Journal of Current Chemical & Pharmaceutical Sciences

J. Curr. Chem. Pharm. Sc.: 2(4), 2012, 244-260
ISSN 2277-2871

CHEMOMETRIC DESCRIPTORS IN THE RATIONALE OF ANTIMALARIAL ACTIVITY OF NATURAL AND SYNTHETIC PRODIGININES

P. SINGH* **and NIDHI SHEKHAWAT**

Department of Chemistry, S. K. Government College, SIKAR – 332001 (Raj.) INDIA

(*Received : 28.07.2012; Revised : 05.08.2012; Accepted : 08.08.2012*)

ABSTRACT

The antimalarial activity of natural and synthetic prodiginines has been quantitatively analyzed in terms of chemometric descriptors. The statistically validated quantitative structure-activity relationship (QSAR) models provided rationale to explain the activity against *Plasmodium falciparum* D_6 strain of these compounds. The descriptors identified through combinatorial protocol in multiple linear regression (CP-MLR) analysis have highlighted the role of various chemometric 2D-descriptors. The main contributing descriptors were the information content index of 5-order neighbourhood symmetry (IC5), the mean topological charge indices of order 5 (JGI5), the Moran autocorrelation – lag 6/ weighted by atomic masses (MATS6m) and the Geary autocorrelation–lag 5/weighted by atomic Sanderson electronegativities (GATS5e). The higher values of the descriptors IC5 and JGI5 and lower values of the descriptors MATS6m and GATS5e are required to further improve the antimalarial activity of a compound. From the evolved strategy, a few potential congeners have been suggested for further investigation. The partial least squares (PLS) analysis has further corroborated the results obtained from CP-MLR study.

Key words: Antimalarial activity, Prodiginines, Chemometric descriptors, Combinatorial protocol in multiple linear regression (CP-MLR) analysis, QSAR.

INTRODUCTION

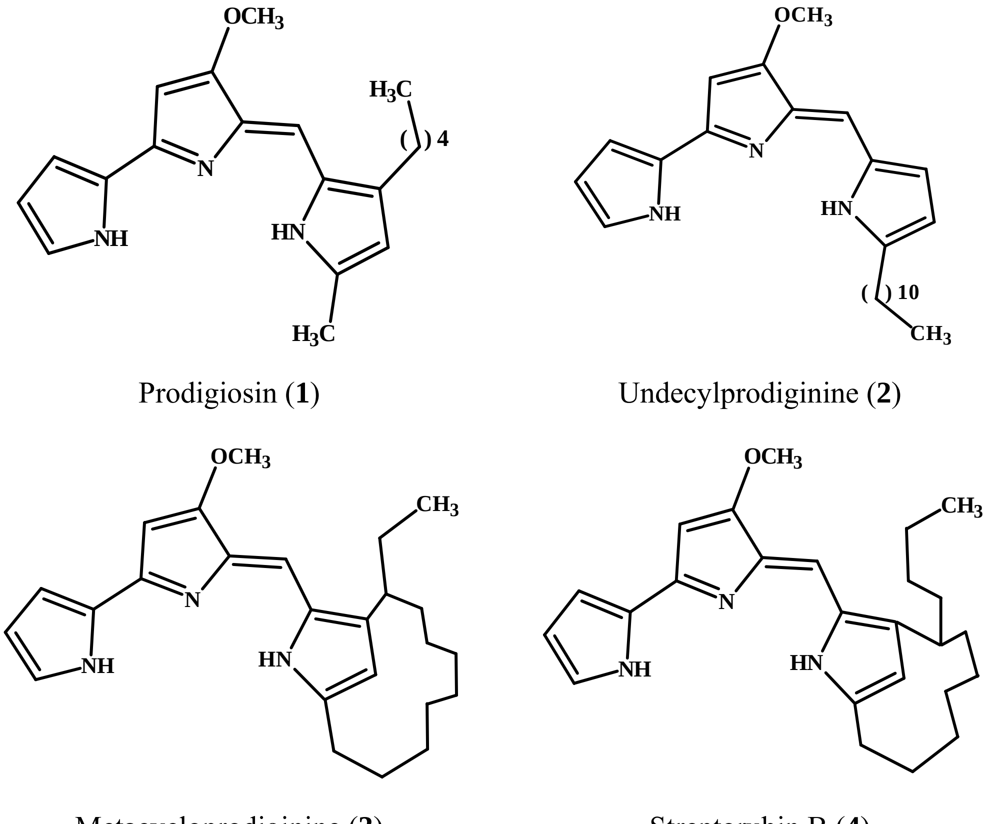
The development of resistance of *Plasmodium falciparum* to conventional antimalarial drugs caused a serious global problem to combat malaria. Despite increased attention to malaria eradication, the disease causes more than a million deaths each year[1]. *P. falciparum*, the protozoan agent responsible for cerebral malaria, is the most worrying parasite, in particular with chloroquine and multi-resistant strains. Besides the worldwide development of chloroquine-resistant *P. falciparum*, resistance has also developed to a variety of quinoline analogues, antifolates, inhibitors of electron transport and perhaps now to artemisinin[2,3]. It is obvious that the haunt for effective novel antimalarial compounds must be comprehensive and must focus on explorations of chemotypes distinct from the prototypes in clinical use. Under prevailing circumstances, the research may be targeted to investigate new pharmacophores and to develop low cost antimalarial compounds, which could significantly contribute to improve the hygienic condition of many developing countries. Since ancient times, natural products have provided great contribution in antimalarial drug discovery, the most notable examples being cinchona alkaloids and artemisinin[4]. Likewise the naturally occurring prodiginines, representing another class, are a family of linear and cyclic oligopyrrole red-

Available online at www.sadgurupublications.com

*Author for correspondence; E-mail: psingh_sikar@rediffmail.com; Ph.: +91 1572 251039; Fax: +91 1572 270039

pigmented compounds, which are produced by actinomycetes and other eubacteria. These compounds possess the antibacterial[5], anticancer[6] and immunosuppressive activity[7]. A few of them induce apoptotic effects, breaking genomic eoxyribonucleic acid (DNA) strands[8]. These compounds are also shown to have potent *in vitro* activity against *Plasmodium* species, at much lower concentration than observed with mammalian cells[9-13]. However, the reported efficacy of naturally occurring prodiginines *in vivo* is associated with toxicity thus hindering their consideration as antimalarial agents.

The earlier studies were limited to naturally occurring prodiginines, leaving open the possibility that some synthetic analogues may have improved *in vitro* activity or *in vivo* efficacy or reduced toxicity. In view of this, Papireddy et al.[14] have recently undertook a more comprehensive assessment of the antiplasmodial activity of prodiginines, initially reassessing the activity of four naturally occurring prodiginines **1-4** (Fig. 1) and subsequently assessing a series of synthetic analogues.



Prodigiosin (**1**)

Undecylprodiginine (**2**)

Metacycloprodiginine (**3**)

Streptorubin B (**4**)

Fig. 1: Natural alkyl prodiginines (Compounds 1-4 listed in Table 1)

However, their structure-activity relationship (SAR) study on synthetic analogues was mainly based on the alteration of substituents at different positions and provided no rationale to reduce the trial-and-error factors. Hence, the present study is aimed at to establish the quantitative SAR (QSAR) between experimental antimalarial activity and chemometric 2D-descriptors which may focus on the molecular structures of the compounds. Such a 2D-QSAR may provide the rationale for drug-design and help to explore the possible mechanism of action at the molecular level. In a congeneric series, where a relative study is being carried out, the 2D-descriptors may play important role in deriving the significant correlations with biological activities of the compounds. The novelty and importance of a 2D-QSAR study is due to its simplicity for the calculations of different descriptors and their interpretation (in physical sense) to explain the inhibition actions of compounds at molecular level.

