	21
20	CH_2OH	OH	H	H	H	CH_2CH_2Br	O	-	H	0.33	0.290381	23
21	CH_2OH	N_3	H	H	H	CH_3	S	-	H	1.07	1.207246	21
22	CH_2OH	OH	H	H	OH	CH_3	NH_2	H	-	-2.01	-2.06099	21
23	CH_2OH	H	-	-	H	H	NH_2	H	-	-1.55	-1.43571	24
24	CH_2OH	OH	H	H	H	CH_3	O	-	H	-0.93	-0.75239	18
25	CH_2OH	OH	H	H	H	Br	O	-	H	-0.29	-0.84905	20
26	CH_2OH	N_3	H	H	H	CH_3	O	-	H	0.05	0.247663	25

Table 2.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 [η=1/2 (HOMO+LUMO)]	H
	Natural Population Analysis	NPA		Softness (S=1/ η)	S
	Electrostatic Potentialc	EP		Electro negativity [χ= -1/2 (HOMO–LUMO)]	*X*
	Highest Occupied Molecular Orbital	HOMO		El Electro philicity (ω=χ^2/2 η)	Ω
	Lowest Unoccupied Molecular Orbital	LUMO		MullikenlCharge	MC
Chemical properties	Partition Coefficient	Log P	Chemical properties	Molecule surface area	SA
	Mass	M		Hydration Energy	HE
	Molecule volume	V		Refractivity	REF

RESULTS AND DISCUSSION

The software package used for conducting MLR analysis was Spss 16. 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 2. After regression analysis, a few suitable models were obtained among which the best model was selected and presented in Eq.1. A small number of molecular descriptors (SAP, LogP, δ_{C2},δ_{N1}) proposed were used to establish a QSAR model. Additional validation was performed on an external data set consisting of 5 organic 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 solubility of the drug compounds is:

LogP=-5.56836(±0.59542)-0.99094(±0.28312)**MC2**+0.01314(±0.00165)**SAP**
+0.044073(±0.0886)**LogP**+ 0.04089(±0.00817)**SAPAC3** Eq.1

N=26 $\mathbf{N_{train}}$=21 $\mathbf{N_{test}}$=5 $\mathbf{R^2_{train}}$= 0.924197 $\mathbf{F_{train}}$=48.76838 $\mathbf{R^2_{test}}$=0.86971
$\mathbf{F_{test}}$=0 $\mathbf{R^2_{adj}}$= 0.905246 $\mathbf{Q^2_{LOO}}$=0.881488 $\mathbf{Q^2_{LGO}}$=0.839088

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.

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 Figure2.

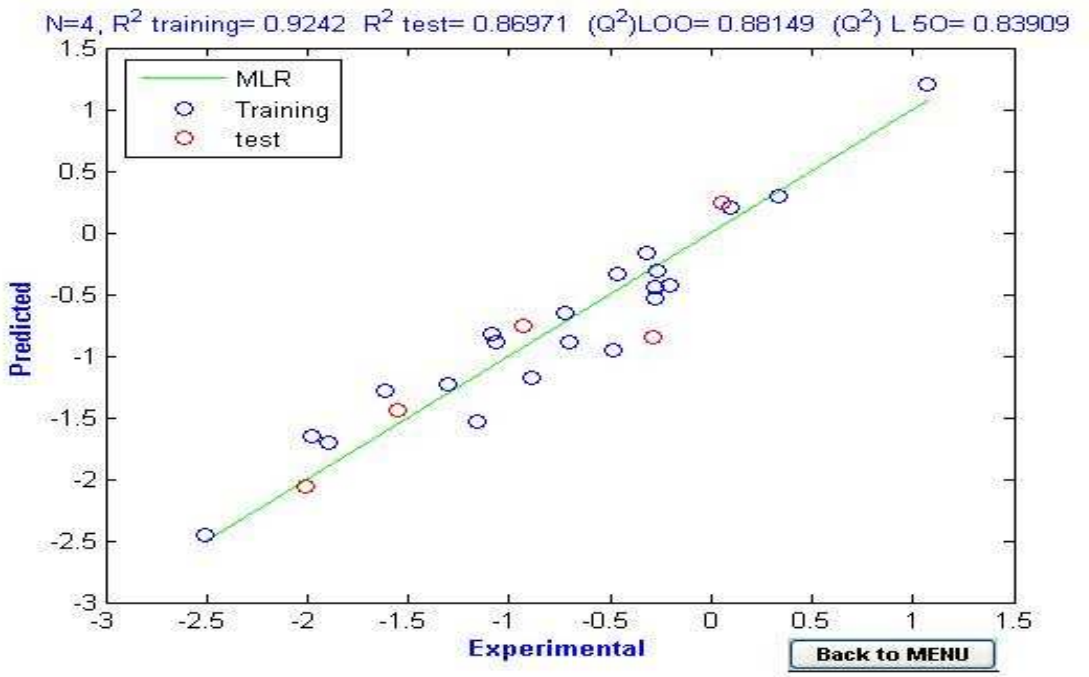


Figure 1.The predicted versus the experimental LogP by MLR.

