QSAR Studies on Ant Malarial Activity of Diverse Set of Compounds for D6 Strain of *P. falciparum*

D. Thakur*[1], Vandana Khatod[2], C.D. Asthana[3] and S. Thakur[4]

ABSTRACT

Malaria spread by sporozoa of p.falciperum and is well known as an infectious lethal disease.Sometimes it causes death if the medicine is not given in time.Only a limited number of drugs can now prevent and cure malaria.For in vitro testing the clones D6 & NF54 of p. felciperum are most often used.These are resistant to present drugs like mefloquine.The data set based on the D6 strains consisting of 57 organic compounds was collected from different sources and their antimalarial activity is predicted by taking different parametric models.

Keywords :*Malaria, sporozoa, antimalarial activity.*

1. INTRODUCTION

Malaria spreads by sporozoa of the genus plasmodium and is well known as an infectious lethal disease. Sometimes it causes death if the medicine is not given in time. The main symptom of malaria is periodic fever, anemia, and enlargement of the liver and spleen. Millions of new clinical cases are being reported with a high percentage of fatalities among children in the tropical and subtropical countries of Asia, Africa, and South America[1].

Only a limited number of drugs can now prevent and cure malaria. One of them is artemisine[2–4]. Also few potentially antimalarial drugs are used as chemotherapeutics which are (i) quinoline derivatives, e.g., primaquine, chloroquine, and mefloquine[5], and/ or (ii) simple sulfonamides, e.g., sulfadoxine[6,7], pyrimidine derivatives; pyrimethamine[8] .

For in vitro testing of the antimalarial activity the clones of *P. falciparum* are most often used. They include: (i) Sierra Leone (D6), Thailand (Thai), and NF54 clones, which are mefloquine resistant and chloroquine sensitive. Likewise (ii) Indochina (W2), and Colombia (FcB1) clones are chloroquine resistant, but mefloquine sensitive.

Pharmacophore search method is very common for synthesis of new compounds. Grigorov et al.[1] correlated the antimalarial activity of a series of synthetic 1,2,4-trioxanes with molecular structure using this technique for which they have used CATALYST package.[9]

Girone´s et al. studied the application, within a quantum similarity framework, of the kinetic energy based quantum similarity measures in the evaluation of the antimalarial activity. The authors used two molecular sets composed of artemisinin derivatives, in which the 50% inhibition of synthesis and reduction of hidrofolate (IC50) in different P. falciparum clone were analysed. Satisfactory correlations were obtained for all antimalarial activities in all the studied

1*. D. Thakur, Department of App. Sc., People's University, Bhopal, India. e-mail :youvisitdilip@gmail.com
2. Vandana Khatod, PGV College, Gwalior, India. e-mail : vandanakhator011@gmail.com
3. C.D. Asthana, PGV College, Gwalior, India. e-mail : asthanac1751@gmail.com
4. S. Thakur, KKITM, Gwalior, India. e-mail : thakurshivakar47@gmail.com

molecular sets. Four-parameter QSAR models that relate IC50 (NF54 clone, r = 0.754) and logIC_{50} (D6 clone, r = 0.767, and W2 clone, r = 0.821) with the principal components (PCs) were proposed by the authors to be used for the prediction of antimalarial activity. The authors observed that the PCs accounting for the maximal variance are not always those best related to the activity.

The data set based on the D6 strain of P. falciparum consisting of 57 organic compounds has been taken from four literature sources [1-4]. The structures used in the present study are presented in Table-1 along with their log IC50 (nM) values. The structures have been drawn using Chemsketch software developed by ACD labs.

2. METHODOLOGY USED

The mol files have been obtained by depleting all the carbon-hydrogenbonds using the chemsketch software[5], then DRAGON software [6] was used for calculating a variety of descriptor, from which useful descriptors are selected by variable selection. Then multiple regression analysis employing statistical significant models were obtained using NCSS software [7] having maximum R2 value [8]. The data set is divided into training set (75%) and test set (25%) using random selection technique. The following parameters are found to be useful for performing multiple regression analysis. All the values of these descriptors along with log IC50 arereported in Table-2.