EXPERIMENTAL

Materials and methods

The natural and synthetic prodiginines along with their antimalarial activity against *P. falciparum* D_6 and Dd_2 strains under present investigation (Table 1) have been taken from the literature[14]. The naturally occurring prodiginines are included in Fig. 1, while generalized structure of synthesized compounds is shown in Fig. 2. The antimalarial activity has been expressed on the negative logarithm as pIC_{50} ($-\log IC_{50}$) on the molar basis and stand as the dependent descriptor for present quantitative analysis. For modeling purpose, the data-set was divided into training- and test-sets to insure external validation of models derived through identified descriptors. Additionally, leave-one-out (LOO) and leave-five-out (L5O) procedures were employed for internal validation of such models derived from the training set.

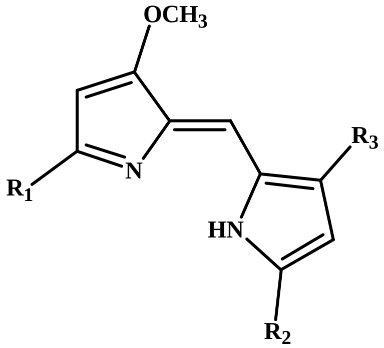


Fig. 2: General structures for Compounds (5-60) listed in Table 1

Table 1: Observed and modeled antimalarial activity of Prodiginines (Fig. 1 and 2 for structures)

S. No.	R_1	R_2	R_3	pIC_{50} (M)				
				D_6				Dd_2
				Obsd.[a]	Calcd.			Obsd.[a]
					Eq. (2)	Eq. (5)	PLS	
1				8.10	7.74	7.63	7.37	ND[b]
2				8.11	5.85	5.81	5.85	ND[b]
3[c]				8.77	8.46	8.41	7.98	ND[b]
4				8.11	8.26	8.18	8.11	ND[b]
5	1*H*-indol-2-yl	CH_3	CH_3	5.37	5.73	5.71	5.71	5.31
6	1*H*-indol-2-yl	n-$C_{11}H_{23}$	H	5.39	6.18	6.14	5.39	5.21
7	Phenyl	n-$C_{11}H_{23}$	H	4.98	6.44	6.10	5.91	4.80
8[c]	Phenyl	CH_3	CH_3	4.71	5.90	5.52	5.26	4.87
9	Furan-2-yl	n-$C_{11}H_{23}$	H	5.54	5.49	5.73	5.67	5.42
10	Thiofuran-2-yl	CH_3	CH_3	< 4.60	4.80	4.52	4.62	< 4.60
11	Thiofuran-2-yl	n-$C_{11}H_{23}$	H	5.23	5.41	5.20	5.34	5.11
12	Furan-2-yl	CH_3	CH_3	< 4.60	4.87	5.10	4.99	< 4.60

Cont…

S. No.	R_1	R_2	R_3	pIC_{50} (M)				
				D_6				Dd_2
				Obsd.[a]	Calcd.			Obsd.[a]
					Eq. (2)	Eq. (5)	PLS	
13	Pyrrol-2-yl	n-C_3H_7	H	5.64	6.24	5.90	5.85	ND[b]
14[c]	Pyrrol-2-yl	n-C_4H_9	H	5.75	7.24	7.04	6.94	5.80
15[c]	Pyrrol-2-yl	n-C_6H_{13}	H	6.43	6.94	6.79	6.71	6.35
16	Pyrrol-2-yl	n-C_8H_{17}	H	7.10	6.96	6.85	6.74	6.89
17	Pyrrol-2-yl	n-$C_{16}H_{33}$	H	6.52[d]	---	---	---	6.40
18	Pyrrol-2-yl	n-$C_{11}H_{22}NH_2$	H	5.77	6.08	6.04	5.98	ND[b]
19[c]	Pyrrol-2-yl	H	$(CH_2)_3COOCH_3$	5.35	6.51	6.42	6.52	ND[b]
20	Pyrrol-2-yl	H	$CH_2CH(CH_3)_2$	6.34	6.95	6.85	6.63	6.64
21	Pyrrol-2-yl	H	n-C_4H_9	7.10	7.54	7.48	7.50	7.75
22[c]	Pyrrol-2-yl	H	n-C_6H_{13}	7.55	7.24	7.21	7.24	8.16
23[c]	Pyrrol-2-yl	H	n-C_8H_{17}	8.34	7.26	7.25	7.29	8.75
24[c]	Pyrrol-2-yl	H	n-$C_{10}H_{21}$	8.10	6.63	6.65	6.77	8.00
25	Pyrrol-2-yl	H	n-$C_{16}H_{33}$	< 4.60	4.04	4.20	4.43	< 4.60
26	Pyrrol-2-yl	H	$C_6H_5CH_2$	7.08	6.89	6.89	7.30	7.07
27	Pyrrol-2-yl	H	4-$OCH_3C_6H_4CH_2$	6.77	7.16	7.51	7.47	6.81
28	Pyrrol-2-yl	H	4-$ClC_6H_4CH_2$	7.19	6.73	6.95	7.22	7.09
29	Pyrrol-2-yl	H	4-$BrC_6H_4CH_2$	7.05	6.81	6.90	7.22	6.97
30	Pyrrol-2-yl	H	2-NaphthylCH_2	7.25	7.64	7.60	8.02	ND[b]
31	Pyrrol-2-yl	CH_3	CH_3	5.05	5.45	5.39	5.36	5.09
32	Pyrrol-2-yl	n-C_6H_{13}	n-C_3H_7	8.35	8.80	8.58	8.30	8.40
33[c]	Pyrrol-2-yl	n-C_8H_{17}	n-C_3H_7	8.54	8.85	8.65	8.40	8.57
34	Pyrrol-2-yl	n-C_3H_7	Cyclohexylethyl	8.77	8.88	8.74	8.51	8.89
35	Pyrrol-2-yl	n-C_6H_{13}	n-C_6H_{13}	8.77	8.18	8.06	8.14	8.96
36	Pyrrol-2-yl	n-C_7H_{15}	n-C_6H_{13}	8.68	8.23	8.11	8.22	8.92
37	Pyrrol-2-yl	n-C_6H_{13}	n-C_8H_{17}	8.31	8.19	8.08	8.19	8.70
38	Pyrrol-2-yl	n-C_7H_{15}	n-C_8H_{17}	8.21	8.12	8.01	8.18	8.54
39	Pyrrol-2-yl	n-C_8H_{17}	n-C_8H_{17}	7.04	7.85	7.75	8.00	6.89
40	Pyrrol-2-yl	Cyclohexylethyl	Cyclohexylethyl	8.28	7.85	7.94	8.29	8.46
41	Pyrrol-2-yl	C_2H_5	4-$ClC_6H_4CH_2$	8.20	7.96	8.18	8.05	8.21
42	Pyrrol-2-yl	n-C_3H_7	4-$ClC_6H_4CH_2$	8.52	8.07	8.10	7.95	8.59
43[c]	Pyrrol-2-yl	n-C_6H_{13}	4-$ClC_6H_4CH_2$	8.70	8.63	8.74	8.75	8.75

