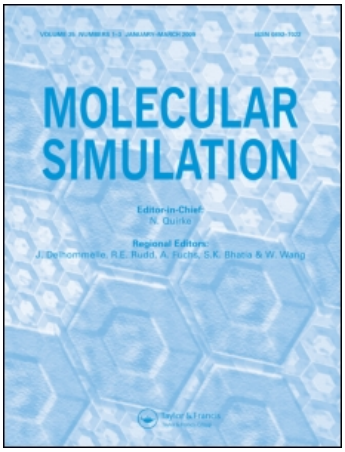


Molecular Simulation
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713644482

Chemometric modelling of antimalarial activity of aryltriazolylhydroxamates

Probir Kumar Ojha[a]; Kunal Roy[a]
[a] Division of Medicinal and Pharmaceutical Chemistry, Drug Theoretics and Cheminformatics Laboratory, Department of Pharmaceutical Technology, Jadavpur University, Kolkata, India

Online publication date: 02 November 2010

To cite this Article Ojha, Probir Kumar and Roy, Kunal(2010) 'Chemometric modelling of antimalarial activity of aryltriazolylhydroxamates', Molecular Simulation, 36: 12, 939 — 952
To link to this Article: DOI: 10.1080/08927022.2010.492835
URL: http://dx.doi.org/10.1080/08927022.2010.492835

Molecular Simulation
Vol. 36, No. 12, October 2010, 939–952

Taylor & Francis
Taylor & Francis Group

Chemometric modelling of antimalarial activity of aryltriazolylhydroxamates

Probir Kumar Ojha and Kunal Roy*

Division of Medicinal and Pharmaceutical Chemistry, Drug Theoretics and Cheminformatics Laboratory, Department of Pharmaceutical Technology, Jadavpur University, Kolkata 700 032, India

(*Received 22 February 2010; final version received 29 April 2010*)

We have performed quantitative structure–activity relationship (QSAR) and quantitative activity–activity relationship (QAAR) studies for aryltriazolylhydroxamates having antimalarial activity data against both chloroquine-sensitive (D6 clone) and chloroquine-resistant (W2 clone) strains of *Plasmodium falciparum* to understand the relationships between the biological activity and molecular properties for the design of new compounds. The QSAR studies were performed using 35 compounds among which 26 molecules were taken using *k*-means clustering technique in the training set for the derivation of the QSAR models and nine molecules were kept as the test-set compounds to evaluate the predictive ability of the derived models. The chemometric tool used for the analysis was the genetic function approximation. The developed models were analysed in terms of their predictive ability, and comparable results were obtained for cross-validated predictive variance (Q^2) and externally predicted variance (R^2_{pred}) values (0.761 and 0.829, respectively, for the D6 model, 0.708 and 0.748, respectively, for the W2 model and 0.984 and 0.982, respectively for the QAAR model). The QSAR models suggest that the number of methylene groups (between the triazolyl and hydroxamate moieties) and partially negatively charged surface areas of the molecules are important parameters for the antimalarial activity.

Keywords: genetic function approximation; *k*-means cluster; quantitative structure–activity relationship; quantitative activity–activity relationship; aryltriazolylhydroxamate

1. Introduction

Despite considerable scientific advances and development of modern drugs, malaria still remains one of the deadliest infectious diseases, and has a tremendous morbidity and mortality impact in the developing world. In 2006, there were an estimated 247 million malaria cases among 3.3 billion people at risk, causing nearly a million deaths, mostly of children under 5 years [1]. In a 2008 report, 109 countries were endemic for malaria, 45 out of them within the World Health Organization African region [1]. There are four major species of the malaria parasite: *Plasmodium falciparum, Plasmodium vivax*, *Plasmodium malariae* and *Plasmodium ovale*. Among the four species, *P. falciparum* is responsible for more than 95% of malaria-related morbidity and mortality [2]. The treatment and prevention of malaria are becoming increasingly ineffective with chloroquine, mefloquine and other frontline drugs [8-aminoquinolines, quinolinemethanols, sulphonamides and dihydrofolate reductase (DHFR) inhibitors] [3]. The treatment and prevention of malaria now includes long-lasting insecticidal nets and artemisinin-based combination therapy (ACT), supported by indoor residual spraying of insecticide and intermittent preventive treatment in pregnancy [1]. These methods are still being used and continue to be the foundations of malaria control and elimination programmes. Unfortunately, the global emergence of resistance of *P. falciparum* to commonly used antimalarial drugs has rendered these drugs useless in many endemic areas [4]. In recent years, most countries have been using ACT. However, there are concerns that resistance against these drugs is already beginning to evolve [5]. As a result, discovering and developing new molecules effective against resistant strains is one of the greatest challenges [6]. Computational tools such as quantitative structure–activity relationships (QSARs) are one of the most important applications of chemometrics, giving information to the medicinal chemist to understand the relationship between biological activity and molecular properties for the design of new compounds acting on a specific target. Thus, QSAR models can be used to design new compounds and to predict the activity of untested compounds [7].

QSARs have been reported for antimalarials by many groups of researchers in recent years. Adane and Bharatam [8] reported a 3D-QSAR study [Comparative molecular field analysis (CoMFA) and comparative molecular similarity indices analysis (CoMSIA)] of cycloguanil derivatives, which are reported as growth inhibitors of *P. falciparum* clone (T9/94 RC17), to investigate the structural requirements for the activities of these compounds and to derive predictive models. Cruz-Monteagudo et al. [9] reported computational modelling tools for the design of potent antimalarial bisbenzamidines exploiting the antimalarial potential of pentamidine using GETAWAY descriptors. Katritzky et al. [10] reported

*Corresponding author. Email: kunalroy_in@yahoo.com

ISSN 0892-7022 print/ISSN 1029-0435 online

DOI: 10.1080/08927022.2010.492835
http://www.informaworld.com

a QSAR model of the antimalarial activity of two diverse sets of compounds for each of two strains, D6 and NF54, of *P. falciparum* by using CODESSA PRO software. Parenti et al. [11] explored a 3D pharmacophore model using CATALYST software on a diverse set of PfDHFR-TS (PfDHFR-TS; TS refers to thymidylate synthase bound to DHFR in *P. falciparum*) inhibitors, including cycloguanil and pyrimethamine derivatives.

In the present study, we have performed a QSAR study of aryltriazolylhydroxamates having activity data against both chloroquine-sensitive (D6 clone) and chloroquine-resistant (W2 clone) strains of *P. falciparum* to understand the relationship between the biological activity and molecular properties of the compounds for the design of new compounds. We have also performed the quantitative activity–activity relationship (QAAR) study by taking the *in vitro* activity data against chloroquine-resistant strain (W2 clone) as the response and those against chloroquine-sensitive strain (D6 clone) as one of the predictor variables to calculate one activity when the other is known. It may be mentioned here that these aryltriazolylhydroxamates belong to the class of histone deacetylase inhibitors. No QSAR reports are available to date on this class of compounds.