The correlation matrix calculated for different parameters and also for activity is reported in Table-3. A perusal of this Table reveals that no auto-correlation exists among any of the parameters and hence they can be used in multi-parametric modeling without any further explanation. The data is divided into training set and test set by random selection method and then subjected to regression analysis using NCSS software and the results obtained for variable selection for multiple regression analysis are summarized in Table-4. The regression parameters and the quality of regression for statistically allowed models are presented in Table-5.

3. RESULTS AND DISCUSSIONS

The statistical details suggest following models:

3.1 One-Parametric Model

The best One-Parametric model is obtained with NRS (number of ring systems) as correlating parameter. The model is as below:

$$\log IC_{50} = -0.8011(\pm 0.1018)\ NRS + 4.8817$$

N=43, Se= 0.6710, R^2= 0.6016,
F-ratio = 61.911, Q= 1.156

Here after n is total number of compounds; is the standard error of estimation; R^2 is the square of correlation coefficient; R^2_{adj} is the adjusted R^2; F is the Fisher's ratio and Q is the Pogliani's quality factor[9] which is the ratio of R/Se. However, this model explains only 75% of variance.

3.2 Two-Parametric Model

When SpMin1_Bh(s) is added to above model a bi-parametric model is obtained with improved statistics (Table-5). Here R^2 changes from 0.6710 to 0.7441. The drastic improvement in R^2 suggests that the bi-parametric model is better than the single-parametric model. The Q-value also confirms the above finding. This model explains 74% of variance.

$$\log IC_{50} = -1.2210(\pm 0.1214)\ NRS + 3.2171(\pm 0.6817)\ SpMin1\_Bh(s) + 1.8484$$

N=43, Se = 0.5449, R^2 = 0.7441, R^2_{adj} = 0.7313,
F-ratio = 58.147, Q = 1.583

The R^2_{adj} changes from 0.5919 to 0.7313 suggesting that added parameter has its fair share in the model.

3.3 Three-Parametric Model

The Q value also changes from 1.583 to 2.003 suggesting that the three parametric model is the best for modeling log IC_{50} activity of present set of compounds. Further addition of MATS4e to the above model, improves the quality of model as the variance changes from 0.7313 to 0.8266. The variance of model is 82% the model is as under:

log IC_{50} = - 0.9923(±0.1143) NRS+ 2.9364 (±0.5720) SpMin1_Bh(s)+ 3.0452 (±0.7066) MATS4e+1.7884

N = 43, Se= 0.4538, R^2 = 0.8266, R^2_{adj} = 0.8133, F-ratio=61.989, Q=2.003

3.4 Four - Parametric Model

When MATS5m is added to three parametric model discussed above a four-parametric model is obtained. All the statistical parameters have improved values. Hence from all the counts one may conclude that this model is better than all the previously reported models. The model is given below:

logIC_{50}= 3.5280(±0.7118) MATS4e-2.0026 (±0.9288)MATS5m-0.9794(±0.1094) NRS+2.6358(±0.5644) SpMin1_Bh(s) +2.1555

N= 52, Se =0.4340, R^2 = 0.8455, R^2_{Adj} = 0.8293, F-ratio = 52.003, Q =2.119

The variation observed in R^2_{Adj} (Table-5) on addition of each variable indicates that added parameters have favourable contributions to the proposed models. The lowest values of SE and also highest value of F-ratio and Q- value further confirm our findings.

Further, conformation is also obtained by estimating log IC_{50} using the above model. The obtained values are reported in Table-6. The estimated values are in good agreement with the observed value suggesting that this model is the best. We plotted a graph between observed and estimated pIC_{50} values using this model such a comparison is shown in Fig.1. The predictive power of model comes out to be 0.5525.

