Available online at www.scholarsresearchlibrary.com

Scholars Research Library

***J. Comput. Method. Mol. Design*, 2011, 1 (1): 14-25**
(http://scholarsresearchlibrary.com/archive.html)

2D-QSAR analysis on 4-Flouro-2-Cyanopyrrolidine derivatives as DPP-IV Inhibitors

Sanmati K. Jain[*1], S. Mallick[2], R. Dubey[3], S. Nag[1] and A. Yadav[1]

*[*1]SLT Institute of Pharmaceutical Sciences, Guru Ghasidas Vishwavidyalaya, Bilaspur (CG)*
[2]Royal College of Pharmacy and Health Sciences, Andhapasara Road, Berhampur, Orissa, India
[3]Institute of Pharmacy, RITEE, Mandir Hasaud, Chhatauna, Raipur, Chhattisgarh, India

---

ABSTRACT

A quantitative structure activity relationship study was performed on a series of 4-flouro-2-cyanopyrrolidine derivatives possessing Dipeptidyl Peptidase IV (DPP IV) inhibitory activity for establishing quantitative relationship between biological activity and their physicochemical / structural properties. Several statistical regression expressions were obtained using stepwise multiple linear regression (MLR), partial least squares regression (PLSR), partial component regression (PCR) analysis. Three statistical significant models were generated (r^2 = 0.830, 0.893, 0.875 for model 1, 2 and 3 respectively) indicating that biological activity is influenced by the descriptors, T_C_N_6, T_C_N_3, MMFF_8, slogp etc.

Key words: 2D-QSAR, Dipeptidyl Peptidase IV inhibitors, 4-flouro-2-cyanopyrrolidine derivatives.

INTRODUCTION

Type 2 diabetes mellitus (T2DM), also known as non-insulin-dependent diabetes mellitus (NIDDM) or adult onset diabetes, is a major worldwide health concern and accounts for more than 90% of cases in industrialized nations [1,2]. T2DM is characterized by fasting and postprandial hyperglycemia and relative insulin insufficiency. If left untreated, hyperglycemia may cause long-term microvascular and macrovascular complications, such as nephropathy, neuropathy and retinopathy, and atherosclerosis. This disease causes significant morbidity and mortality at considerable expense to patients, their families and society [3]. Thus, there is an imperative need for potent therapeutic approaches for glycemic control. Treatment of NIDDM currently centers on biguanides, sulphonylureas, α-glucosidase inhibitor, insulin sensetizer, and insulin secretagogues [4]. Now the discovery of small-molecule inhibitors of the enzyme

dipeptidyl peptidase-IV (DPP-IV) is providing a major field for the development of new drugs to treat type 2 diabetes.

Dipeptidyl peptidase IV (also known as CD 26) is a membrane-associated serine protease which is widely distributed in mammalian tissues and body fluids [5,6]. It modulates the biological activity of several peptide hormones, chemokines and neuropeptides by specifically cleaving X-proline or X-alanine dipeptides from the N-terminus. DPP-IV cleaves and inactivate the incretin hormone, glucagon like peptide (GLP-1), which is an important stimulator of insulin secretion [7,8]. Inhibition of DPP-IV increases the level of circulating GLP-1 and thus increases insulin secretion [9,10], which in turn control the glucose level in type 2 diabetes. 4-Flouro-2-cyanopyrrolidine derivatives [11] reported to have dipeptidyl peptidase IV inhibitory activity. No Quantitative structure activity relationship (QSAR) study was reported in the literature on 4-flouro-2-cyanopyrrolidine derivatives. Therefore, QSAR analysis was performed for establishing quantitative relationship between biological activity of 4-flouro-2-cyanopyrrolidine derivatives and their physicochemical / structural properties. The aim of the present work is to generate best predictive and validated QSAR model.

QSAR analysis on 4-flouro-2-cyanopyrrolidine derivatives (Figure-1) was performed for DPP-IV inhibitory activity ($\log 1/IC_{50}$) as dependent and different physicochemical / structural parameters as independent variables using the software, QSAR Plus {Molecular Design Suite (MDS)} [12].