Cont…

S. No.	**R_1**	**R_2**	**R_3**	**pIC_{50} (M)**				
				D_6				**Dd_2**
				Obsd.[a]	**Calcd.**			**Obsd.**[a]
					Eq. (2)	**Eq. (5)**	**PLS**	
44	Pyrrol-2-yl	n-C_7H_{15}	4-$ClC_6H_4CH_2$	8.55	8.79	8.90	8.86	8.66
45[c]	Pyrrol-2-yl	n-C_8H_{17}	4-$ClC_6H_4CH_2$	7.80	8.79	8.91	8.90	7.92
45	Pyrrol-2-yl	4-$ClC_6H_4CH_2$	Cyclohexylethyl	8.41	8.56	8.64	8.57	8.54
47[c]	Pyrrol-2-yl	n-C_6H_{13}	4-$FC_6H_4CH_2$	9.05	8.70	9.07	8.97	9.05
48	Pyrrol-2-yl	n-C_8H_{17}	4-$FC_6H_4CH_2$	8.89	8.86	9.23	9.12	8.92
49	Pyrrol-2-yl	n-C_6H_{13}	4-$BrC_6H_4CH_2$	8.54	8.70	8.69	8.75	8.55
50[c]	Pyrrol-2-yl	n-C_8H_{17}	4-$BrC_6H_4CH_2$	8.40	8.86	8.87	8.90	8.54
51	Pyrrol-2-yl	4-$ClC_6H_4CH_2$	4-$ClC_6H_4CH_2$	8.22	7.37	7.54	7.50	8.32
52	Pyrrol-2-yl	4-$FC_6H_4CH_2$	4-$FC_6H_4CH_2$	8.25	7.60	8.09	7.91	8.24
53	Pyrrol-2-yl	4-$BrC_6H_4CH_2$	4-$BrC_6H_4CH_2$	7.85	7.44	7.41	7.46	7.96
54	Pyrrol-2-yl	4-$FC_6H_4CH_2$	4-$ClC_6H_4CH_2$	8.22	8.07	8.43	8.25	8.22
55	Pyrrol-2-yl	4-$BrC_6H_4CH_2$	4-$ClC_6H_4CH_2$	8.08	8.00	8.06	8.00	8.11
56[c]	Pyrrol-2-yl	4-$BrC_6H_4CH_2$	4-$FC_6H_4CH_2$	8.24	8.15	8.46	8.29	8.29
57[c]	Pyrrol-2-yl	2,4-$Cl_2C_6H_4CH_2$	2,4-$Cl_2C_6H_4CH_2$	7.90	8.39	8.51	8.45	7.96
58	Pyrrol-2-yl	2,6-$F_2C_6H_4CH_2$	2,6-$F_2C_6H_4CH_2$	7.83	7.60	7.68	7.95	7.74
59	Pyrrol-2-yl	3-$FC_6H_4CH_2$	3-$FC_6H_4CH_2$	8.29	8.69	8.21	8.12	8.17
60	Pyrrol-2-yl	2-$ClC_6H_4CH_2$	2-$ClC_6H_4CH_2$	8.44	8.70	8.61	8.75	8.31

[a] IC_{50} represents the concentration of a compound required to bring out 50% inhibition of D_6 and Dd_2 strains of *P. falciparum*.

[b] ND; Activity is not determined. [c] Compound in the test-set, [d] 'Outlier' compound

The selection of compounds for test-set has been made through SYSTAT[15] using the single linkage hierarchical cluster procedure involving the Euclidean distances of the activity, pIC_{50} values. Nearly 25% of the compounds, from total population, were selected from the generated cluster tree in such a way to keep them at a maximum possible distance from each other. In SYSTAT, by default, the normalized Euclidean distances are computed to join the objects of cluster. The normalized distances are root mean-squared distances. The single linkage uses distance between two closest members in clustering. It generates long clusters and provides scope to choose objects at different intervals. Due to this reason, a single linkage clustering procedure was applied.

Molecular descriptors

The structures of the compounds under study have been drawn in 2D ChemDraw[16] using the standard procedure. These structures were converted into 3D objects using the default conversion procedure implemented in the CS Chem3D Ultra. The generated 3D-structures of the compounds were subjected to energy minimization in the MOPAC component, using the AM1 procedure for closed shell systems, implemented in the CS Chem3D Ultra. This will ensure a well defined conformer relationship across the compounds of the study. All these energy minimized structures of individual compounds have been ported to DRAGON software[17] for computing the descriptors corresponding to 0D-, 1D- and 2D-classes. Table 2 provides the definition and scope of these descriptor-classes in addressing the structural features, which were employed in present QSAR work. The combinatorial protocol in multiple linear regression (CP-MLR) computational procedure[18] has been used for present work in developing QSAR models. Prior to application of the CP-MLR procedure, all those descriptors which are inter-correlated beyond 0.90 and showing a correlation of less than 0.1 with the biological endpoints (descriptor versus activity, $r < 0.1$) were excluded. The remaining descriptors, able to address the biological activity of these compounds, served as the database (pool). The descriptors of this database have been scaled[19] so that the values of each descriptor would lies between 0 and 1. The scaled descriptors would then have equal potential to influence the QSAR models and none of them dominate simply because of its higher or lower pre-scaled values compare to the other descriptors.

Table 2: Descriptor classes used for the analysis of antimalarial activity of Prodiginines

Descriptor class (acronyms)	Definition and scope
Constitutional (CONST)	Dimensionless or 0D descriptors; independent from molecular connectivity and conformations
Topological (TOPO)	2D-descriptor from molecular graphs and independent conformations.
Molecular walk counts (MWC)	2D-descriptors representing self-returning walks counts of different lengths
Modified Burden eigenvalues (BCUT)	2D-descriptors representing positive and negative eigen values of the adjacency matrix, weights the diagonal elements and atoms
Galvez topological charge indices (GLVZ)	2D-descriptors representing the first 10 eigen values of corrected adjacency matrix
2D-autocorrelations (2DAUTO)	Molecular descriptors calculated from the molecular graphs by summing the products of atom weights of the terminal atoms of all the paths of the considered path length (the lag)
Functional groups (FUNC)	Molecular descriptors based on the counting of the chemical functional groups
Atom-centred fragments (ACF)	Molecular descriptors based on the counting of 120 atom-centred fragments, as defined by Ghose-Crippen
Empirical (EMP)	1D-descriptors represent the counts of non-single bonds, hydrophilic groups and ratio of the number of aromatic bonds and total bonds in an H-depleted molecule
Properties (PROP)	1D-descriptors representing molecular properties of a molecule

Model development

The CP-MLR is a 'filter'-based variable selection procedure for model development in QSAR studies[18]. Its procedural aspects and implementation are discussed in some of our recent publications[20-25]. The thrust of this procedure is in its embedded 'Filters'. They are briefly as follows: Filter-1 seeds the variables by way of limiting inter-parameter correlations to predefined level (upper limit ≤ 0.79); Filter-2 controls the variables entry to a regression equation through t-values of coefficients (threshold value ≥ 2.0); Filter-3 provides comparability of equations with different number of variables in terms of square root of adjusted multiple correlation coefficient of regression equation, r-bar; Filter-4 estimates the consistency of the equation in terms of cross-validated Q^2 with leave-one-out (LOO) cross-validation as default option (threshold value $0.3 \leq Q^2 \leq 1.0$). All these filters make the variable selection process efficient and lead to a unique solution. In order to collect the descriptors with higher information content and explanatory power, the threshold of filter-3 was successively incremented with increasing number of descriptors (per equation) by considering the r-bar value of the preceding optimum model as the new threshold for next generation.