2. Material and methods

2.1 Data-set

In the present study, a data-set of 35 aryltriazolylhydroxamates [12] having antimalarial activity data (Table 1) against both chloroquine-sensitive (D6 clone) and chloroquine-resistant (W2 clone) strains of *P. falciparum* was selected. The *P. falciparum* growth inhibition was reported for all the compounds by a parasite lactate dehydrogenase assay using Malstat reagent [13]. In our QSAR study, we have used the negative logarithm of the reported biological activity ($pIC_{50} = -\log IC_{50}$) against both D6 and W2 clones. The general structure of aryltriazolylhydroxamates is given in Table 1. It is found that a methylene spacer containing three to eight methylene fragments separates the aryltriazolyl from the hydroxamate moiety. The aryltriazolyl ring is substituted with different aromatic and heteroaromatic rings.

2.2 Quantitative structure–activity relationship

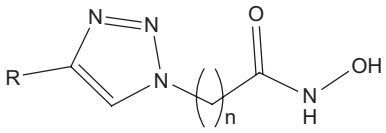
2.2.1 Descriptors

The analyses were performed using electronic (dipole-mag, HOMO, LUMO and Sr), spatial (radius of gyration, Jurs descriptors, Area, PMI-mag, Density, Vm), thermodynamic (ALogP, ALogP98, MolRef, MR, LogP), structural (H-bond donor, H-bond acceptor, Rotlbonds) as well as different topological parameters (E-state index, kappa shape index, molecular connectivity index, subgraph count, Balaban, Wiener and Zagreb indices). All the descriptors were calculated using Descriptor + module of the Cerius2 version 4.10 software [14]. The categorical list of the descriptors used in the development of QSAR models is reported in Table 2. Finally, we have tried to develop a QAAR model by taking the *in vitro* activity data against chloroquine-resistant strain (W2 clone) as the response and those against chloroquine-sensitive strain (D6 clone) as one of the predictor variables.

2.2.2 Cluster analysis

The goal of any QSAR modelling is to develop a model which should be capable of making accurate and reliable predictions of biological activities of new compounds. However, internal validation does not ascertain that the model will perform well on a new set of data. In many cases, appropriate external data-set is not available for prediction purpose. Thus, the whole data-set is divided into a training set and a test set or external evaluation set. The models developed from the training set are externally validated using the test-set molecules. Predictive capacity of a model for new chemical entities is influenced by chemical nature of the training-set molecules used for the development of the model [15–17]. The test-set molecules will be predicted well when they are structurally very similar to the training-set molecules. The reason is that the model has captured all features common to the training-set molecules. There are different techniques available for division of the data-set into training and test sets such as statistical molecular design, self-organising map, clustering, Kennard–Stone selection, sphere exclusion, etc. [18]. In the present study, we have used the clustering technique as the method for training-set selection. Cluster analysis [19] is a technique to arrange the objects into groups. There are two types of clustering: (i) hierarchical clustering and (ii) non-hierarchical clustering. Hierarchical clustering techniques proceed by either a series of successive mergers or a series of successive divisions. Agglomerative hierarchical methods start with the individual objects. Thus, there are initially as many clusters as objects. The most similar objects are first grouped and these initial groups are merged according to their similarities. Eventually, as the similarity decreases, all subgroups are fused into a single cluster. Divisive hierarchical methods work in the opposite direction. An initial single group of objects is divided into two subgroups such that the objects in one subgroup are far from the objects in the other. These subgroups are then further divided into dissimilar subgroups; the process continues until there are as many subgroups as objects – that is, until every object forms a group. Non-hierarchical clustering techniques are designed to group items, rather than variables, into a collection of k clusters. The number of clusters, k, may either be specified in advance or determined as part of the clustering procedure. One of the important non-hierarchical techniques is k-means

Table 1. Structural features, observed and calculated antimalarial activity values of aryltriazolylhydroxamate derivatives



.

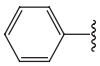
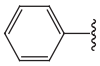
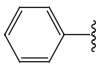
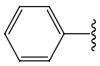
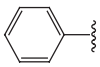
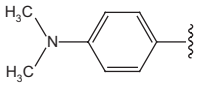
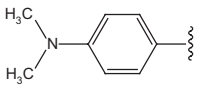
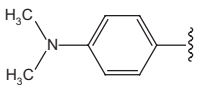
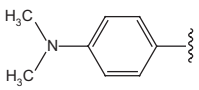
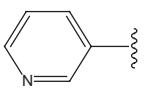
Sl. no.	*R*	*n*	Observed antimalarial activity [pIC_{50} (IC_{50} in mmol)]		Antimalarial activity calculated from different models		
			D6 clone	W2 clone	Equation (1)	Equation (2a)	Equation (3)
1		4	3.319	3.336	3.233	3.589	3.282
2[a]		5	3.494	3.324	3.369	3.317	3.467
3		6	3.880	4.071	3.806	3.550	3.878
4[a]		7	2.671	2.639	2.542	2.408	2.593
5		8	2.077	2.061	1.885	2.120	1.985
6		5	3.669	3.916	3.872	4.247	3.787
7		6	4.131	4.015	4.093	3.763	4.145
8[a]		7	3.295	3.271	3.313	3.250	3.257
9		8	2.300	2.338	2.516	2.652	2.356
10		5	3.197	3.128	2.961	3.116	3.152

Table 1 – *continued*

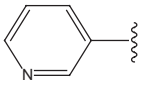
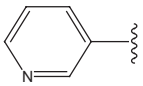
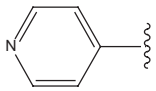
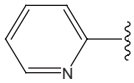
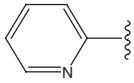
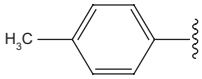
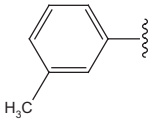
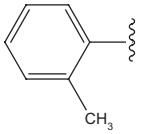
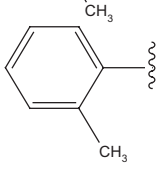
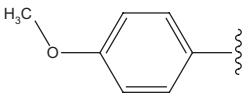
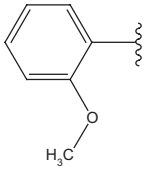
Sl. no.	*R*	*n*	Observed antimalarial activity [pIC_{50} (IC_{50} in mmol)]		Antimalarial activity calculated from different models		
			D6 clone	W2 clone	Equation (1)	Equation (2a)	Equation (3)
11		6	3.583	3.553	3.358	3.281	3.562
12		7	2.307	2.255	2.257	2.211	2.206
13[a]		5	3.150	3.318	2.878	2.984	3.102
14		5	2.446	2.644	2.981	3.225	2.626
15		6	3.623	3.859	3.401	3.430	3.877
16		5	3.346	3.314	3.629	3.464	3.310
17		5	3.560	3.627	3.404	3.204	3.538
18		5	2.363	2.170	3.159	2.961	2.266
19		6	3.401	3.334	3.641	3.237	3.369
20		5	3.369	3.529	3.450	3.386	3.571
22		5	3.307	3.494	3.057	3.387	3.505