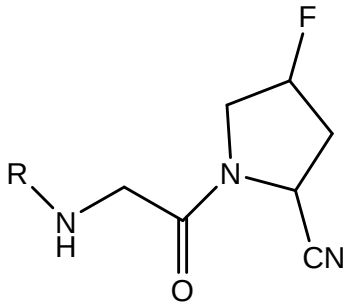


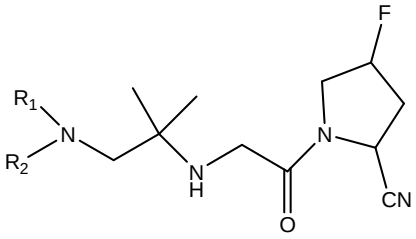
Figure-1: General structure of 4-flouro-2-cyanopyrrolidine derivatives selected

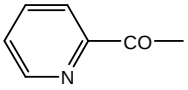
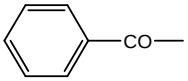
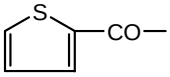
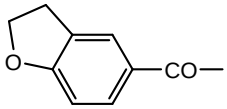
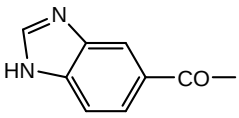
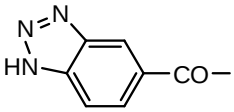
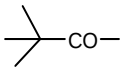
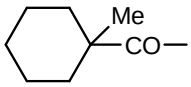
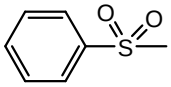
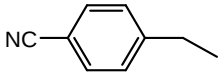
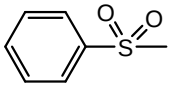
MATERIALS AND METHODS

Experimental

***Data set*:** A dataset of 40 molecules has been taken from the literature[11]. Selected data set, their biological activity are shown in Table-1, 2 and 3. Biological data's represented as IC_{50} (nM) were converted into log ($1/IC_{50}$) for computational work.

Table -1: General structure of 4-fluoro-2-cyanopyrrolidine derivatives and their biological activities (data set of 15 molecules)



S.No.	Compound	R_1	R_2	IC_{50} (nM)	Log(1/IC_{50})
1	10a		H	8.2	8.086
2	10b		H	4.5	8.347
3	10c		H	5.4	8.268
4	10d		H	2.9	8.538
5	10e		H	5.7	8.244
6	10f		H	1.5	8.824
7	10g		H	5.8	8.237
8	10h		H	13	7.886
9	10i	$(4\text{-Cl-Ph})_2$CHCO-	H	109	6.963
10	10j		H	75	7.125
11	10l		H	39	7.409
12	11a		Me	31	7.509
13	11b	Et	Et	252	6.599

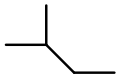
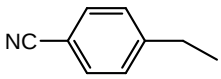
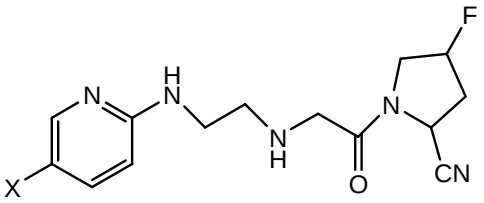
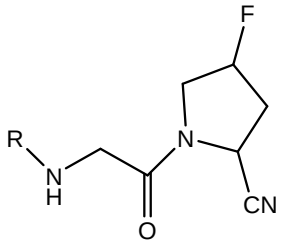
14	11c		PhCO-	7.3	8.137
15	11d		PhCO-	6.2	8.208

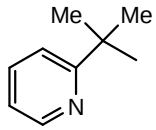
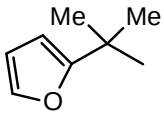
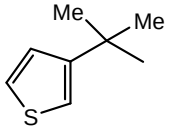
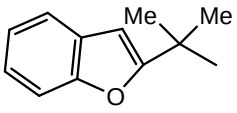
Table - 2: General structure of the compounds of 4-fluoro-2-cyanopyrrolidine derivatives and their biological activities (data set of 4 molecules)