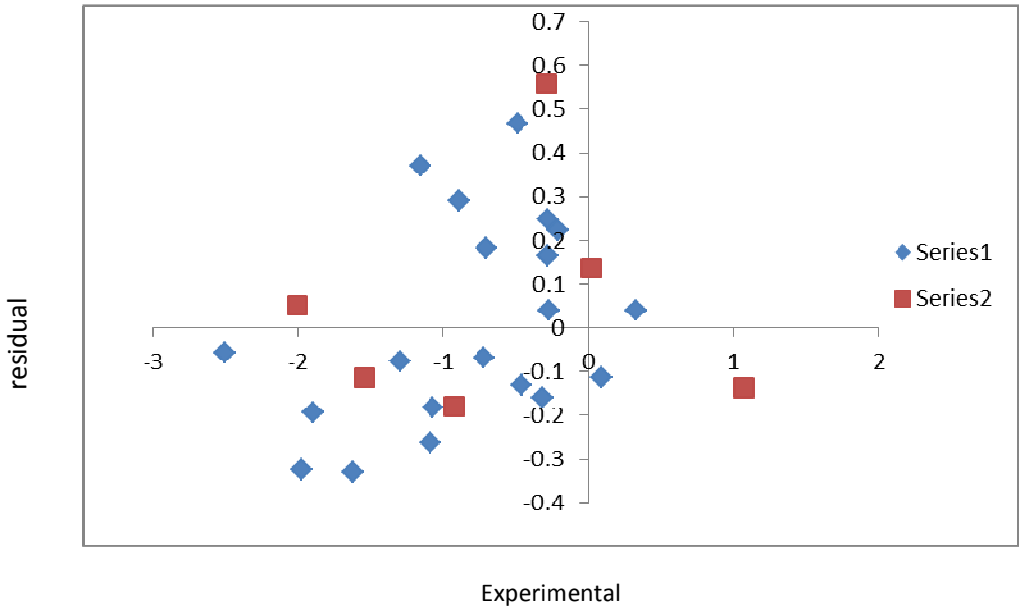


Figure 2.The residual versus the experimental LogP by MLR. (See colour version of this figure online at *www.informahealthcare.com/enz)*

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 41.24% of the overall variables, as follows: PC1 = 26.04% and PC2 =15.2 %. 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-col linearity between the above seven descriptors were detected by calculating their variation inflation factors (VIF), which can be calculated as follows:

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

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 [22].

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:

$$MF_j=\frac{\beta \sum_{i=1}^{i=n} dij}{\sum_j^m \beta j \sum_i^n \beta ij} \quad (2)$$

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 five parameters selected by the MLR method.

Descriptor	Chemical meaning	MF_a	VIF_b
Constant	Intercept	0	0
MC_2	Mulliken charge	0.03303	1.27729
SAP	Surface area approx	0.931501	1.221225
LogP	Partition coefficient	-0.0572	1.21832
$SAPAC_3$	Surface area approx atomic contribution3	0.092665	1.080532

[a]Mean effec
[b]Variation inflation factors

Also, in order to assess the robustness of the model, the Y-randomisation test was applied in this study [23–24]. 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

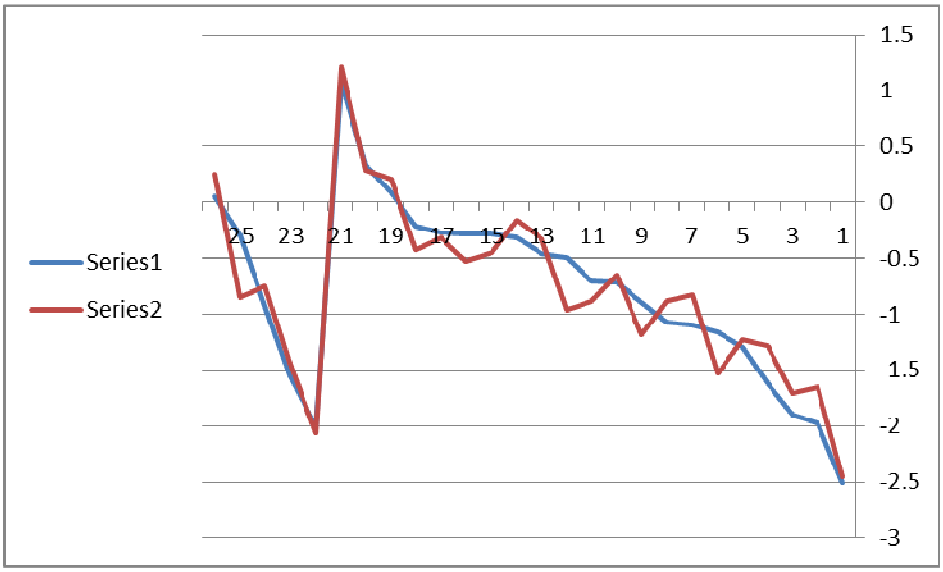
No	Q^2	R^2
1	0.16196	0.121513
2	0.09857	0.077405
3	0.067934	0.122382
4	0.249642	0.091604
5	0.04081	0.314066
6	0.084249	0.017966
7	0.022438	0.245614
8	0.068989	0.341942
9	0.036214	0.370282
10	0.008476	0.286468

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 b3lyp/6-31+G method.**

	MC_2	**SAP**	**LogP**	**$SAPAC_3$**
MC_2	1	0	0	0
SAP	-0.40125	1	0	0
LogP	-0.45311	0.364457	1	0
$SAPAC_3$	0.263683	-0.29924	-0.24723	1

Figure 3 has showed that results were obtained from equation B3LYP/6-31+G** to the experimental values.