Table 1 – *continued*

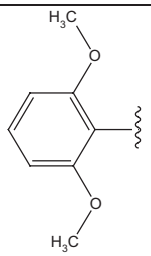
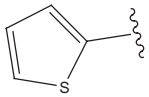
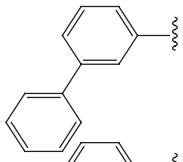
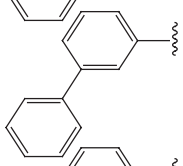
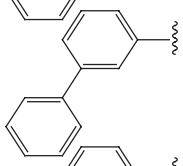
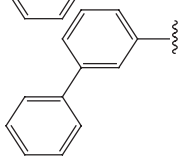
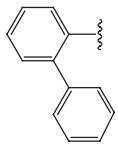
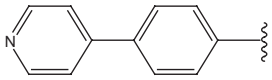
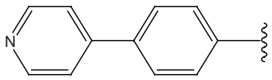
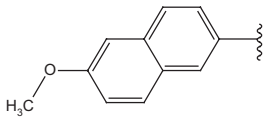
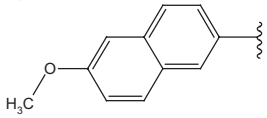
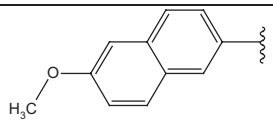
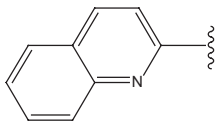
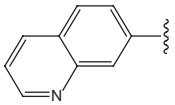
Sl. no.	*R*	*n*	Observed antimalarial activity [pIC_{50} (IC_{50} in mmol)]		Antimalarial activity calculated from different models		
			D6 clone	W2 clone	Equation (1)	Equation (2a)	Equation (3)
23		5	2.223	2.109	2.198	2.049	2.117
24[a]		5	3.785	3.885	3.612	3.622	3.777
25		5	3.971	4.130	3.762	3.652	3.975
26		6	4.164	4.164	3.885	3.330	4.182
27[a]		7	2.654	2.765	3.131	2.847	2.575
28		8	2.069	2.089	2.190	2.321	2.111
29		5	2.222	2.130	2.407	2.148	2.116
30[a]		5	3.888	3.869	3.937	3.898	3.886
32		7	3.632	3.474	3.515	3.415	3.614
33[a]		5	4.005	4.058	4.153	4.643	4.011
34		6	3.890	3.842	4.313	4.266	3.889

Table 1 – *continued*

Sl. no.	*R*	*n*	Observed antimalarial activity [pIC_{50} (IC_{50} in mmol)] D6 clone	W2 clone	Antimalarial activity calculated from different models Equation (1)	Equation (2a)	Equation (3)
35		7	3.533	3.365	3.486	3.657	3.509
36		5	3.994	3.981	3.407	4.251	4.003
37[a]		5	3.792	3.994	3.625	4.346	3.805

[a] Test set compounds.

Table 2. Categorical list of the descriptors used in the development of QSAR models.

Category of descriptors	Name of the descriptors
Topological	Jx, $^{1}\kappa$, $^{2}\kappa$, $^{3}\kappa$, $^{1}\kappa_{am}$, $^{2}\kappa_{am}$, $^{3}\kappa_{am}$, ϕ, SC-0, SC-1, SC-2, SC-3_P, SC-3_C, $^{1}\chi$, $^{2}\chi$, $^{3}\chi_{p}$, $^{3}\chi_{c}$, $^{0}\chi^{v}$, $^{1}\chi^{v}$, $^{2}\chi^{v}$, $^{3}\chi_{p}^{v}$, $^{3}\chi_{c}^{v}$, Wiener, Zagreb, S_sCH$_3$, S_ssCH$_2$, S_aaCH, S_aaaC,S_sssCH, S_dssC, S_aasC, S_ssssC, S_sssN, S_sOH, S_dO, S_ssO, S_ssS, S_aaS
Structural	MW, Rotlbonds, Hbond acceptor, Hbond donor
Physico-chemical	AlogP, AlogP98, MR, MolRef, LogP
Electronic	Dipole-mag, Sr
Spatial	RadOfGyration, Jurs-SASA, Jurs-PPSA_1, Jurs-PNSA_1, Jurs-DPSA_1, Jurs-PPSA_2, Jurs-PNSA_2, Jurs-DPSA_2, Jurs-PPSA_3, Jurs-PNSA_3, Jurs-DPSA_3, Jurs-FPSA_1, Jurs-FNSA_1, Jurs-FPSA_2, Jurs-FNSA_2, Jurs-FPSA_3, Jurs-FNSA_3, Jurs-WPSA_1, Jurs-WNSA_1, Jurs-WPSA_2, Jurs-WNSA_2, Jurs-WPSA_3, Jurs-WNSA_3, Jurs-RPCG, Jurs-RNCG, Jurs-RPCS, Jurs-RNCS, Jurs-TPSA, Jurs-TASA, Jurs-RPSA, Jurs-RASA, Area, Vm, Density, PMI-mag

clustering [20], which has been used in the present study. The *k*-means algorithm assigns each item to the cluster having the nearest centroid (mean). In its simplest version, the process is composed of three steps [21]:

(1) partitioning the items into *k* initial clusters;
(2) assigning an item to the cluster whose centroid (mean) is nearest (distance is usually computed using Euclidean distance with standardised observations) followed by recalculation of the centroid for the cluster receiving the new item and for the cluster losing the item;
(3) step 2 is repeated until no more reassignments take place.

In this report, the total data-set ($n = 35$) was divided into internal evaluation (training) set ($n = 26$) and external evaluation (test) set ($n = 9$) (75 and 25%, respectively, of the total number of compounds) based on clusters obtained from *k*-means clustering applied on a standardised topological and structural descriptor matrix by using SPSS software [22]. All the parameters were standardised to values between 0 and 1, and the whole data-set was clustered into four subgroups from each of which 25% of compounds were selected as the members of the test set. Serial numbers of compounds under different clusters are shown in Table 3.

Table 3. Serial numbers of compounds under different clusters.