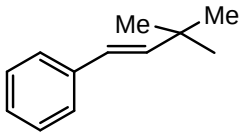
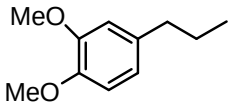
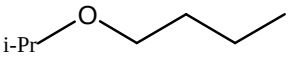
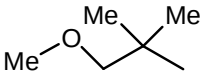
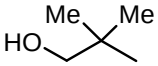
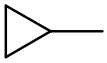
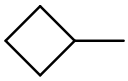
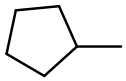
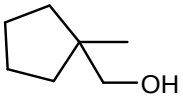
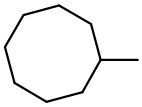
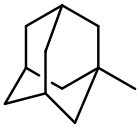
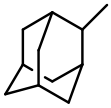
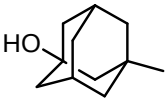


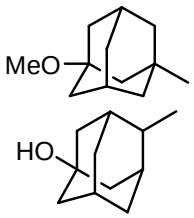
S.No.	Compound	X	IC_{50} (nM)	Log(1/ IC_{50})
16	2a	CN	1.1	8.959
17	2b	H	2.8	8.553
18	2c	Cl	2.7	8.569
19	2d	$CONH_2$	2.4	8/620

Table - 3: General structure of the compounds of 4-fluoro-2-cyanopyrrolidine derivatives and their biological activities *(data set of 21 molecules)*



S.No.	Compound	R	IC_{50} (nM)	Log(1/ IC_{50})
20	12a		22	7.658
21	12b		8.6	8.066
22	12c		12	7.921
23	12d		88	7.056

24	12e		16	7.796
25	12f		27	7.569
26	12g	t-Bu-	2.9	8.538
27	12h	i-Pr-	7.8	8.108
28	12i		8.7	8.060
29	12j		13	7.886
30	12k		4.6	8.337
31	12m		33	7.481
32	12n		3.1	8.509
33	12o		2.6	8.585
34	12p		3.3	8.481
35	12q		3.3	8.481
36	12r		4.1	8.387
37	12s		3.1	8.509
38	12t		3.1	8.509
39	12u		8.3	8.081

40	12v		5.3	8.276

***QSAR Analysis*:** Structure of the compounds of selected series were drawn in the 2D Draw application option of QSAR Plus and then exported to QSAR Plus window. Energy minimizations of the compounds were done by using UFF force field method until the root mean square (rms) gradient reached 0.01 kcal/mol A° before they were undertaken for 2D QSAR studies, followed by batch optimization. Number of descriptors (physicochemical, alignment independent and atom type independent) was calculated after optimization of the data set molecules. Calculated descriptors and biological activity were taken as independent and dependent variables respectively. Various types of physicochemical descriptors were calculated: Individual (H-Acceptor count, H-Donor count , X logP, SMR, polarizablity, etc.), retention index (Chi), atomic valence connectivity index (ChiV), Path count, Chi chain, Chiv chain, Chain Path Count, Cluster, Path cluster, Kapa, Element count (H, N, C, S, O, Cl, Br, I), Estate numbers (SsCH3 Count, SdCH2 Count, StCH count etc.), Estate contribution (SsCH3-index., SdCH2-index, StCH index) and Polar surface area.

More than 200 alignment independent descriptors were also calculated using the following attributes. A few examples are T_2_O_7, T_2_N_5, T_2_2_6, T_C_O_1, T_O_Cl_5 etc.

Structural descriptors	Selected Attributes
*Topological	2
Range	3
Min – 0	T (any)
Max. - 7	C
	N
	O
	F
	S
	Cl
	Br
	I

Generation of training and test set of compounds: In order to evaluate the QSAR model externally, data set was divided into training and test set using Random selection method, Manual data selection method and Sphere Exclusion methods. Training set is used to develop the QSAR model for which biological activity data are known. Test set is used to challenge the QSAR model developed based on the training set to assess the predictive effectiveness of the model which is not included in model generation.

Random selection: In order to construct and validate the QSAR models, both internally and externally, the data sets were divided into training [90%-60% (90%, 85%, 80%, 75%, 70%, 65% and 60%) of total data set] and test sets [10%-40% (10%, 15%, 20%, 30%, 35% and 40%) of total data set] in a random manner. 10 trials were run in each case.

Manual data selection: Whole range of activities was sorted through ascending & descending order and every 4^{th}, 5^{th}, 6^{th},7^{th}, 8^{th}, 9^{th} and 10^{th} compound assigned to the test set.