The proposed models were validated by the leave-one-out cross validation procedure and the parameters obtained thereby are reported in Table-7. PRESS (predicted residual sum of squares) appears to be the most important cross validation parameter accounting for a good estimate of the real predictive error of the models. Its value less than SSY (sum of squares of response value) indicates that the model predict better than the chance and can be considered statistically significant. In our case PRESS << SSY indicating that all the models obtained are statistically significant. The ratio PRESS/SSY, can be used to calculate approximate confidence intervals of prediction of new compounds. To be a reasonable and significant QSAR model the ratio PRESS/SSY should be less than 0.4 (PRESS/SSY <0.4) and the value of this ratio smaller than 0.1, indicates an excellent model. A close observation of Table-7 shows that all the models have the PRESS/SSY ratio more or less or nearer to 0.1, indicating that all the proposed models have best predicting capacity. R^2_{cv} is the cross validated squared correlation coefficient.

The highest R^2_{cv} value (0.817) for four-variable model finally confirms our results. The two important cross-validation parameters uncertainty in prediction (S_{PRESS}) and predictive squared error (PSE) also favour our results. The lowest value of S_{PRESS} (0.434) and PSE (0.408) further conforms above findings.

In order to explain whether or not the proposed models are free from collinearity, we have calculated the VIF (variance inflation factor), Eigen values (ëi), condition number (k), tolerance (T) for all the independent parameters used in the proposed models and the same are recorded in Table-8. The parameters whose VIF value is greater than 10 will show collinearity. A perusal of Table-8 shows that in all the cases VIF values are less than 10, which means that all the proposed models reported by us are free from the defect of collinearity. Similarly if ëi, (Eigen value) is found to be greater than 5 then the model will suffer from collinearity. The Table-8 shows that for all the parameters ë$_i$ is less than 5. Therefore, from this point of view also, proposed models are free from the defect of collinearity. Another test for collinearity is condition number if its value is found to be >100 and then the collinearityexists but the values reported in Table-8 indicates that the values are always <100. Similarly Tolerance value equal to 1 or less indicates absence of collinearity. Table-8 indicates that all the above mentioned parameters or models discussed in the study are free from defect of collinearity. The ridge traces are recorded in Fig.2 and Fig.3 respectively.

4. CONCLUSION

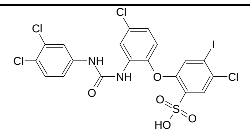
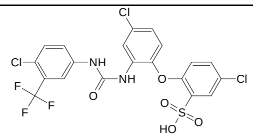
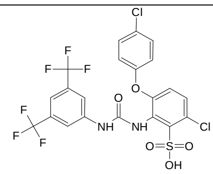
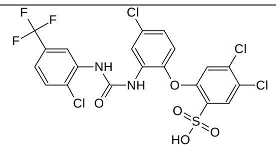
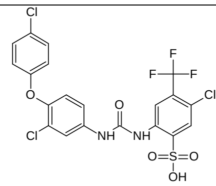
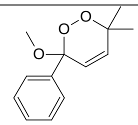
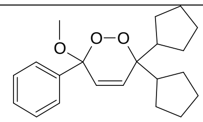
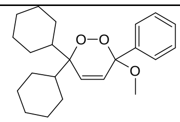
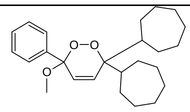
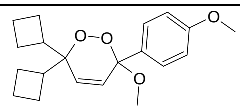
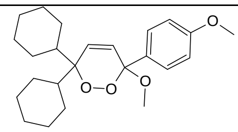
On the basis of the obtained results we have made the following conclusions:

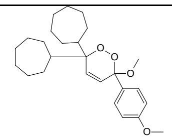
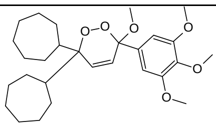
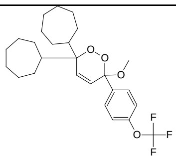
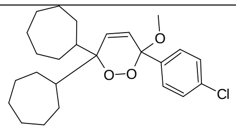
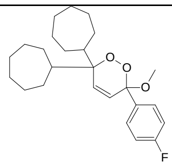
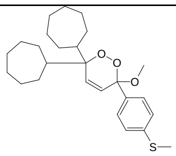
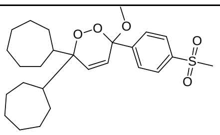
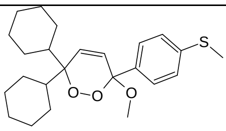
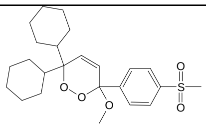
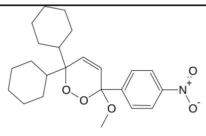
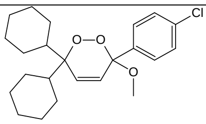
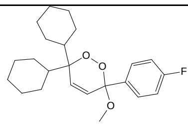
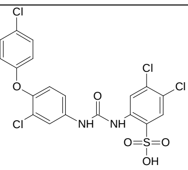
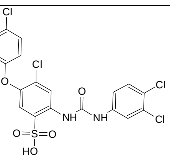
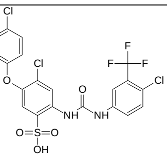
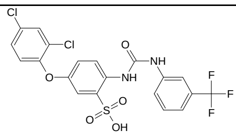
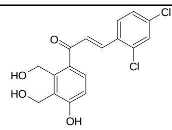
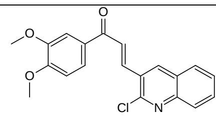
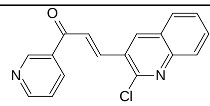
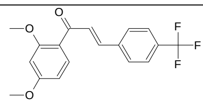
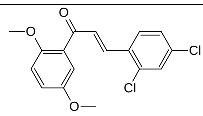
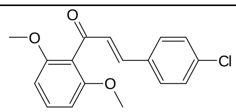
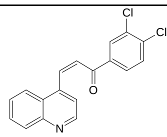
The best model obtained suggests that the NRS, 2D autocorrelation descriptors and Burden eigenvalues play important role in prediction of biological activity of present set of compounds.

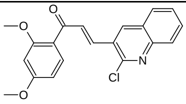
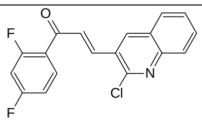
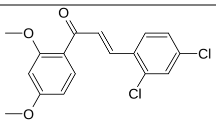
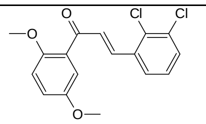
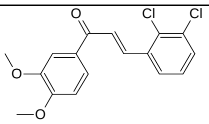
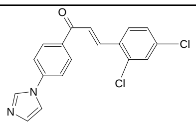
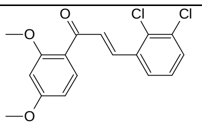
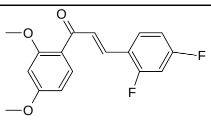
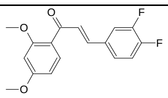
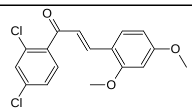
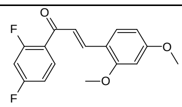
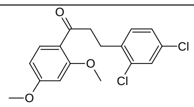
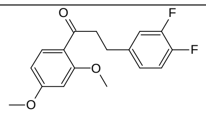
$$3.5280(\pm0.7118)\ MATS4e-2.0026\ (\pm0.9288)\ MATS5m-0.9794(\pm0.1094)\ NRS+\ 2.6358\ (\pm0.5644)\ SpMin1\_Bh\ (s)\ +2.1555$$