Cluster number	Serial number of compounds
1	**4, 5, 12, 19, 20, 22, 36, 37**
2	**6, 7, 8, 9, 23**
3	**25, 26, 27, 28, 29, 30, 32, 33, 34, 35**
4	**1, 2, 3, 10, 11, 13, 14, 15, 16, 17, 18, 24**

2.2.3 *Genetic function approximation*

Genetic algorithms are derived from an analogy with the evolution of DNA [23]. The GFA algorithm was initially anticipated by (i) Holland's genetic algorithm and (ii) Friedman's multivariate adaptive regression splines (MARS) algorithm. In this algorithm, an individual or model is represented as 1D string of bits. A distinctive feature of GFA is that it produces a population of models (e.g. 100), instead of generating a single model, as do most other statistical methods. The genetic algorithm makes superior models to those developed using stepwise regression techniques because it selects the basis functions genetically. Descriptors, which are selected by this algorithm, are subjected to multiple linear regression (MLR) for generation of models. A 'fitness function' or lack of fit (LOF) is used to estimate the quality of an individual or model, so that best individual or model receives the best fitness score. The error measurement term LOF is determined by the following equation:

$$\mathrm{LOF} = \frac{\mathrm{LSE}}{\left(1 - \frac{c + d*p}{M}\right)^2}.$$

In the above equation, c is the number of basis functions (other than the constant term); d is the smoothing parameter (adjustable by the user); M is the number of samples in the training set; LSE is the least-squares error; and p is the total number of features contained in all basis functions.

Once models in the population have been rated using the LOF score, the genetic crossover operation is repeatedly performed. Initially, two good models are probabilistically selected as parents and each parent is randomly cut into two pieces and a new model (child) is generated using a piece from each parent. After many mating steps, i.e. genetic crossover-type operation, average fitness of individuals (models) in the population increases as good combination of genes are discovered and spread through the population. It can build not only linear models but also higher order polynomials, splines and Gaussians. In our present work, linear and spline terms have been used. For the development of GFA models, Cerius2 4.10 version [14] has been used. The mutation probabilities were kept at 50% with 5000 iterations. Smoothness (d) factor was kept at 1.00. The initial equation length value was selected as 4 and the length of the final equations was not fixed.

2.2.4 *Partial least squares*

The basic concept of partial least-squares (PLS) regression was originally developed by Wold [24,25]. PLS regression is extensively used in CoMFA and CoMSIA. Recently, PLS has been used by combination with other mathematical methods [e.g. genetic/PLS (G/PLS), factor analysis-PLS (FA-PLS) and orthogonal signal correction-PLS (OSC-PLS)] to give better performance in QSAR–QSPR analyses. PLS regression is an extension of the MLR model, which can handle data with strongly correlated and/or noisy or numerous X variables [26]. It gives a reduced solution, which is statistically more robust than MLR. The linear PLS model finds 'new variables' (latent variables or X scores), which are linear combinations of the original variables. To avoid overfitting, a strict test for the significance of each consecutive PLS component is necessary and then stopping when the components are non-significant. For PLS, the 'leave-one-out' (LOO) method was used for cross-validation to obtain the optimum number of components. Cross-validation is a reliable and commonly used method for selecting optimum number of components [27]. However, recently, it has been shown that from the viewpoint of external predictability, the choice of variables for PLS based on internal validation may not be optimum [28]. Application of PLS allows the construction of larger QSAR equations while still avoiding overfitting and eliminating most variables. Cross-validation ensures that the QSAR equations are selected based on their ability to predict the data rather than to fit the data [29]. In the present report, PLS analysis has been done to remove the intercorrelation problem with the descriptors selected by the GFA-MLR technique.

2.2.5 *Validation methods*

The robustness of the models should be verified by using different validation criteria. For the validation of QSAR models, usually four strategies [30] are adopted: (i) internal validation or cross-validation, (ii) validation by dividing the data-set into a training set and a test set, (iii) data randomisation or *Y*-scrambling and (iv) true external validation by the application of a model on new external data. In the present study, due to the lack of true external evaluation set, the total data-set has been divided into internal evaluation (training) set and external evaluation (test) set. Therefore, we have applied only the first three validation techniques. Most of the QSAR modelling methods implement the LOO or leave-many-out (LMO) cross-validation procedure, which are internal validation techniques. The outcome from the cross-validation procedure is cross-validated R^2 (LOO-Q^2 or LMO-Q^2), which is used as a criterion of both robustness and predictive ability of the model. In this paper, we have performed the LOO cross-validation method as the internal validation tool. The cross-validated determination coefficient (LOO-Q^2) is calculated according to this equation:

$$Q^2 = 1 - \frac{\sum (Y_{\mathrm{obs(training)}} - Y_{\mathrm{pred(training)}})^2}{\sum (Y_{\mathrm{obs(training)}} - \overline{Y}_{\mathrm{training}})^2}.$$

In the above equation, $\overline{Y}_{\text{training}}$ represents the average activity value of the training set while $Y_{\text{obs(training)}}$ and $Y_{\text{pred(training)}}$ represent, respectively, observed and predicted activity values of training-set compounds. Often, a high Q^2 value ($Q^2 > 0.5$) is considered as a proof of high-predictive ability of the model [31].

Models are generated based on the training-set compounds and predictive capacity of the models is judged based on the predictive R^2 (R^2_{pred}) values calculated according to the following equation [32]:

$$R^2_{\text{pred}} = 1 - \frac{\sum(Y_{\text{obs(test)}} - Y_{\text{pred(test)}})^2}{\sum(Y_{\text{obs(test)}} - \overline{Y}_{\text{training}})^2}.$$

In the above equation, $Y_{\text{pred(test)}}$ and $Y_{\text{obs(test)}}$ indicate, respectively, predicted and observed activity values of the test-set compounds, and $\overline{Y}_{\text{training}}$ indicates the mean activity value of the training-set compounds. The value of R^2_{pred} for an acceptable model should be greater than 0.5.

It has been previously shown [28,32] that the R^2_{pred} may not be sufficient to indicate external predictivity of a model. The value of R^2_{pred} is mainly controlled by the term $\sum(Y_{\text{obs(test)}} - \overline{Y}_{\text{training}})^2$, i.e. the sum of squared differences between the observed values of test-set compounds and the mean observed value of the training-set compounds. Thus, it may not truly reflect the predictive capability on a new data-set. In addition, the square of the correlation coefficient (r^2) between the observed and predicted values of the test-set compounds does not necessarily mean that the predicted values are very near to the corresponding observed activity values (there may be a considerable numerical difference between the values, though maintaining an overall good intercorrelation). Therefore, for a better indication of external predictive potential of the model, a modified r^2 ($r^2_{\text{m(test)}}$) has been introduced by the following equation [28,33]:

$$r^2_{\text{m(test)}} = r^2\left(1 - \sqrt{r^2 - r_0^2}\right).$$

The value of $r^2_{\text{m(test)}}$ should be greater than 0.5 for an acceptable model.