Y-Randomization

In order to discover any chance correlations associated with the models obtained through CP-MLR, each cross-validated model has been put to a randomization test[26,27] by repeated randomization of the activity to ascertain the chance correlations, if any, associated with them. For this, every model has been subjected to 100 simulation runs with scrambled activity. The scrambled activity models with regression statistics better than or equal to that of the original activity model have been counted, to express the percent chance correlation of the model under scrutiny.

Model validation

Validation of the derived model is necessary to test the prediction and generalization of the method. In the present study, the data set has been divided into training-set for model development and test-set for external prediction. Goodness of fit of the models was assessed by examining the multiple correlation coefficient (r), the standard deviation (s), the F-ratio between the variances of calculated and observed activities (F). The internal validation of derived model was ascertained through the cross-validated index, Q^2, from leave-one-out and leave-five-out procedures. The LOO method creates a number of modified data sets by taking away one compound from the parent data set in such a way that each observation has been removed once only. Then one model is developed for each reduced data set and the response values of the deleted observations are predicted from these models. In leave-five-out procedure a group of five compounds is randomly kept outside the analysis each time in such a way that all compounds, for once, become the part of the predictive groups. A value greater than 0.5 of Q^2-index hints towards a reasonable robust model.

Predictive power of a model

The predictive power of a derived model is based on test-set compounds. The squared correlation coefficient between the observed and predicted values of compounds from test-set, r^2_{Test}, has been calculated to ascertain the same. A value greater than 0.5 of r^2_{Test} suggests that the model obtained from training-set has a reliable predictive power.

Applicability domain

The utility of a QSAR model is based on its accurate prediction ability for new compounds. A model is valid only within its training domain and new compounds must be assessed as belonging to the domain before the model is applied. The applicability domain is assessed by the leverage values for each compound[28,29]. The Williams plot (the plot of standardized residuals versus leverage values, h) can then be used for an immediate and simple graphical detection of both the response outliers (Y outliers) and structurally influential chemicals (X outliers) in the model. In this plot, the applicability domain is established inside a squared area within ± x × (standard deviations) and a leverage threshold h^*. The threshold h^* is generally fixed at $3 (k + 1)/n$ (n is the number of training-set compounds and k is the number of model parameters) whereas x = 2 or 3. Prediction must be considered unreliable for compounds with a high leverage value ($h > h^*$). On the other hand, when the leverage value of a compound is lower than the threshold value, the probability of accordance between predicted and observed values is as high as that for the training-set compounds.

Partial least squares analysis

Partial Least Squares[30-32] (PLS) linear regression is a method suitable for overcoming the problems in MLR related to multicollinear or over-abundant descriptors. This is a modeling technique where information in the descriptor matrix X is projected onto a small number of latent variables (LV) called PLS components, which are linear combination of the original variables. The matrix Y is simultaneously used in estimating the "latent" variables in X that will be most relevant to predict the Y variables. All descriptor variables are preprocessed by autoscaling, using weights based on the variables' standard deviation and the data are mean-centered prior to PLS processing. Scaling of descriptors is necessary because the values have different orders of magnitude.

Cross-validation was employed to select the used optimum number of LVs. With cross-validation, some samples were kept out of the calibration and used for prediction. The process was repeated so that each of the samples was kept out once. The predicted values of left-out samples were then compared to the observed values using predicted residual sum of squares (PRESS). The PRESS obtained in the cross-validation was calculated each time that a new LV was added to the model. The optimum number of LVs was concluded as the first local minimum in the PRESS versus LV plot.

RESULTS AND DISCUSSION

A total number of 487 descriptors, belonging to 0D-2D classes of DRAGON, have been computed for 60 compounds listed in Table 1. Next, the descriptors which were inter-correlated above 0.90 and exhibited correlation less than 0.1 with biological activities have been eliminated in the initial stage. The remaining 75 descriptors able to address antimalarial activity against *P. falciparum* strain D_6 have been scaled and collated in a pool for CP-MLR analyses. A test-set has been selected through SYSTAT and the same was used for external validation of derived models. Fifteen compounds (S. Nos. **3**, **8**, **14**, **15**, **19**, **22**, **23**, **24**, **33**, **43**, **45**, **47**, **50**, **56** and **57**; Table 1) were identified for the test-set while remaining compounds constitute the training-set which was then used for the development of statistical significant models. A number of models in two-, three- and four-descriptors have been derived in succession. In doing so, filter-3 was in turn incremented with increasing number of descriptors (per equation) by considering the r-bar value of the preceding optimum model as the new threshold for next generation.

From preliminary study of quantifying antimalarial activity (*P. falciparum* strain D_6) in terms of molecular descriptors, compound **17** (Table 1) appeared to behave indifferently from other compounds of the series. In fact, it was the lone compound bearing a long aliphatic chain (n-$C_{16}H_{33}$) at R_2 (Table 1, Fig. 2) which appeared to be unsuitable for proper binding at the receptor site. This compound has been treated as the "outlier". The training-set was then employed to explore predictive models. 12 Models in two-descriptors, 2 models in three-descriptors and 3 models in four-descriptors only remained statistically significant and the models, able to account for highest variances in observed activity, have been documented through Equations (1)-(5).

$$pIC_{50} = 7.577 - 2.204\ (0.403)\ MSD + 2.694\ (0.420)\ IC5 - 1.824\ (0.439)\ MATS\ 6\ m$$

$$n = 44,\ r = 0.883,\ s = 0.668,\ F\ (3, 40) = 47.209,\ Q^2_{LOO} = 0.734,\ Q^2_{L5O} = 0.731,$$

$$r^2_{randY}\ (s.d.) = 0.263\ (0.098),\ r^2_{Test} = 0.686 \qquad \ldots(1)$$

$$pIC_{50} = 4.883 + 3.221\ (0.358)\ IC5 + 2.609\ (0.371)\ JGI5 – 2.258\ (0.382)\ MATS\ 6\ m$$
$$n = 44,\ r = 0.910,\ s = 0.592,\ F\ (3, 40) = 64.099,\ Q^2_{LOO} = 0.790,\ Q^2_{L5O} = 0.780,$$
$$r^2_{randY}\ (s.d.) = 0.254\ (0.096),\ r^2_{Test} = 0.616 \qquad …(2)$$

$$pIC_{50} = 7.415 – 1.956\ (0.354)\ MSD + 2.871\ (0.366)\ IC5 – 1.960\ (0.514)\ D/Dr09$$
$$– 1.769\ (0.379)\ MATS\ 6\ m$$
$$n = 44,\ r = 0.916,\ s = 0.578,\ F\ (4, 39) = 51.007,\ Q^2_{LOO} = 0.799,\ Q^2_{L5O} = 0.788,$$
$$r^2_{randY}\ (s.d.) = 0.289\ (0.096),\ r^2_{Test} = 0.694 \qquad …(3)$$