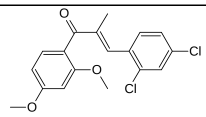
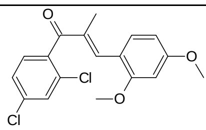
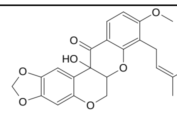
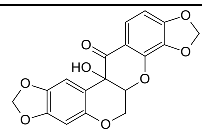
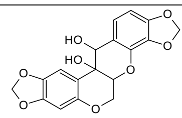
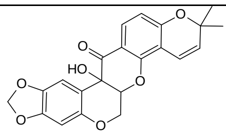
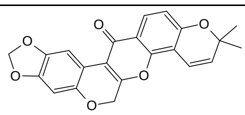
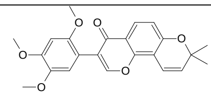
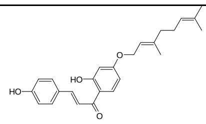
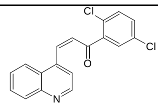
1. Negative coefficient of NRS suggests that number of ring systems plays a dominant role in deciding the antimalarial activity of present set of compounds. Therefore, in designing the new compounds, the structure be modified in such a way so as to get a lower value of NRS (less number of rings).
2. Both Burden eigen value(SpMin1_Bh(s)) and Moran auto correlation (MATS4e) have Positive coefficients, therefore molecules having high value of Burden eigen values and MATS4e should be preferred in designing new compounds to get better activity.
3. Negative coefficient of MATS5m suggests that lower value of MATS5m will enhance the activity.

Table - 1 : Structures and Their Antimalerial Activity Log IC50 Values Used in Present Study

Compd. No	Structure	Log IC_{50}
1*		1.73
2		1.58
3		1.08
4		1.37
5		3.04
6		2.28
7*		2.45
8		2.34
9*		2.2
10		2.26
11		2.32

12		2.08
13		1.79
14*		1.76
15		1.93
16		1.89
17		1.49
18*		2.2
19		1.75
20		1.66
21		2
22		2.3
23		1.68
24		1.7
25*		1.62
26		1.58
27		1.53
28*		2.28
29		2.63
30*		2.86
31		3.4
32		2.83
33*		2.96
34		3.09

35		3.23
36*		3.11
37		2.99
38		2.89
39		3.04
40*		3.29
41		2.95
42		3.11
43		3.29
44		3.32
45		3.53
46		4.28
47*		4.18

48		4.2
49*		3.8
50		4.15
51*		4.85
52		4.78
53		5
54		4.29
55		4.59
56		4.44
57		2.11

*represents test set

Table-2 : Calculated Values of The Descriptors Used in Present Study

Compd. No.	NRS	MATS4e	MATS5m	SpMin1_Bh(s)
1	3	-0.083	0.079	1.147
2	3	-0.075	0.027	1.135
3	3	-0.229	-0.052	1.154
4	3	-0.233	0.126	1.135
5	3	-0.102	0.028	1.174
6	2	-0.258	-0.131	1.17
7	4	-0.153	-0.268	1.43
8	4	-0.124	-0.116	1.577
9	4	-0.106	0.02	1.498
10	4	-0.238	-0.222	1.607
11	4	-0.133	-0.101	1.577
12	4	-0.112	0.022	1.498
13	4	-0.043	0.012	1.498
14	4	-0.058	0.023	1.498
15	4	-0.117	0.033	1.498
16	4	-0.104	0.03	1.498
17	4	-0.115	0.027	1.498
18	4	-0.116	0.018	1.498
19	4	-0.134	-0.074	1.577
20	4	-0.14	-0.083	1.577
21	4	-0.148	-0.122	1.577
22	4	-0.141	-0.07	1.577
23	4	-0.126	-0.102	1.577
24	3	-0.141	0.075	1.146
25	3	-0.136	0.047	1.151
26	3	-0.099	0.092	1.152
27	3	-0.189	-0.025	1.151
28	2	-0.032	-0.031	1.222
29	2	-0.196	-0.083	1.221
30	2	-0.134	-0.052	1.221
31	2	-0.067	-0.014	1.142
32	2	-0.004	0.046	1.121
33	2	0.256	-0.015	1.122
34	2	-0.189	0.059	1.237
35	2	0.059	-0.014	1.221
36	2	0.087	-0.12	1.221
37	2	0.123	0.018	1.121
38	2	-0.09	-0.062	1.121
39	2	-0.216	-0.152	1.121
40	3	-0.067	0.009	1.14
41	2	0.037	-0.09	1.121
42	2	0.117	0.038	1.111
43	2	0.002	0.038	1.112
44	2	0.099	0.011	1.123
45	2	0.15	0.02	1.118
46	2	0.092	-0.026	1.256
47	2	0.006	-0.017	1.25
48	2	0.073	0.012	1.149
49	2	0.051	0.004	1.15
50	1	0.029	0.096	1.304
51	1	-0.064	0.034	1.281
52	1	-0.047	0.041	1.336
53	1	0.044	0.027	1.296
54	1	0.093	0.046	1.246
55	2	0.087	-0.051	1.226
56	2	0.17	0.117	1.261
57	2	-0.031	0.193	1.237