Sphere Exclusion method: In this method initially data set were divided into training and test set using sphere exclusion method. In this method dissimilarity value provides an idea to handle training and test set size. It needs to be adjusted by trial and error until a desired division of training and test set is achieved. Increase in dissimilarity value results in increase in number of molecules in the test set.

Statistical computation: The different statistical methods which were used to generate QSAR models are stepwise multiple linear regression (MLR), partial least squares regression (PLSR), partial component regression (PCR). Following statistical parameters were considered to select the statistical significance of QSAR models: squared correlation coefficient (r^2), F-test (F-test for statistical significance of the model), cross-validated squared correlation coefficient (q^2) and predicted correlation coefficient.

RESULTS AND DISCUSSION

When all the 40 molecules of the selected series were subjected to various regression analysis, the following significant QSAR models with equations were obtained for DPP-IV inhibitory activity (Table-4).

Table-4: List of predictive QSAR models generated from various regression methods

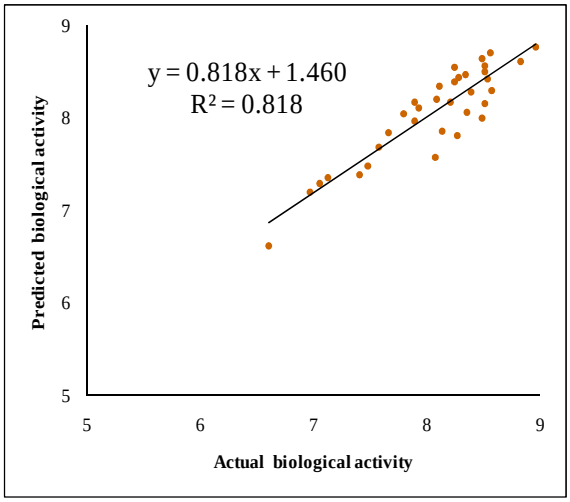
Model No.	Method	Equation
01	Random 85% / Trial-9 / MLR	pIC_{50} = -1.411 MMFF_8 - 0.454 slogp + 0.214 T_C_N_6 - 0.177 T_T_O_2 - 0.326MMFF_22 - 0.348 MMFF_38 n = 34 Degree of freedom = 27 F test = 20.324 r^2 = 0.818 q^2 = 0.656 pred_ r^2= 0.708 r^2se = 0.260 q^2 se = 0.357 pred_ r^2se = 0.256
02	Manual / Trial-2 / PLS	pIC_{50} = - 1.612 MMFF_8 - 0.342 slogp + 0.073 T_C_N_3 - 0.141 T_T_O_2 - 3.293 chi3chain + 0.052 SssCH2E-Index - 0.437 T_3_N_7 + 10.278 n = 35 Degree of freedom = 31 F test = 50.243 r^2= 0.829 q^2 = 0.640 pred_ r^2 = 0.788 r^2se = 0.236 q^2se = 0.342 pred_ r^2se = 0.236
03	Manual / Trial-3 / PLS	pIC_{50} = -1.593 MMFF_8 - 0.310 slogp + 0.117 T_C_N_3 - 0.156 T_C_O_3 - 3.565 chi3chain + 0.056 SssCH2count - 0.370 MMFF_38 + 10.0116 n = 35 Degree of freedom = 31 F test = 56.999 r^2= 0.846 q^2 = 0.623 pred_ r^2 = 0.734 r^2se = 0.221 q^2se = 0.346 pred_ r^2se = 0.296

In the above QSAR models, n is the number of data points used in the model, pred_ r^2 is the predicted correlation coefficient and se is the standard error of estimate (smaller is better).

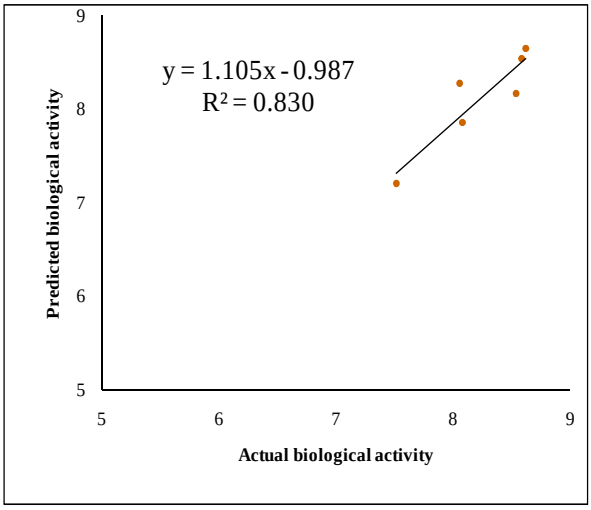
Model-01 explains 82% (r2 = 0.82) of the total variance in the training set as well as it has internal (q2) and external (pred_r2) predicative ability of 65% and 70% respectively. Model-02 explains 83% (r2= 0.83) of the total variance in the training set as well as it has internal (q2) and external (pred_r2) predicative ability of 64% and 79% respectively. Model-03 explains 85% (r2=