$$pIC_{50} = 5.748 + 3.133\ (0.351)\ IC5 – 2.801\ (0.502)\ D/Dr09 + 2.360\ (0.416)\ JGI6$$
$$– 2.273\ (0.371)\ MATS\ 6\ m$$
$$n = 44,\ r = 0.918,\ s = 0.571,\ F\ (4, 39) = 52.560,\ Q^2_{LOO} = 0.781,\ Q^2_{L5O} = 0.792,$$
$$r^2_{randY}\ (s.d.) = 0.303\ (0.102),\ r^2_{Test} = 0.605 \qquad …(4)$$

$$pIC_{50} = 5.578 + 3.083\ (0.345)\ IC5 + 2.546\ (0.353)\ JGI5 – 2.348\ (0.364)\ MATS\ 6\ m$$
$$– 0.912\ (0.392)\ GATS5e$$
$$n = 44,\ r = 0.921,\ s = 0.561,\ F\ (4, 39) = 54.739,\ Q^2_{LOO} = 0.809,\ Q^2_{L5O} = 0.806,$$
$$r^2_{randY}\ (s.d.) = 0.288\ (0.108),\ r^2_{Test} = 0.668 \qquad …(5)$$

In above models, the values given in the parentheses (in regression equation) are the standard errors of the regression coefficients. The r^2_{randY}(s.d.) is the mean squared correlation coefficient of the regressions in the activity (Y) randomization study with its standard deviation from 100 simulations. In the randomization study (100 simulations per model), none of the identified models has shown any chance correlation.

Models in three-descriptors (Equations 1-2) and in four-descriptors (Equations 3-5) have accounted, respectively, up to 83% and 85% of variances in observed antimalarial activity against *P. falciparum* strain D_6 and the F-values for them remained significant at 99% level [$F_{3,40}$ (0.01) = 4.313 and $F_{4,39}$ (0.01) = 3.843]. The indices Q^2_{LOO} and Q^2_{L5O} (> 0.5) have accounted for internal robustness of these models, while the index r^2_{Test} greater than 0.5 revealed that the specified test-set is fully accountable for external validation of above models. Above models have shared 7 descriptors among them and the class, brief description, average regression coefficient and total incidences, for individual descriptor, are given in Table 3. The descriptors appeared in individual model have been found poorly intercorrelated among themselves.

The binding of a compound to a receptor site depends on its shape, size and on a variety of factors, such as the electronic, steric, hydrophobicity, lipophilicity, polarizability, solubility etc. Therefore, in a QSAR study the strategy for encoding molecular information, either explicitly or implicitly, should account for these physicochemical effects. These effects may be interpreted in terms of molecular descriptors of a compound. A major step in constructing the QSAR models is to find a set of molecular descriptors that represents variation in the structural properties, hence the biological activities, of the molecules.

The descriptor MSD, representing the mean square distance (Balaban), is derived by applying different algebraic operators to the distance matrix which collects topological distances between pair of atoms. The topological distance between two atoms is the length (i.e., number of involved bonds) of the shortest path between the two atoms. The descriptor can be sensitive to one or more structural features of the molecule such as size, shape, symmetry, branching and cyclicity and can also encode chemical information concerning atom type and bond multiplicity. The descriptor IC5, denoting the mean information content (neighbourhood symmetry of 5-order), is obtained for a H-included molecular graph. This index represents a measure of structural complexity per vertex. The descriptor D/Dr09, is the distance/detour ring index of order 9. It is calculated by summing up distance/detour quotient matrix row sums of vertices belonging to single rings in the molecule. This is considered as a special substructure descriptor reflecting local geometrical environments in the complex cyclic system. The descriptors JGI5 and JG6 are the mean topological charge indices of order 5 and order 6, respectively. The descriptor, MATS6m is the Moran autocorrelation – lag 6/weighted by atomic masses. Finally, the descriptor, GATS5e is the Geary autocorrelation – lag 5/weighted by atomic Sanderson electronegativities. The autocorrelation descriptors, representing the topological structure of the compounds, explain how the values of certain functions, at intervals equal to the lag (path), are correlated. The computation of these descriptors involves the summations of different autocorrelation functions corresponding to different structural lags and leads to different autocorrelation vectors corresponding to the lengths of the sub-structural fragments. Due to their greater applicability, physicochemical properties, such as, atomic masses, atomic van der Waals volumes, atomic Sanderson electronegativities and atomic polarizabilities were inserted as weighting components. As a result, these descriptors address the topology of the structure or parts thereof in association with a specific physicochemical property.

Table 3: Identified descriptors[a] along with their physical meaning, average regression coefficient and incidence[b], in modeling the antimalarial activity of Prodiginines

S. No.	Descriptor	Descriptor class	Physical meaning	Average regression coefficient (incidence)
1	MSD	TOPO	Mean square distance index (Balaban)	-1.956 (1)
2	IC5	TOPO	Information content index (neighbourhood symmetry of 5-order	3.029 (3)
3	D/Dr09	TOPO	Distance/ detour ring index of order 9	-2.381 (2)
4	JGI5	GLVZ	Mean topological charge index of order 5	2.546 (1)
5	JGI6	GLVZ	Mean topological charge index of order 6	2.360 (1)
6	MATS6m	2DAUTO	Moran autocorrelation – lag 6/weighted by atomic masses	-2.130 (3)
7	GATS5e	2DAUTO	Geary autocorrelation – lag 5/weighted by atomic Sanderson electronegativities	-0.912 (1)

[a] The descriptors have been identified from the models, emerged from CP-MLR protocol with a training-set of 44 for the antimalarial activity of prodiginines.

[b] The average regression coefficient of the descriptor corresponding to all four-descriptor models and the total number of its incidence. The arithmetic sign of the coefficient represents the actual sign of the regression coefficient in the models.

The signs of the regression coefficients suggest the direction of influence of explanatory variables in a given model. For example in Equations (2) and (5), the regression coefficients associated to descriptors IC5 and JGI5 have positive signs. These descriptors impart positive influence and their higher values would be conducive in improving the antimalarial activity of a compound. On the other hand, the regression coefficients of descriptors, MATS6m and GATS5e have negative signs, thus imparting negative impact on activity. For a given compound, the lower values of these descriptors would help in improving its antimalarial activity. Equation (2) and Equation (5), being highest significant amongst three- and four-descriptor models, have been used to calculate the pIC_{50} values of all the compounds and the same are listed in Table 1 for the sake of comparison with observed ones. A close agreement between them has been observed. Moreover, the graphical display showing the variation of observed versus calculated activities (training- and test-sets) is given in Figure 3 to illustrate the goodness of fit for these two models.

Compound **10**, **12** and **25**, having their observed pIC_{50} values less than 4.60 and remained the part of data set, have been evaluated correctly. Their calculated pIC_{50} values, using Equations (2) and (5), remained nearly in parity with the observed ones.