Previously, the concept r^2_{m} was applied only to the test-set prediction [28,33], but it can as well be applied for the training set if one considers the correlation between the observed and LOO-predicted values of the training-set compounds [34]. More interestingly, this can be used for the whole set considering LOO-predicted values for the training set and predicted values for the test-set compounds. The advantages of such consideration are as follows: (i) unlike external validation parameters (R^2_{pred}, etc.), the $r^2_{\text{m(overall)}}$ statistic is not based only on a limited number of test-set compounds. It includes prediction for both test-set and training-set (using LOO predictions) compounds. Thus, this statistic is based on the prediction of a comparably large number of compounds. In many cases, the test-set size is considerably small, and the regression-based external validation parameter may be less reliable and highly dependent on individual test-set observations. In such cases, the $r^2_{\text{m(overall)}}$ statistic may be advantageous and (ii) in many cases, comparable models are obtained where some models show comparatively better internal validation parameters and some other models show relatively superior external validation parameters. This may create problem in selecting the final model. The $r^2_{\text{m(overall)}}$ statistic may be used for the selection of the best predictive models from among the comparable models. For the present QSAR study, we have determined r^2_{m} values for both training (based on LOO-predicted values) and test sets and also for the whole set for the reported models, and the results are shown in Table 4.

Further statistical significance of the relationships between the antimalarial activity (D6 and W2 clones) and descriptors was obtained from randomisation (*Y*-randomisation) tests. There are two types of randomisation tests: process randomisation and model randomisation. In the case of the process randomisation, the values of the dependent variable are randomly scrambled and variable selection is done freshly from the whole descriptor matrix. The process randomisation has been performed at 95% confidence level. In the case of the model randomisation, the *Y* column entries are scrambled and new QSAR models are developed using the same set of variables as present in the non-randomised model. The model randomisation has been performed at 99% confidence level. If the score of the non-random QSAR model is significantly better than that of the random models, then that model should be considered as a statistically robust model [35,36]. We have used a parameter R^2_{p} [35,36], which penalises the model R^2 for a small difference between the square of the mean correlation coefficient (R^2_{r}) of randomised models and the square of the correlation coefficient (R^2) of the non-randomised model. The above-mentioned novel parameter can be calculated by the following equation:

$$R^2_{\text{p}} = R^2\sqrt{R^2 - R^2_{\text{r}}}.$$

This novel parameter R^2_{p} ensures that the models developed are not obtained by chance. For an acceptable QSAR model, the value of R^2_{p} should be greater than 0.5. The values of R^2, R^2_{r} and R^2_{p} based on the process and model randomisation tests for different models are reported in Table 5.

2.3 Software

MINITAB [37] was used for PLS regression. Cerius2 version 4.10 [14] was used for the calculation of descriptors and genetic analyses. STATISTICA [38] was

Table 4. Statistical quality and validation parameters of different models.

Equation nos.	Response variable	Type of model	Descriptors	Equation quality									
				R^2	R^2_a	F	S	PRESS	Q^2	$r^2_{m(LOO)}$	R^2_{pred}	$r^2_{m(test)}$	$r^2_{m(overall)}$
(1)	pIC_{50} (D6 clone)	GFA Spline	JursFPSA-1, RadOfGyration, S-ssCH$_2$, Jurs-RNCS	0.826	0.793	24.90	0.320	2.952	0.761	0.590	0.829	0.731	0.608
(2)	pIC_{50} (W2 clone)	GFA Linear	PHI, Jurs-RNCS, Jurs-WNSA-1, Jurs-WPSA-2, RadOfGyration	0.822	0.777	18.42	0.350	4.005	0.708	0.536	0.748	0.540	0.541
(2a)		PLS	PHI, Jurs-RNCS, Jurs-WNSA-1, Jurs-WPSA-2, RadOfGyration	0.752	0.705	15.91	0.402	4.941	0.639	0.602	0.738	0.507	0.583
(3)	pIC_{50} (W2 clone) (QAAR)	GFA Spline	pIC_{50}_D6, Sr,CHI-V-3_C	0.987	0.986	576.14	0.088	0.213	0.984	0.959	0.942	0.882	0.949

Table 5. Results of randomisation tests.

Equation No.	Response variable	Descriptor type	Model type	R^2	Process randomisation		Model randomisation	
					R^2_r	R^2_p	R^2_r	R^2_p
(1)	pIC_{50} (D6 clone)	Topological with 3D descriptors (D6 clone)	GFA Linear + Spline	0.826	0.293	0.516	0.143	0.607
(2)	pIC_{50} (W2 clone)	Topological with 3D descriptors (W2 clone)	GFA Linear	0.822	0.197	0.569	0.194	0.571
(3)	pIC_{50} (W2 clone) (QAAR)	Topological with 3D descriptors	GFA Linear + Spline	0.987	0.288	0.817	0.098	0.924

used to determine the LOO-predicted values of the training-set compounds and correlation matrix of the descriptors.

3. Results and discussion

Initially, we have performed the QSAR study using a pool of topological, structural, thermodynamic, spatial and electronic descriptors. In this QSAR study, we have used the *in vitro* activity data against both chloroquine-sensitive (D6 clone) and chloroquine-resistant (W2 clone) strains of *P. falciparum.* Though both GFA and G/PLS techniques were tried to develop the QSAR models, the former generated better quality equations which are reported here. We have subsequently tried to develop a QAAR model taking the *in vitro* activity data against chloroquine-resistant strain (W2 clone) as the response and those against chloroquine-sensitive strain (D6 clone) as one of the predictor variables.

3.1 QSAR with the in vitro *activity data of chloroquine-sensitive (D6 clone)* P. falciparum

Using the GFA spline technique, the following equation was obtained with acceptable cross-validated predictive variance (Q^2) and externally predicted variance (R^2_{pred}):

$$\begin{aligned}
\text{pIC}_{50} &= 2.9092(\pm 1.417) \\
&\quad + 1.25(\pm 0.1626)\,\text{RadOfGyration} \\
&\quad - 0.78(\pm 0.1063) < \text{S_ssCH}_2 - 4.98041 \\
&\quad > -4.32(\pm 1.025)\,\text{Jurs} - \text{FPSA_2} \\
&\quad - 0.170(\pm 0.06550)\,\text{Jurs} - \text{RNCS} \\
n_{\text{Training}} &= 26,\ R^2 = 0.826,\ R_{\text{a}} = 0.793, \\
F &= 24.90(\text{df}4, 21),\ s = 0.320, \\
\text{PRESS} &= 2.952,\ Q^2 = 0.761, \\
r^2_{\text{m(LOO)}} &= 0.590,\ n_{\text{Test}} = 9,\ R^2_{\text{pred}} = 0.829, \\
r^2_{\text{m(test)}} &= 0.731,\ r^2_{\text{m(overall)}} = 0.608. \qquad (1)
\end{aligned}$$

The model above could explain 79.3% of the variance (adjusted coefficient of variation). The cross-validated predictive variance (Q^2) was found to be 76.1%. The predictive ability of the model was evaluated by means of predictive R^2 (R^2_{pred}) for the test-set compounds, and the resulting R^2_{pred} value of 0.829 shows the good predictive ability of the model. Using the standardised variable matrix for regression, the significance level of the descriptors was found to be in the following order: RadOfGyration, S_ssCH$_2$, Jurs-fractional charged partial positive surface area (FPSA-2) and Jurs-relative negative-charge surface area (RNCS). The observed and calculated pIC$_{50}$ (D6 clone) values according to Equation (1) are given in Table 1.