0.85) of the total variance in the training set as well as it has internal (q2) and external (pred_r2) predicative ability of 62% and 73% respectively.

Graphs were plotted between the actual and the predicted biological activities of training and test set for model 1, 2 and 3 shown in Figure-2, 3, and 4 respectively.

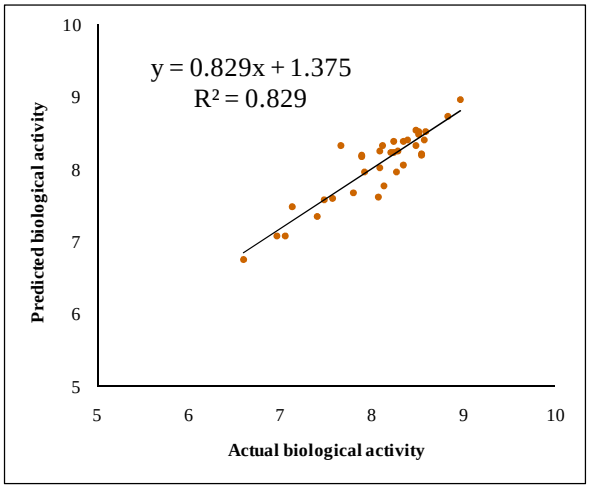


(Training set)

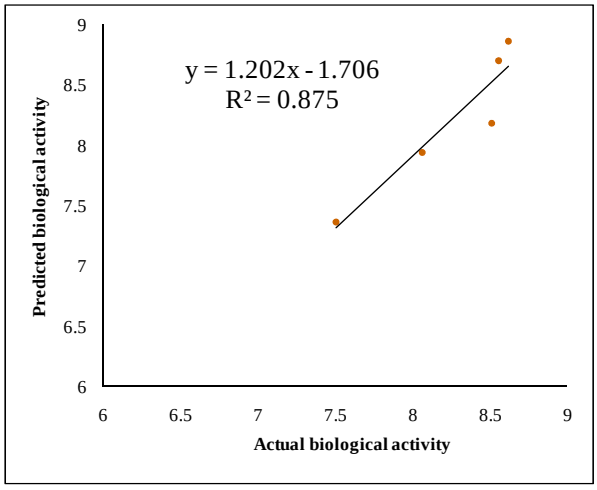


(Test set)

Figure-2: Graph between actual and predicted biological activity of training and test set for Model-1.



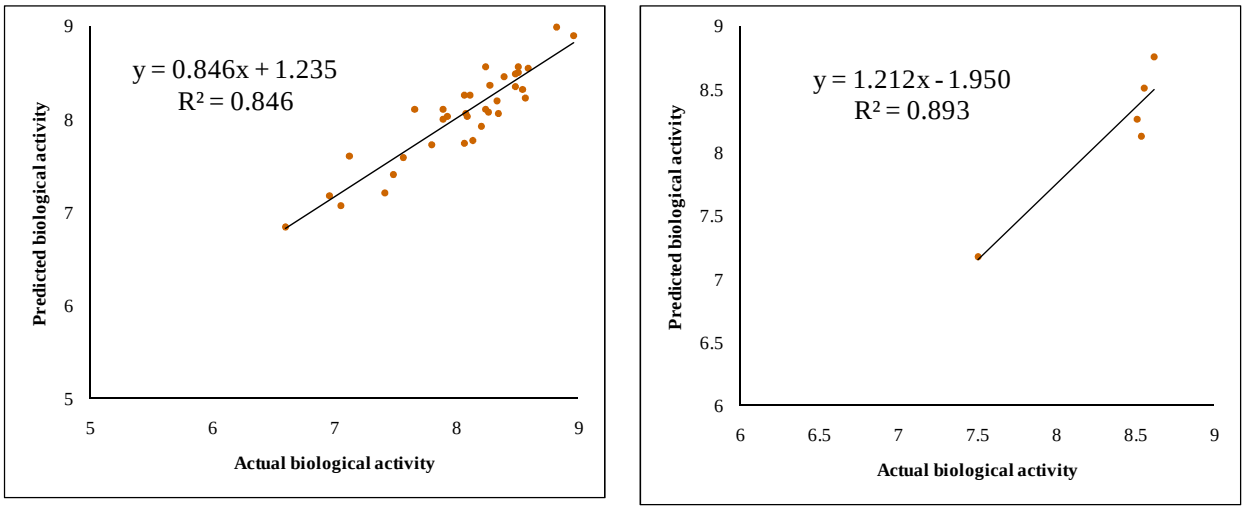
(Training set)