Further, the PLS analyses have been performed on 7 identified descriptors related to antimalarial activity (*P. falciparum* strain D_6) of the compounds and the results are summarized in Table 4. In the study, the descriptors were autoscaled (zero mean and unit standard deviation) to provide each one of them equal weightage. In the PLS cross-validation, two-components remained optimum for 7 descriptors and they have explained, 86% of variances in the observed antimalarial activity. The PLS equation of optimum two-components and MLR-like PLS coefficients of identified descriptors, for antimalarial activity against *P. falciparum* strain D_6, is given in Table 4. The calculated activity values of training- and test-set compounds remained in close agreement to that of the observed ones and are listed in Table 1. For comparison, the plot between observed and calculated activities (through PLS analyses) for the training- and test-set compounds is given in Fig. 3. Fig. 4 shows a plot of the fraction contribution of normalized regression coefficients of these descriptors to the activity (Table 4). In decreasing level of significance, 7 descriptors, being the part of Equations (2) and (5) have been arrange as IC5, MATS6m, MSD, D/Dr09, JGI5, GATS5e and JGI6 for the inhibition of D_6 strain of *P. falciparum*. Similar conclusions have been observed from MLR-like coefficients of the PLS model for this activity. Further the descriptors, IC5, JGI5 and JGI6 have positive contribution to antimalarial activity while the descriptors MATS6m, MSD, D/Dr09 and GATS5e have negative contribution to it. The descriptors, in a given significant model, having positive contributions will augment the activity and their higher values are desirable to further improve it. On the other hand, the descriptors having negative contributions will diminish the activity and their lower or more negative values may, therefore, improve it.

Table 4: PLS and MLR-like PLS models from the descriptors of seven parameter CP-MLR models for antimalarial activity of Prodiginines

A : PLS equation	
PLS components	PLS coefficient (s.e.)[a]
Component-1	0.744 (0.052)
Component-2	-0.417 (0.070)
Constant	7.257

B : MLR-like PLS equation

S. No.	Descriptor	MLR-like coefficient (f. c.)[b]	Order
1	MSD	-0.232(-0.146)	3
2	IC5	0.516(0.324)	1
3	D/Dr09	-0.184(-0.115)	4
4	JGI5	0.141(0.089)	5

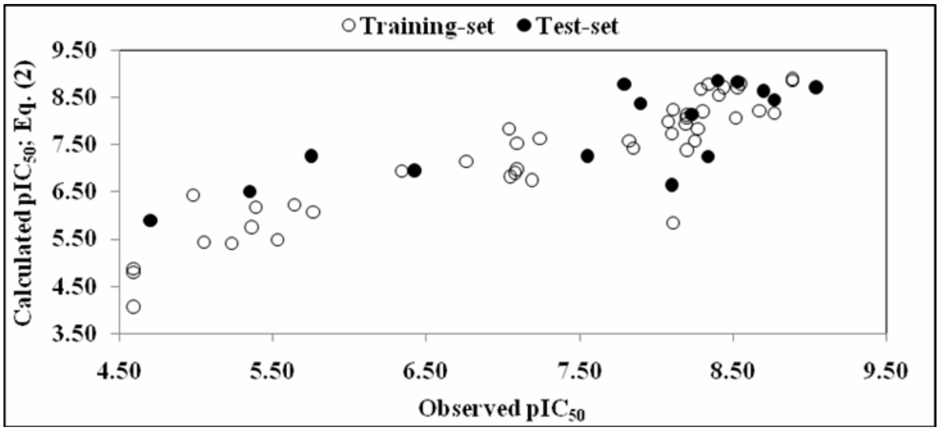
Cont…

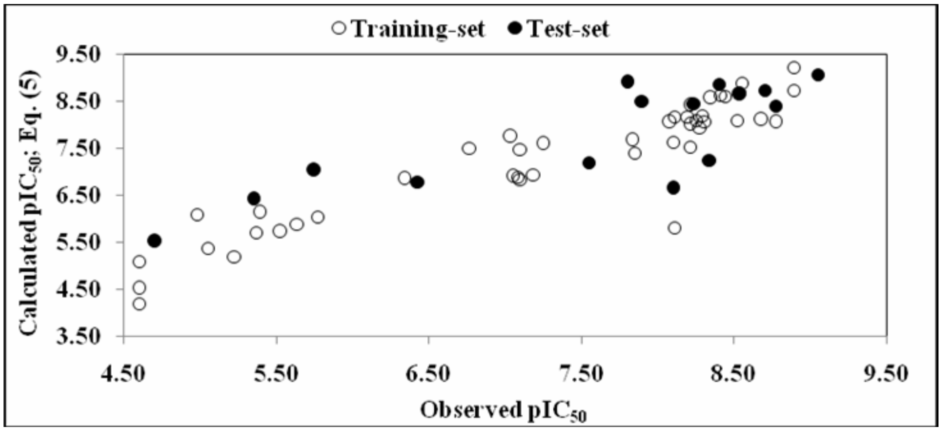
5	JGI6	0.045(0.028)	7
6	MATS6m	-0.387(-0.243)	2
7	GATS5e	-0.087(-0.054)	6
	Constant	7.034	

C: PLS regression statistics	Values
n	44
r	0.925
s	0.534
F	121.787
Q^2_{LOO}	0.832
Q^2_{L5O}	0.824
r^2_{Test}	0.685

[a]Regression coefficient of PLS factor and its standard error.

[b]Coefficients of MLR-like PLS equation in terms of descriptors for their original values; f.c. is fraction contribution of regression coefficient, computed from the normalized regression coefficients obtained from the autoscaled (zero mean and unit s.d.) data





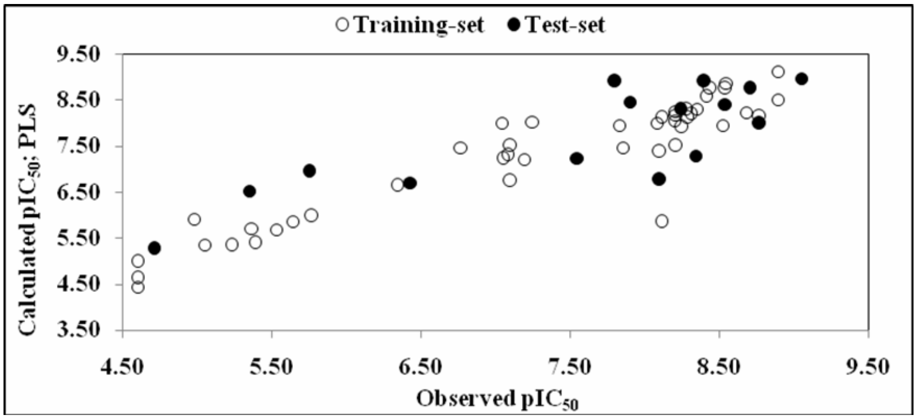


Fig. 3: Plot of observed versus caculated pIC_{50} values relating to antimalarial activity against D_6 strain of *P. Falciparum* for training-set and test-set compounds

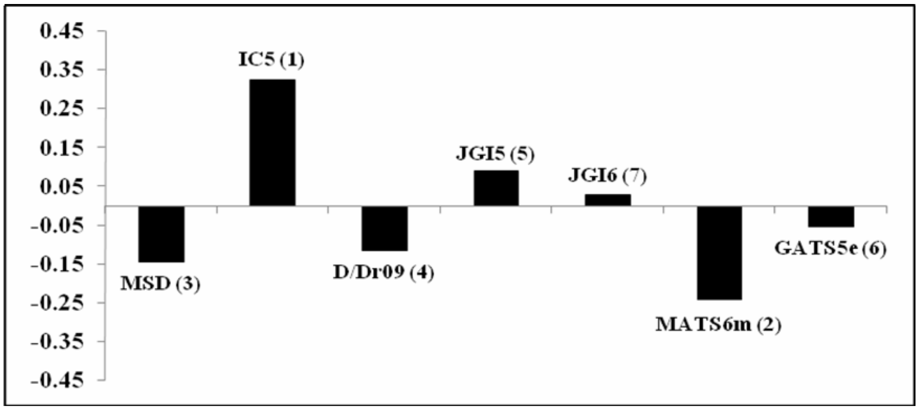


Fig. 4: Plot of fraction contribution of MLR-like PLS coefficients (normalized) against 7 identified descriptors (Table 4) associated with antimalarial activity against D_6 strain of *P. Falciparum*