NRS = number of ring systems (Ring descriptors)
MATS4e = Moran autocorrelation of lag 4 weighted by Sanderson electro negativity (2D autocorrelations)
MATS5m = Moran autocorrelation of lag 5 weighted by mass (2D autocorrelations)
SpMin1_Bh(s) = smallest eigenvalue n. 1 of Burden matrix weighted by I-state (Burden eigenvalues)

Table - 3: Correlation Matrix

	Log IC_{50}	NRS	MATS4e	MATS5m	SpMin1_Bh(s)
Log IC_{50}	1				
NRS	-0.763	1			
MATS4e	0.621	-0.525	1		
MATS5m	0.109	-0.291	0.315	1	
SpMin1_Bh(s)	-0.305	0.718	-0.354	-0.341	1

Table - 4: Variable Selection for Multiple Regression Analysis (Training Set)

Model Size	R^2	R^2-Change	Variable Names
1	0.602	0.602	NRS
2	0.744	0.142	NRS, SpMin1_Bh(s)
3	0.827	0.083	NRS, MATS4e, SpMin1_Bh(s)
4	0.846	0.019	NRS, MATS4e, MATS5m, SpMin1_Bh(s)

Table - 5 : Regression Parameters and Quality of Correlation for the Obtained Models

Model No	Parameters used	A_i = (1----5)	B	Se	R^2	R^2_{Adj}	F-ratio	Q=R/Se
1	MATS5m	2.0490(±1.9563)	2.7563	1.0491	0.0261	-	1.097	0.154
2	SpMin1_Bh(s)	-1.0870(±0.8605)	5.0739	1.0101	0.0971	-	4.410	0.308
3	MATS4e	6.1616(±0.9936)	3.1213	0.7636	0.4840	-	38.453	0.911
4	NRS	-0.8011(±0.1018)	4.8817	0.6710	0.6016	0.5919	61.911	1.156
5	MATS5m NRS	-1.5696(±1.3283) -0.8455(±0.1081)	4.9876	0.6678	0.6150	0.5958	31.953	1.174
6	MATS4e NRS	3.4582(±0.8973) -0.5830(±0.1046)	4.5130	0.5801	0.7095	0.6950	48.843	1.452
7	NRS SpMin1_Bh(s)	-1.2210(±0.1214) 3.2171(±0.6817)	1.8484	0.5449	0.7441	0.7313	58.147	1.583
8	MATS4e MATS5m NRS	4.1344(±0.8667) -3.0739(±1.1146) -0.6275(±0.0983)	4.6484	0.5374	0.7569	0.7382	40.475	1.619
9	MATS4e NRS SpMin1_Bh(s)	3.0452(±0.7066) -0.9923(±0.1143) 2.9364(±0.5720)	1.7884	0.4538	0.8266	0.8133	61.989	2.003
10	MATS4e MATS5m NRS SpMin1_Bh(s)	3.5280(±0.7118) -2.0026(±0.9288) -0.9794(±0.1094) 2.6358(±0.5644)	2.1555	0.4340	0.8455	0.8293	52.003	2.119