(Test set)

Figure-3: Graph between actual and predicted biological activity of training and test set for Model-2.

The three models with their r^2 value produced from actual vs. predicted activity graph is shown in Table-5. That table represents the best 3 models which are having $r^2 > 0.83$, indicating their predictive ability in predicting the activity of the test set molecules.



(Training set) (Test set)

Figure-4: Graph between actual and predicted biological activity of training and test set for Model-3.

Table-5: List of the best 3 models with their r^2 value produced from actual versus predicted activity graph for different models

MODEL	r^2	
	For Training set	For Test set
01	0.818	0.830
02	0.829	0.875
03	0.846	0.893

Table-6: Correlation matrix for descriptors used in model-01

	MMFF_8	slogp	T_C_N_6	T_T_O_2	MMFF_22	MMFF_38
MMFF_8	1					
slogp	-0.00974	1				
T_C_N_6	-0.07162	0.493321	1			
T_T_O_2	-0.23511	0.048164	0.205274	1		
MMFF_22	-0.04352	-0.2008	-0.29211	-0.16371	1	
MMFF_38	-0.10381	-0.25094	0.054381	-0.28636	-0.07228	1

Table-7: Correlation matrix for descriptors used in model-02

	MMFF_8	*slogp*	*T_C_N_3*	*T_T_O_2*	*chi3chain*	*SssCH2E- index*	*T_3_N_7*
MMFF_8	1						
slogp	-0.01668	1					
T_C_N_3	0.070598	0.154441	1				
T_T_O_2	-0.22536	0.060939	0.176035	1			
chi3chain	-0.0527	0.164791	-0.2157	-0.19595	1		
SssCH2E-index	0.01094	0.105143	-0.09282	-0.21441	-0.11157	1	
T_3_N_7	0.469697	0.215082	0.367109	-0.0707	-0.0527	-0.10865	1

The correlation matrix for model-1, 2 and 3 are given in Table-6, 7 and 8. The correlation matrix used to see the mutual correlation among the parameters used in the model.

Table-8: Correlation matrix for descriptors used in model-03

	MMFF_8	slogp	T_C_N_3	T_C_O_3	chi3chain	SssCH2count	MMFF_38
MMFF_8	1						
slogp	-0.01681	1					
T_C_N_3	0.068665	0.150063	1				
T_C_O_3	-0.26004	0.126602	0.320663	1			
chi3chain	-0.04222	-0.20727	-0.2312	-0.18115	1		
SssCH2count	0.026458	0.038826	0.016653	-0.09518	-0.02458	1	
MMFF_38	-0.08843	-0.19159	0.209774	-0.27883	-0.0616	-0.18659	1

The correlation matrix shows that descriptors have low inter-correlation value. The contribution charts representing contribution of various descriptors in biological activity for model 1, 2 and 3 are shown in Figure-5, 6, and 7 respectively.