The exploration of new antimalarial compounds with improved activity profiles prior to their actual synthesis is one of the important aspects of a QSAR study. This will minimize the time and cost associated with identifying new leads. A virtual screening was performed on present antimalarial compounds by insertion, deletion and substitution of different substituents on the original molecules and the effects of the structural modifications on the biological activity were investigated. A few new compounds have been suggested for further biological investigation (Table 5). These predicted compounds have higher pIC_{50} values compared to the highest active compounds in the original data-set (Table 1).

Table 5: Predicted compounds and their modeled antimalarial activity against D_6 strain of *P. Falciparum* (Fig. 2, R_1 = Pyrrol-2-yl)

S. No.	R_2	R_3	Eq. (2)	Eq. (5)	PLS
1	n-C_6H_{13}	Cyclohexylethyl	9.41	9.34	9.29
2	n-C_7H_{15}	Cyclohexylethyl	9.42	9.34	9.36
3	n-C_7H_{15}	4-$FC_6H_4CH_2$	8.86	9.23	9.08
4	n-C_8H_{17}	Cyclohexylethyl	9.43	9.35	9.40
5	n-C_8H_{17}	4-$ClC_6H_4CH_2$	8.79	8.91	8.90
6	n-C_8H_{17}	4-$BrC_6H_4CH_2$	8.86	8.87	8.90

On analyzing the applicability domain (AD) in the Williams plot (Fig. 5) of the model based on the whole data set (Table 6; Eqs. 2a and 5a), one compound (**17**; Table 1) has been identified as an obvious 'outlier' for the antimalarial activity against *P. falciparum* strain D_6 if the limit of normal values for the Y outliers (response outliers) was set as 3 × (standard deviation) units. None of the compounds was found to have leverage (h) values greater than the threshold leverages (h*). For both the training-set and test-set, the suggested models match the high quality parameters with good fitting power and the capability of assessing external data. Furthermore, almost all of the compounds were within the applicability domain of the proposed models and were evaluated correctly.

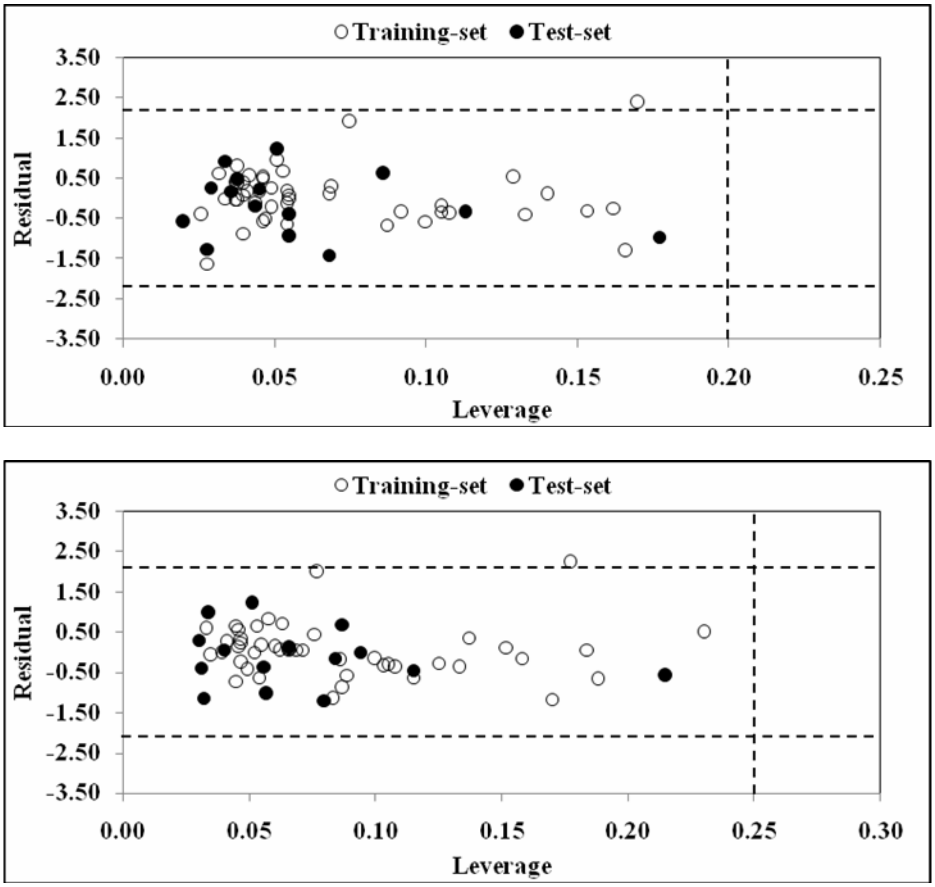


Fig. 5: Williams plot for antimalarial activity (D_6) of training-set and test-set compounds (Table 1). The horizontal dotted line refers to the residual limit ± 3.0 × (standard deviation) and the vertical dotted line represents threshold leverage, h* (= 0.200 and 0.250 for three- and four-descriptor models respectively)

Table 6: Models derived from whole data set (n = 60) of prodiginines for antimalarial activity against D_6 strain of *P. falciparum*

Model	r	s	F	Q^2_{LOO}	Q^2_{L5O}	Eq.
pIC_{50} = 5.046 + 3.385 (0.389) IC5 + 2.048 (0.439) JGI5 – 1.971 (0.437) MATS6m	0.848	0.739	47.638	0.659	0.656	2(a)
pIC_{50} = 6.091 + 3.018 (0.389) IC5 + 2.007 (0.415) JGI5 – 2.153 (0.417) MATS6m – 1.269 (0.451) GATS5e	0.868	0.698	42.120	0.694	0.685	5(a)

A considerable number of synthetic prodiginines have been tested alongside chloroquine (CQ) against multidrug-resistant (MDR) Dd_2 strain. The activity profile has been included, as pIC_{50} (Dd_2), in Table 1 on molar basis. In order to infer any significant change in the activity of each analogue between their in vitro antimalarial activity against *P. falciparum* pansensitive D_6 with CQ as a reference drug, attempt has been made to find a relationship between their pIC_{50} values.

To this effect, a highly significant correlation between pIC_{50}s of D_6 and Dd_2 was obtained and the same is given through Equation (6).

$$pIC_{50}(Dd_2) = 1.034\ (0.050)\ pIC_{50}(D_6) - 0.197$$
$$r = 0.992,\ s = 0.176,\ F\ (1, 50) = 3274.773 \qquad \ldots(6)$$

Above equation has reflected upon the fact the two activities are exactly inter-correlated and present synthetic prodiginines are equally effective against *P. falciparum* pansensitive D_6 and MDR Dd_2.