Compd. No.	Obs. log IC_{50}	Est. log IC_{50}	Residual
1	1.73	1.79	-0.06
2	1.58	1.89	-0.31
3	1.08	1.555	-0.475
4	1.37	1.135	0.235
5	3.04	1.896	1.144
6	2.28	2.633	-0.353
7	2.45	2.004	0.446
8	2.34	2.189	0.151
9	2.2	1.772	0.428
10	2.26	2.078	0.182
11	2.32	2.128	0.192
12	2.08	1.747	0.333
13	1.79	2.011	-0.221
14	1.76	1.936	-0.176
15	1.93	1.707	0.223
16	1.89	1.759	0.131
17	1.49	1.726	-0.236
18	2.2	1.741	0.459
19	1.75	2.07	-0.32
20	1.66	2.067	-0.407
21	2	2.117	-0.117
22	2.3	2.037	0.263
23	1.68	2.154	-0.474
24	1.7	1.59	0.11
25	1.62	1.677	-0.057
26	1.58	1.72	-0.14
27	1.53	1.634	-0.104

28	2.28	3.367	-1.087
29	2.63	2.89	-0.26
30	2.86	3.046	-0.186
31	3.4	2.998	0.402
32	2.83	3.045	-0.215
33	2.96	4.087	-1.127
34	3.09	2.672	0.418
35	3.23	3.651	-0.421
36	3.11	3.962	-0.852
37	2.99	3.549	-0.559
38	2.89	2.958	-0.068
39	3.04	2.694	0.346
40	3.29	1.968	1.322
41	2.95	3.462	-0.512
42	3.11	3.462	-0.352
43	3.29	3.059	0.231
44	3.32	3.484	-0.164
45	3.53	3.633	-0.103
46	4.28	3.884	0.396
47	4.18	3.547	0.633
48	4.2	3.459	0.741
49	3.8	3.4	0.4
50	4.15	4.523	-0.373
51	4.85	4.259	0.591
52	4.78	4.45	0.33
53	5	4.693	0.307
54	4.29	4.696	-0.406
55	4.59	3.837	0.753
56	4.44	3.886	0.554
57	2.11	2.961	-0.851

Table-7 : Cross Validated Parameters for the Obtained Models

Model No	Parameters used	PRESS/ SSY	R^2_{CV}	S_{PRESS}	PSE
7	NRS SpMin1_Bh(s)	0.344	0.656	0.544	0.525
9	MATS4e NRS SpMin1_Bh(s)	0.210	0.790	0.454	0.432
10	MATS4e MATS5m NRS SpMin1_Bh(s)	0.183	0.817	0.434	0.408

Table-8 : Ridge Regression Parameters for the Obtained Models

Model No	Parameters used	VIF	T	λ_i	k
7	NRS	2.1603	0.4629	1.7328	1.0000
	SpMin1_Bh(s)	2.1603	0.4629	0.2671	6.4900
9	MATS4e	1.4317	0.6985	0.6883	3.0300
	NRS	2.7538	0.3631	2.0865	1.0000
	SpMin1_Bh(s)	2.1886	0.4569	0.2251	9.2700
10	NRS	2.7621	0.3620	2.3936	1.0000
	MATS4e	1.5890	0.6293	0.7678	3,1200
	MATS5m	1.3174	0.7591	0.6247	3.8300
	SpMin1_Bh(s)	2.3309	0.4290	0.2137	11.2000

VIF= variance inflation factor
T = tolerance
ë$_i$ =Eigen values
k = condition number