RadOfGyration (Å) is a measure of the size of an object, a surface or an ensemble of points. It is calculated as the root-mean-square distance of the objects/parts from either its centre of gravity or an axis. This can be calculated by the following equation:

$$\text{RadOfGyration} = \sqrt{\left(\sum \frac{(x_i^2 + y_i^2 + z_i^2)}{N}\right)}.$$

Here, N is the number of atoms and x, y, z are the atomic coordinates relative to the centre of mass. This reflects that the shape and size of the molecules play an important role in the antimalarial activity. RadOfGyration has the favourable contribution towards the antimalarial activity as evidenced by the positive regression coefficient. If the value of RadOfGyration is higher, then the antimalarial activity will be higher (e.g. compounds **7**, **26** and **33**). Compounds **14**, **18** and **29** have lower values of RadOfGyration and their corresponding antimalarial activity values are in the lower range. However, in the case of compounds **9**, **27** and **28**, the antimalarial activity values are in the lower range in spite of the higher value of RadOfGyration due to a longer methylene spacer (*vide infra*). Compounds **7**, **26** and **33** have 6, 6 and 5 methylene spacer groups, respectively, while compounds **9**, **27** and **28** have 8, 7 and 8 methylene spacer groups, respectively, separating the 'active site' zinc binding hydroxamate moiety and the aryltriazolyl group. Therefore, the chain length of the methylene spacer groups should be fixed to an optimum of 5 or 6. For this reason, compounds **9**, **27** and **28** are showing lower activity in spite of higher values of RadOfGyration.

The S_ssCH$_2$ descriptor indicates the sum of *E*-state values of methylene carbons (—CH$_2$—) present in the aryltriazolylhydroxamate derivatives and the value depends on the number of such fragments present. The negative regression coefficient of the spline term $\langle \text{S\_ssCH}_2 - 4.98041 \rangle$ indicates that the numerical value of the *E*-state parameter S_ssCH$_2$ should be lower than 4.98041 for better antimalarial activity. It has been observed that compounds **7**, **26** and **33** showed better antimalarial activity since corresponding S_ssCH$_2$ values are lower than 4.98041. Compounds **5**, **9** and **28** showed lower activity as the corresponding S_ssCH$_2$ values are higher than 4.98041. The value of *E*-state parameter S_ssCH$_2$ will be higher if the number of methylene fragments is greater. In the case of compounds **5**, **9** and **28**, the number of methylene spacer groups is 8 and in the case of compounds **7**, **26** and **33**, the number of methylene spacer groups is 6, 6 and 5, respectively. Therefore, for better antimalarial activity, the number of methylene fragments should be 5 or 6.

Jurs-FPSA_2 is the fractional charged partial positive surface area. It can be calculated as the total charge-weighted positive-charge surface area (Jurs-PPSA_2)

divided by the total molecular solvent-accessible surface area (Jurs-SASA):

$$\text{Jurs-FPSA\_2} = \frac{\text{Jurs-PPSA\_2}}{\text{Jurs-SASA}}.$$

PPSA_2 is the partial positive SASA multiplied by the total positive charge Q^+, i.e.

$$\text{PPSA\_2} = Q^{+} * \sum\nolimits_{a+} \text{SA}_a^{+}.$$

Jurs-FPSA_2 has an unfavourable contribution towards the antimalarial activity as evidenced from the negative regression coefficient. This implies that a decrease in the total positive charge may enhance the activity. It has been observed that the compounds **2** and **24** have lower values of Jurs-FPSA_2 and their corresponding inhibitory activity values are in the highest range. In the case of compounds **9** and **23**, the values of Jurs-FPSA_2 are higher and the corresponding activity values are in the lowest range. Thus, for better antimalarial activity, the molecules should have a less fractional charged partial positive-charge surface area.

Jurs-RNCS is the relative negative-charge surface area: it is the solvent-accessible surface area of the most negatively charged atom divided by the relative negative charge (RNCG), i.e.

$$\text{RNCS} = \frac{\text{SA}^{-}_{\max}}{\text{RNCG}}.$$

Jurs-RNCS has the unfavourable contribution towards the antimalarial activity as evidenced by the negative regression coefficient. This implies that the decrease in RNCS may enhance the activity (e.g. compounds **6**, **7** and **33**). The high numerical values of Jurs-RNCS explain the reduced activity of compounds **14** and **18**.

3.2 QSAR with the in vitro *activity data of chloroquine-sensitive (W2 clone)* P. falciparum

Using the GFA linear technique, the following equation was obtained with acceptable cross-validated predictive variance (Q^2) and externally predicted variance (R^2_{pred}):

$$\begin{aligned}
\text{pIC}_{50} &= 2.407(\pm 1.143) \\
&\quad - 0.009949(\pm 0.002157)\,\text{Jurs-WPSA_2} \\
&\quad + 0.035515(\pm 0.005975)\,\text{Jurs-WNSA_1} \\
&\quad + 1.3906(\pm 0.2377)\,\text{RadOfGyration} \\
&\quad - 0.48698(\pm 0.07055)\,\text{Jurs-RNCS} \\
&\quad - 0.6765(\pm 0.211)\phi \\
n_{\text{Training}} &= 26,\ \ R^2 = 0.822,\ \ R^2_a = 0.777, \\
F &= 18.42(\text{df}5, 20),\ \ s = 0.350, \\
\text{PRESS} &= 4.00539,\ \ Q^2 = 0.708, \\
r^2_{\text{m(LOO)}} &= 0.536,\ \ n_{\text{Test}} = 9,\ \ R^2_{\text{pred}} = 0.748, \\
r^2_{\text{m(test)}} &= 0.540,\ \ r^2_{\text{m(overall)}} = 0.541.
\end{aligned} \tag{2}$$

The above model could explain 77.7% of the variance (adjusted coefficient of variation). The cross-validated predictive variance (Q^2) was found to be 70.8%. The predictive ability of the model was evaluated by means of predictive R^2 (R^2_{pred}) for the test-set compounds and the resulting R^2_{pred} value of 0.748 shows the good predictive ability of the model. Using the standardised variable matrix for regression, the significance level of the descriptors was found to be in the following order: Jurs-surface-weighted charged partial positive surface areas (WPSA_2), Jurs-surface-weighted charged partial negative surface areas (WNSA_1), RadOfGyration, Jurs-RNCS and ϕ. In the above model, some variables are highly intercorrelated (Jurs-WPSA_2, RadOfGyration and ϕ). Therefore, we have performed PLS using descriptors selected by the GFA technique to eliminate the intercorrelation problem. In the case of the PLS technique, the following equation was obtained with acceptable cross-validated predictive variance (Q^2) and externally predicted variance (R^2_{pred}). The observed and calculated pIC_{50} (W2 clone) values according to Equation (2a) are given in Table 1:

$$\begin{aligned}
\text{pIC}_{50} &= 3.85951 - 0.00475\,\text{Jurs-WPSA_2} \\
&\quad + 0.02118\,\text{Jurs-WNSA_1} \\
&\quad + 1.34327\,\text{RadOfGyration} \\
&\quad - 0.34038\,\text{Jurs-RNCS} - 1.10993\phi \\
n_{\text{Training}} &= 26,\ \ R^2 = 0.752,\ \ R^2_a = 0.705, \\
F &= 15.91(\text{df}4, 21),\ \ s = 0.402, \\
\text{PRESS} &= 4.9407,\ \ Q^2 = 0.639, \\
r^2_{\text{m(LOO)}} &= 0.602,\ \ n_{\text{Test}} = 9,\ \ R^2_{\text{pred}} = 0.738, \\
r^2_{\text{m(test)}} &= 0.507,\ \ r^2_{\text{m(overall)}} = 0.583.
\end{aligned} \tag{2a}$$