CONCLUSION

The activity against *P. falciparum* D_6 strain of natural and synthetic prodiginines has been quantitatively analyzed in terms of chemometric descriptors. The statistically validated QSAR models provided rationales to explain the antimalarial activity of these congeners. The descriptors identified through CP-MLR analysis have highlighted the role of the mean square distance index (Balaban) (MSD), the information content index (neighbourhood symmetry of 5-order (IC5), the distance/ detour ring index of order 9 (D/Dr09), the mean topological charge indices of order 5 and order 6 (JGI5 and JGI6), the Moran autocorrelation – lag 6/ weighted by atomic masses (MATS6m) and the Geary autocorrelation – lag 5/ weighted by atomic Sanderson electronegativities (GATS5e). The statistical significant models in three- and four-descriptors have been derived to explain the antimalarial activity against *P. falciparum* D_6 and the main contributors in the highest significant models were the descriptors IC5, JGI5, MATS6m and GATS5e. For a compound to be more potent, the higher values of descriptors IC5 and JGI5 and lower values of descriptors MATS6m and GATS5e are conducive. The statistics emerged from the test-set have validated the identified significant models. A few new compounds, having activity more the highest active congener, have been suggested for further exploration. PLS analysis has further confirmed the dominance of the CP-MLR identified descriptors. Applicability domain analysis revealed that the suggested models have acceptable predictability. Except one "outlier" (S.No. **17**, Table 1), all the compounds remained within the applicability domain of the proposed models and were evaluated correctly.

ACKNOWLEDGEMENT

The financial support provided by the University Grants Commission, New Delhi to PS, under the scheme of Emeritus Fellowship, is thankfully acknowledged. Authors are also thankful to the Institution for providing necessary facilities to complete this study.

REFERENCES

1. B. M. Greenwood, D. A. Fidock, D. E. Kyle, S. H. Kappe, P. L. Alonso, F. H. Collins and P. E. Duffy, J. Clin. Invest, **118**, 1266 (2008).
2. J. E. Hyde, Trends Parasitol., **21**, 494 (2005).
3. A. M. Dondorp, F. Nosten, P. Yi, D. Das, A. P. Phyo, J. Tarning, K. M. Lwin, F. Ariey, W. Hanpithakpong, S. J. Lee, P. Ringwald, K. Silamut, M. Imwong, K. Chotivanich, P. Lim, T. Herdman,

S. S. An, S. Yeung, P. Singhasivanon, N. P. Day, N. Lindegardh, D. Socheat and N. J. White, N. Engl. J. Med., **361**, 455 (2009).

4. K. Kaur, M. Jain, T. Kaur and R. Jain, Bioorg. Med. Chem., **17**, 3229 (2009).
5. F. Alihosseini, K. S. Ju, J. Lango, B. D. Hammock and G. Sun, Biotechnol. Prog., **24**, 742 (2008).
6. J. Regourd, A. Al-Sheikh Ali and A. Thompson, J. Med. Chem., **50**, 1528 (2007).
7. R. D. Alessio, A. Bargiotti, O. Carlini, F. Colotta, M. Ferrari, P. Gnocchi, A. Isetta, N. Mongelli, P. Motta, A. Rossi, M. Rossi, M. Tibolla and E. Vanotti, J. Med. Chem., **43**, 257 (2000).
8. T. F. Ho, C. J. Ma, C. H. Lu, T. T. Tsai, Y. H. Wei, J. S. Chang, J. K. La, P. J. Cheuh, C. T. Yeh, P. C. Tang, J. H. T. Chang, J. L. Ko, F. S. Liu, H. C. E. Yen and C. C. Chang, Toxicol. Appl. Pharmacol., **225**, 318 (2007).
9. A. J. Castro, Nature, **213**, 903 (1967).
10. N. N. Gerber, J. Antibiot. (Tokyo), **28**, 194 (1975).
11. D. E., Jr., Davidson, D. O. Johnsen, P. Tanticharoenyos, R. L. Hickman and K. E. Kinnamon, Am. J. Trop. Med. Hyg., **25**, 26 (1976).
12. M. Isaka, A. Jaturapat, J. Kramyu, M. Tanticharoen and Y. Thebtaranonth, Antimicrob. Agents Chemother., **46**, 1112 (2002).
13. J. E. H. Lazaro, J. Nitcheu, R. Z. Predicala, G. C. Mangalindan, F. Nesslany, D. Marzin, G. P. Concepcion and B. Diquet, J. Nat. Toxins, **11**, 367 (2002).
14. K. Papireddy, M. Smilkstein, J. X. Kelly, Shweta, S. M. Salem, M. Alhamadsheh, S. W. Haynes, G. L. Challis and K. A. Reynolds, J. Med. Chem., **54**, 5296 (2011).
15. SYSTAT, Version 7.0; SPSS Inc 444 North Michigan Avenue, Chicago IL, 60611.
16. Chemdraw Ultra 6.0 and Chem. 3D Ultra, Cambridge Soft Corporation, Cambridge, USA. http://www.cambridgesoft.com
17. Dragon Software (Version 1.11-2001) by R. Todeschini and V. Consonni, Milano, Italy. http//www.talete.mi.it/dragon.htm
18. Y. S. Prabhakar, QSAR Comb. Sci., **22**, 583 (2003).
19. A. Golbraikh, A. Tropsha, J. Mol. Graph. Model, **20**, 269 (2002).
20. S. Sharma, Y. S. Prabhakar, P. Singh and B. K. Sharma, Eur. J. Med. Chem., **43**, 2354 (2008).
21. S. Sharma, B. K. Sharma, S. K. Sharma, P. Singh and Y. S. Prabhakar, Eur. J. Med. Chem., **44**, 1377 (2009).
22. B. K. Sharma, P. Pilania, P. Singh and Y. S. Prabhakar, SAR QSAR Environ. Res., **21**, 169 (2010).
23. B. K. Sharma, P. Singh, K. Sarbhai and Y. S. Prabhakar, SAR QSAR Environ. Res., **21**, 369 (2010).
24. B. K. Sharma, P. Pilania, K. Sarbhai, P. Singh and Y. S. Prabhakar, Mol. Divers., **14**, 371 (2010).
25. B. K. Sharma, P. Singh, M. Shekhawat, K. Sarbhai and Y. S. Prabhakar, SAR QSAR Environ. Res., **22**, 365 (2011).
26. S. –S. So and M. Karplus, J. Med. Chem., **40**, 4347 (1997).
27. Y. S. Prabhakar, V. R. Solomon, R. K. Rawal, M. K. Gupta and S. B. Katti, QSAR Comb. Sci., **23**, 234 (2004).

28. P. Gramatica, QSAR Comb. Sci., **26**, 694 (2007).

29. L. Eriksson, J. Jaworska, A. P. Worth, M. T. D. Cronin, R. M. McDowell and P. Gramatica, Environ. Health Persp., **111**, 1361 (2003).

30. S. Wold, Technometrics, **20**, 397 (1978).

31. N. Kettaneh, A. Berglund and S. Wold, Comput. Stat. Data Anal., **48**, 69 (2005).

32. L. Stahle and S. Wold, In: G. P. Ellis and G. B. West, (Eds.), “Progress in Medicinal Chemistry”, **Vol. 25**, Elsevier Science Publishers, BV (1988) p. 291.