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

CONCLUSION

Predictive QSAR model which is based on molecular descriptors is proposed in this study to correlate the LogP of drug compounds. In this article, a QSAR study of 26 anti-HIV drugs was performed based on the theoretical molecular descriptors calculated by the GAUSSIAN software and selected. Application of the developed model to a testing set of 5 compounds demonstrates that the new model is reliable with good predictive accuracy and simple formulation. We have developed here a useful QSAR equation derived from theoretical descriptors associated with LogP. A MLR is successfully presented for prediction LogP of various drug compounds with diverse chemical structures using a linear quantitative structure– activity relationship. A model with high statistical quality and low prediction errors was obtained. The model could predict the LogP of the drug compounds accurately. Development of quantitative structure property/activity relationships (QSPR/QSAR) on theoretical descriptors is a powerful tool not only for prediction of the chemical, physical and biological properties/activities of compounds, but also for deeper understanding of the detailed mechanisms of interactions in complex systems that predetermine these properties/activities.

REFERENCES

[1]KobraZarei and MortezaAtabati *Journal of the Chinese Chemical Society*, **2009**, *56*, 206-213.
[2] School of pharmacy and chemistry , Liverpool john moores university .
[3] P.J. Taylor, C. Hansch, P.G. Sammes, J.B. Taylor, Comprehensive Medicinal Chemistry, Pergamon Press, Oxford, **1990**.
[4] I. Moriguchi, S. Hirono, I. Nakagome, H. Hirano, *Chem. Pharm. Bull*. 42 (**1994**) 976.
[5] W.M. Meylan, P.H. Howard, *J. Pharm. Sci*. 84 (**1995**) 83.
[6] V.K. Gombar, K. Enslein, *J. Chem. Inf. Comput. Sci.* 36 (**1996**) 1127.
[7] S.C. Basak, B.D. Gute, G.D. Grunwald, *J. Chem. Inf. Comput. Sci*. 36 (**1996**) 1054.
[8] J.J. Huuskonen, D.J. Livingstone, I.V. Tetko, *J. Chem. Inf. Comput. Sci*. 40 (**2000**) 947.
[9] I.V. Tetko, V.Y. Tanchuk, A.E.P. Villa, *J. Chem. Inf. Comput. Sci*. 41 (**2001**) 1407.
[10] L. Molnar, G.M. Keseru, A. Papp, Z. Gulyas, F. Darvas, *Bioorg. Med. Chem. Lett.* 14 (**2004**) 851.
[11] A.F. Duprat, T. Huynh, G. Dreyfus, *J. Chem. Inf. Comput. Sci*. 38 (**1998**) 586.
[12] H.Farghali,L.Novotny,M.Ryba,J.Berank and I. Janku. *Bio chem.Pharmacol*.33(4),655-662(**1984**).
[13] N.El,Tayar,R.-S.Tasi,P.Vallat,C.Altomare andB.Testa. *J.Chromatogr*.556,181-194(**1991**).
[14] L.Novotny,H.Farghali,M.Ryba,I.Janku and J.Beranek. *Cancer Chermother. Pharmacol*. 13(3).195-199(**1984**).
[15] J.Balazarini,M.Cools and E.DeClerq. *Biochem.Biophys.Res.Commun*.158(2),413-422 (**1989**).
[16] Pomona College Medicinal Chemistry Project, Claremont, CA 91711,Logp Database. (C,Hansch and A.Leo), July **1987** Edition.
[17] http://chem.sis.nlm.nih.gov/chemidplus/chemidlite.jsp.
[18] S.L.Glym and M.Yazdanian. *J.Pharm.Sci*.87(3),306-310(**1998**).
[19] C.Meier,A.Lomp,A.Meerbach and P.Watzler. *J.Med.Chem*.45(23),5157-5172(**2002**).
[20] A.P.Cheung and D.Kenney. *J.Chromatogr*. 506,119-131(**1990**).
[21] H.ford,C.L.Merski and J.A.kelley. *J.liq.chromatogr*.14(18),3356-3386(**1991**).

[22] G. Espinosa, D. Yaffe, Y. Cohen, A. Arenas, F. Giralt, *J. Chem. Inf. Comput. Sci*. **2000**, 40,859–879.
[23] C.Hansh, A.Leo and D.Hokman, American chemical society,Washington(**1995**).
[24] 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**).