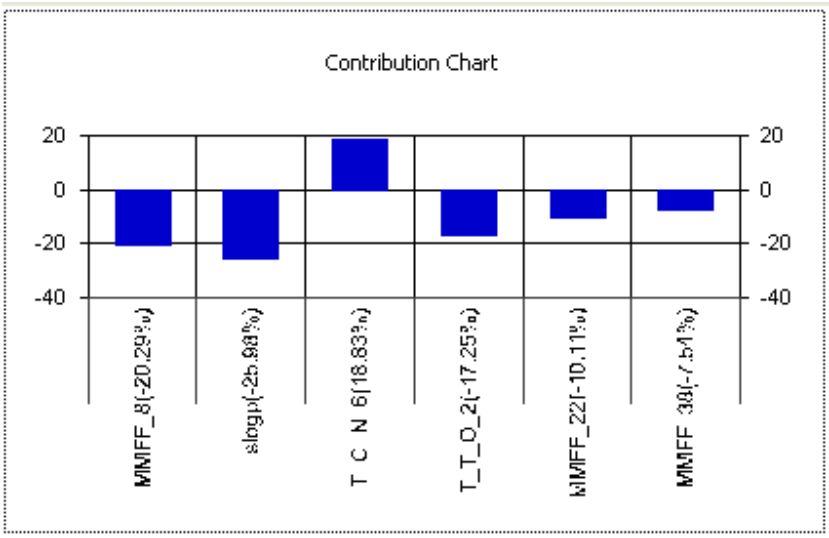


Figure-5: Contribution chart of various descriptors in biological activity for Model-01.

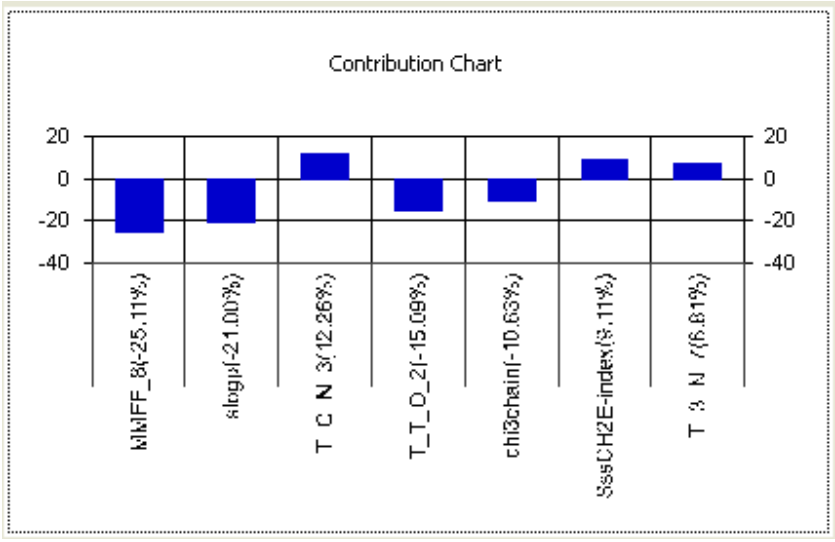


Figure-6: Contribution chart of various descriptors in biological activity for Model-02.

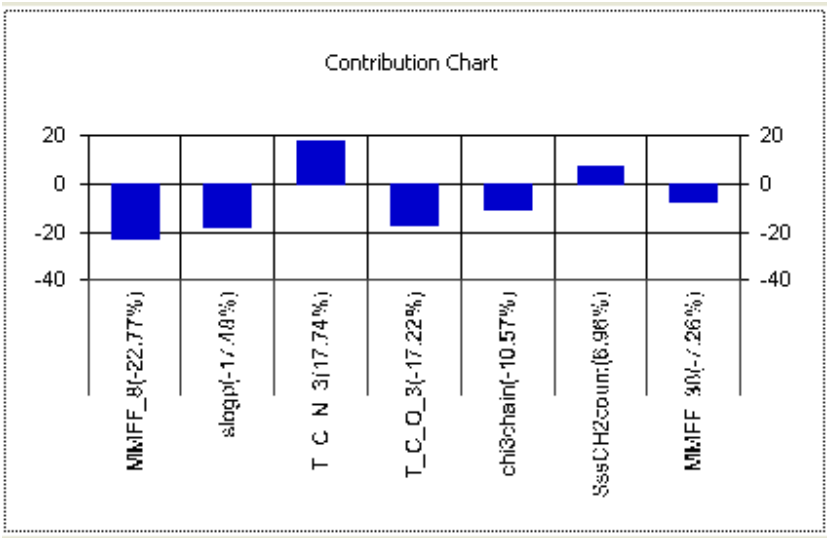


Figure-7: Contribution chart of various descriptors in biological activity for Model-03.