Jurs-WPSA_2 is the surface-weighted charged partial positive surface areas. This is calculated as the total charge weighted positive surface area (PPSA_2) multiplied by the total molecular solvent-accessible surface area (SASA) and divided by 1000, i.e.

$$\text{WPSA\_2} = \frac{\text{PPSA\_2} \times \text{SASA}}{1000}.$$

The negative regression coefficient indicates that Jurs-WPSA_2 has the unfavourable contribution towards the antimalarial activity. This implies that decreases in the total charge weighted positive surface area (e.g. compounds **3**, **25**, **26** and **24**) may enhance the antimalarial activity. It has been observed that compounds **9** and **28** show poor activity due to higher values of Jurs-PPSA_2 and Jurs-SASA.

Jurs-WNSA_1 is the surface-weighted charged partial negative surface areas. It can be calculated as the partial negative surface area (PNSA_1) multiplied by the total molecular solvent-accessible surface area (SASA) and

divided by 1000, i.e.

$$\text{Jurs-WNSA\_1} = \frac{\text{Jurs-PNSA\_1} \times \text{Jurs-SASA}}{1000}.$$

PNSA_1 is the sum of the SASAs of all negatively charged atoms, i.e.

$$\text{PNSA\_1} = \sum\nolimits_{a^-} \text{SA}_a^-,$$

where the sum is restricted to negatively charged atoms a^-.

The positive regression coefficient of Jurs-WNSA_1 indicates that it has a favourable contribution towards the antimalarial activity. This implies that increases in the PNSA_1 may enhance the antimalarial activity. It has been observed that compounds **25**, **26**, **33** and **34** show good antimalarial activity due to higher numerical values of Jurs-PNSA_1. In the case of compounds **9** and **23**, the low numerical values explain the reduced antimalarial activity.

Similar to Equation (1), Equation (2a) explains higher activity values of compounds **7** and **33** and lower activity values of compounds **18** and **29** due to, respectively, higher and lower values of RadOfGyration.

Similar to Equation (1), Equation (2a) shows that a decrease (e.g. compounds **6** and **7**) and increase (e.g. compounds **4** and **18**) in the RNCS may, respectively, enhance and reduce the antimalarial activity of the corresponding compounds.

The parameter ϕ is the Kier molecular flexibility index: a measure of molecular flexibility derived from the Kier alpha-modified shape descriptors ${}^1\kappa_\alpha$ and ${}^2\kappa_\alpha$:

$$\phi = \frac{{}^1\kappa_\alpha \times {}^2\kappa_\alpha}{A},$$

where the Kier shape indices calculated from the H-depleted molecular graph depend on the heteroatoms by the parameter α [39]. The parameter ${}^1\kappa_\alpha$ encodes information about the count of atoms and relative cyclicity of molecules, while ${}^2\kappa_\alpha$ encodes information about branching or relative spatial density of molecules. The atom count A allows comparisons among isomers. The negative regression coefficient of ϕ indicates that decreases in the molecular flexibility may enhance the antimalarial activity (e.g. **3**, **24**, **36** and **37**) and increases in the molecular flexibility may reduce the antimalarial activity (e.g. **5**, **9**, **27** and **28**).

3.3 *Quantitative activity–activity relationship*

We have also performed the QAAR study for the development of models by taking one response (pIC_{50}_W2) as the dependent variable and another as one of the independent (pIC_{50}_D6) variables. Models were generated by using genetic function approximation (GFA) with spline option as the statistical tool. The mutation probability was kept at 50% with 5000 iterations. In the case of the GFA spline technique, the following equation was obtained with acceptable cross-validated predictive variance (Q^2) and externally predicted variance (R^2_{pred}). From this model, one activity can be calculated when the other is known:

$$\begin{aligned}
pIC_{50}\text{_W2} = &-0.24585(\pm 0.08697) \\
&+ 1.06282(\pm 0.02565)\, pIC_{50}\text{_D6} \\
&+ 0.7405(\pm 0.1768)\langle 0.705712 - \text{Sr}\rangle \\
&+ 14.493(\pm 3.519)\langle 0.301701 - {}^3\chi_c^v\rangle \\
n_{\text{Training}} = 26,\ & R^2 = 0.987,\ R_a^2 = 0.986, \\
F = 576.14(\text{df}3, 22),\ & s = 0.088, \\
\text{PRESS} = 0.213,\ & Q^2 = 0.984, \\
r^2_{\text{m(LOO)}} = 0.959,\ & n_{\text{Test}} = 9,\ R^2_{\text{pred}} = 0.942, \\
r^2_{\text{m(test)}} = 0.882,\ & r^2_{\text{m(overall)}} = 0.949.
\end{aligned} \tag{3}$$

The above GFA model could explain 98.6% of the variance (adjusted coefficient of variation). The cross-validated predictive variance (Q^2) was found to be 98.4%. The predictive ability of the model was evaluated by means of predictive R^2 (R^2_{pred}) for the test-set compounds and the resulting R^2_{pred} value of 0.942 shows the good predictive ability of the model. Using the standardised variable matrix for regression, the significance level of the descriptors was found to be in the following order: pIC_{50}_D6, Sr and ${}^3\chi_c^v$. The observed and calculated pIC_{50} (W2 clone) values according to Equation (3) are given in Table 1.

The antimalarial activity against the chloroquine-resistant strain W2 clone is favoured by the parameters such as Sr (superdelocalisability) and ${}^3\chi_c^v$ (molecular connectivity index). The parameter ${}^3\chi_c^v$ is the third-order valance-modified cluster connectivity index, indicating the impact of branching. The positive regression coefficients of the spline term $\langle 0.705712\text{-Sr}\rangle$ and $\langle 0.301701 - {}^3\chi_c^v\rangle$ indicate that the numerical values of Sr and ${}^3\chi_c^v$ should be less than 0.705712 and 0.301701, respectively, for the antimalarial activity against W2 clone. It has been observed that compounds **26**, **36** and **37** showed better antimalarial activity against W2 clone, since their corresponding Sr values are lower than 0.705712. Compounds **9**, **12** and **14** show poor antimalarial activity due to the higher value of Sr (greater than 0.705712). A value of ${}^3\chi_c^v$ less than 0.301701 accounts for better antimalarial activity of compound **15**.