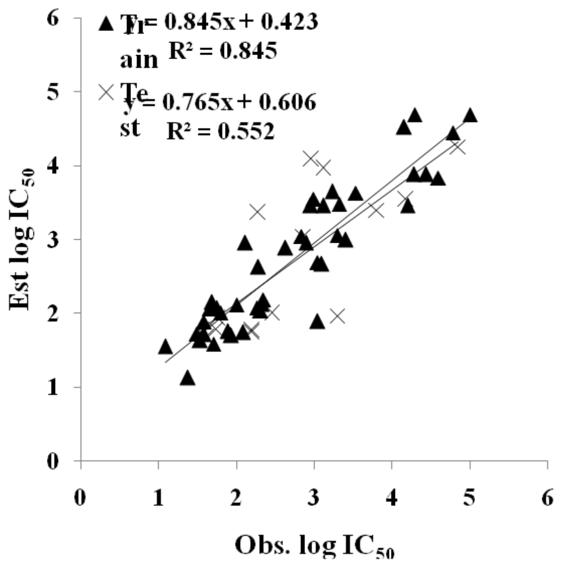


Fig.1 : Correlation Between Observed and Estimated Log IC50 Using Model 10

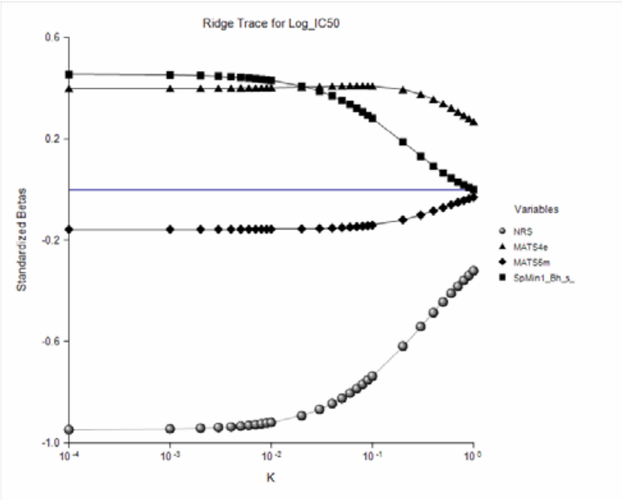


Fig. 2 : Ridge Trace for Five - Parametric Model

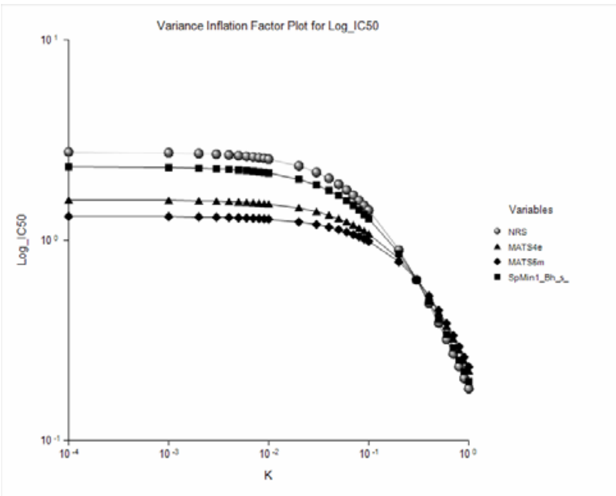


Fig.3.VIF Plot for Four- Parametric Model

REFERENCES

[1] Girones, X.; Gallegos, A.; Carbo-Dorca, R;. J. Comput.- Aided Mol. Des. 2001, 15, 1053.

[2] Jiang, S.; Prigge, S.; Wei, L.; Gao, Y.-E.; Hudson, T.; Gerena, L.; Dame, J.; Kyle, D.;Antimicrob. Agents Chemother. 2001, 45, 2577.

[3] Li, R.; Kenyon, G.; Cohen, F.; Chen, X.; Gong, B.; Dominigues, J.; Davidson, E.; Kurzban, G.; Miller, R.; Nuzum, E.; Rosenthal, P.; McKerrow, J.; J. Med. Chem. 1995, 38, 5031.

[4] Yenesew, A.; Derese, S.; Midiwo, J.; Oketch, H.; Lisgarten, J.; Palmer, R.; Heydenreich, M.; Peter, M.; Akala, H.; Wangui, J.; Liyala, P.; Waters, N.; Phytochemistry 2003, 64, 773.

[5] Chemsketchver 12.0, Software for calculation of physiological parameters. *www.acdlabs.com* (assessedfeb 2010)

[6] Dragon Software for calculation of Topological Indoses,www.disat . unimib,it.

[7] NCSS, statistical analysis software, www. ncss. com (Accessed May 2005)

[8] S. Cheterjee, A.S. Hadi, B. Price; Regression analysis by Examples,3rd.ed.; Wiley & Sons: New York; 6, 141-153

[9] L. Pogliani.; Structure Property Relationship Of Amino Acids and Some Dipeptide, Amino acids, 1994, 6,141-153.