Interpretation of the result:

Model-01: From the Table-4, it can be seen that in the present study MLR (coupled with stepwise forward-backward variable selection) led to a statistical significant model. The developed MLR model reveals that the descriptor, T_C_N_6 (the count of number of carbon atoms separated from nitogen atom by 6 bonds in the molecule) plays an important role (18.83%) in determining DPP-4 inhibitory activity. The other descriptors like MMFF_8 (indicating 8 times that atom type has occurred in the given molecule, -20.29%) and slogp (-25.98%) are inversely proportional to activity.

Model-02: From the Table-4, the present study PLS (coupled with stepwise forward- backward variable selection) led to a statistical significant model. The developed PLS model reveals that the descriptors T_C_N_3 (the count of number of carbon atoms separated from nitogen atom by 3 bonds in the molecule) plays an important role (12.26%) in determining DPP-4 inhibitory activity. The other descriptors like MMFF_8 (-25.11%) and slogp (-21.00%) are inversely proportional to activity.

Model-03: From the Table-4, the present study PLS (coupled with stepwise forward- backward variable selection) led to a statistical significant model. The developed PLS model reveals that the descriptors T_C_N_3 (17.74%) & SssCH2Count (any carbon atom attached with two non-hydrogen atoms by two single bonds, 6.96%) plays an important role in determining DPP-4 inhibitory activity. The other descriptors like MMFF_8 (-22.77%), T_C_O_3 (the count of number of carbon atoms separated from oxygen atom by 3 bonds in the molecule, -17.22%) and slogp (-17.48%) are inversely proportional to activity.

From Figure 2, 3 and 4, it is seen that the plots of observed vs predicated activity for different models provide an idea about how well the models were trained and how well they predict the activity of the external test set.

CONCLUSION

In present study an attempt has been made to identify the necessary structural and substituent requirements. From the present QSAR analysis, three best models were generated among which any one can be used for predicting the activity of the newly designed compounds in finding some more potent molecules. Finally, it is concluded that the work presented here will play an important role in understanding the relationship of physiochemical parameters with structure and biological activity. By studying the QSAR model one can select the suitable substituent for active compounds with maximum potency.

Acknowledgement
Authors are thankful to the Head, S.L.T. Institute of Pharmaceutical Sciences, Guru Ghasidas Vishwavidyalaya, Bilaspur(C.G.), India, for providing the necessary facilities for this work. One of the author (SKJ) is thankful for AICTE for the award of research project under RPS scheme.

REFERENCES

[1] S. Nordhoff, A. Feurer, O. Hill, *Bioorg. Med. Chem. Lett.*, **200**6, 16, 1744.
[2] S. A. Ross, E. A. Gulve, M. Wang, *Chem. Rev,*. **2004**, 104, 1255.
[3] J. M. Lenhard, W. K. Gottschalk, *Advanced Drug Delivery Reviews*, **2002,** 54, 1199.
[4] L. K. Soni, A. K. Gupta, S.G. Kaskhedikar, *E-Journal of Chemistry*, **2004**, 1, 170.
[5] J. Zeng, G. Liu, Y. Tang, H. Jiang, *J Mol. Model*, **2007**, 13, 993.
[6] A. Yaron, F. Naider, *Crit Rev Bioche. Mo.l Bio.*, **1993**, 28, 31.
[7] S. K. Jain, N. Raj, H. Rajak, *J. Inst. Chemists (India)*, **2008**, 80, 77.
[8] D.J. Drucker, *Diabetes*, **1998**, 47, 159-169.
[9] C. F. Deacon, T. E. Hughes, J. J. Holst, *Diabetes*, **1998**, 47, 764-769.
[10] R.A. Pederson, H.A. White, R.P. Pauly, C.R.P Mclntosh, H.U. Demuth, *Diabetes*, **1998**, 47, 1253-1258.
[11] H. Fukushima, A. Hiratate, M. Takahashi, A. Mikami, K. Kitano, S. Chonan, *Bioorg. Med. Chem.* **2008,** 16(7), 4093-4106.
[12] VLife Sciences Technology Pvt. Ltd. Pune-411045, Web: www.vlifesciences.com.