4. Overview and conclusion

In this paper, we have explored QSAR studies of 35 aryltriazolylhydroxamates as antimalarial agents using

electronic (dipole-mag and Sr), spatial (RadOfGyration, Jurs descriptors, Area, PMI-mag, Density, Vm), thermodynamic (ALogP, ALogP98, MolRef, MR, LogP), structural (H-bond donor, H-bond acceptor, Rotlbonds) along with topological descriptors. QSAR models were developed by taking activity data against both chloroquine-sensitive (D6 clone) and chloroquine-resistant (W2 clone) strains of *P. falciparum*. Finally, the QAAR model was developed by taking the *in vitro* activity data against the chloroquine-resistant strain (W2 clone) as the response and those against the chloroquine-sensitive strain (D6 clone) as one of the predictor variables. The whole data-set ($n = 35$) was divided into a training set ($n = 26$) and a test set ($n = 9$) by using *k*-means clustering. Statistical qualities of different models are given in Table 4. From the different QSAR models, it can be concluded that (i) the number of methylene spacer groups should be optimum (5 or 6). The number of methylene spacer groups greater than 6 and less than 5 will lead to poor antimalarial activity. (ii) The total charge-weighted positive surface area (PPSA-2) of the molecules should be less. (iii) The partial negative surface area (PNSA-1) of the molecules should be greater. This suggests that the negative charge distributed over a large surface area may enhance the antimalarial activity. (iv) Flexibility of the molecules should be less. This suggests that the large number of methylene spacer groups and single bonds will lead to poor antimalarial activity. (v) The QAAR studies suggest that electronic parameter Sr (superdelocalisability) and molecular connectivity index $^3\chi_c^v$ are important for better antimalarial activity against the chloroquine-resistant strain.

Acknowledgements

Financial assistance from the UGC (New Delhi) in the form of a fellowship to P.K.O. is thankfully acknowledged.

References

[1] World Malaria Report (2008) WHO, Geneva, Switzerland. Available at http://www.who.int/malaria/publications/atoz/9789241563697/en/index.html.
[2] M.C. Murray and M.E. Perkins, *Chemotherapy of malaria*, Annu. Rep. Med. Chem. 31 (1996), pp. 141–150.
[3] J.A. Vroman, M.A. Gaston, and M.A. Avery, *Current progress in the chemistry, medicinal chemistry and drug design of artemisinin based antimalarials*, Curr. Pharm. Des. 5 (1999), pp. 101–138.
[4] A. Gregson and C.V. Plowe, *Mechanisms of resistance of malaria parasites to antifolates*, Pharmacol. Rev. 57 (2005), pp. 117–145.
[5] P.E. Duffy and C.H. Sibley, *Are we losing artemisinin combination therapy already?* Lancet 366 (2005), pp. 1908–1909.
[6] R.G. Ridley, *Medical need, scientific opportunity and the drive for antimalarial drugs*, Nature (London) 41 (2002), pp. 5686–5693.
[7] M.P. Gonzalez, C. Teran, L. Saiz-Urra, and M. Teijeira, *Variable selection methods in QSAR: An overview*, Curr. Top. Med. Chem. 8 (2008), pp. 1606–1627.
[8] L. Adane and P.V. Bharatam, *3D-QSAR analysis of cycloguanil derivatives as inhibitors of A16V+S108T mutant* Plasmodium falciparum *dihydrofolate reductase enzyme*, J. Mol. Graph. Model 28 (2009), pp. 357–367.
[9] M. Cruz-Monteagudo, F. Borges, M.P. Gonzalez, and M.N.D.S. Cordeiro, *Computational modeling tools for the design of potent antimalarial bisbenzamidines: Overcoming the antimalarial potential of pentamidine*, Bioorg. Med. Chem. 15 (2007), pp. 5322–5339.
[10] A.R. Katritzky, O.V. Kulshyn, I. Stoyanova-Slavova, D.A. Dobchev, M. Kuanar, D.C. Fara, and M. Karelson, *Antimalarial activity: A QSAR modeling using CODESSA PRO software*, Bioorg. Med. Chem. 14 (2006), pp. 2333–2357.
[11] M.D. Parenti, S. Pacchioni, A.M. Ferrari, and G. Rastelli, *3D quantitative structure–activity relationship analysis of a set of* Plasmodium falciparum *dihydrofolate reductase inhibitors using a pharmacophore generation approach*, J. Med. Chem. 47 (2004), pp. 4258–4267.
[12] V. Patil, W. Guerrant, P.C. Chen, B. Gryder, D.B. Benicewicz, S.I. Khan, B.L. Tekwani, and A.K. Oyelere, *Antimalarial and antileishmanial activities of histone deacetylase inhibitors with triazole-linked cap group*, Bioorg. Med. Chem. 18 (2010), pp. 415–425.
[13] M.T. Makler, J.M. Ries, J.A. Williams, J.E. Bancroft, R.C. Piper, B.L. Gibbins, and D.J. Hinriches, *Parasite lactate dehydrogenase as an assay for* Plasmodium falciparum *drug sensitivity*, Am. J. Trop. Med. Hyg. 48 (1993), pp. 739–741.
[14] *Cerius2 Version 4.10, a product of Accelrys, Inc.*, San Diego, USA; software available at http://www.accelrys.com/cerius2.
[15] L. Eriksson, J. Jaworska, A.P. Worth, M.T.D. Cronin, M.M. Robert, and P. Gramatica, *Methods for reliability and uncertainty assessment and for applicability evaluations of classification and regression-based QSARs*, Environ. Health Perspect. 111 (2003), pp. 1361–1375.
[16] R. Guha and P.C. Jurs, *Determining the validity of a QSAR model – A classification approach*, J. Chem. Inf. Model 45 (2005), pp. 65–73.
[17] J.T. Leonard and K. Roy, *On selection of training and test sets for the development of predictive QSAR models*, QSAR Comb. Sci. 25 (2006), pp. 235–251.
[18] K. Roy, *On some aspects of validation of predictive QSAR models*, Expert Opin. Drug Discov. 2 (2007), pp. 1567–1577.
[19] B.S. Everitt, S. Landau, and M. Leese, *Cluster Analysis*, Edward Arnold, London, 2001.
[20] E.R. Dougherty, J. Barrera, M. Brun, S. Kim, R.M. Cesar, Y. Chen, M. Bittner and J.M. Trent, *Inference from clustering with application to gene-expression microarrays*, J. Comput. Biol. 9 (2002), pp. 105–126.
[21] R.A. Johnson and D.W. Wichern, *Applied Multivariate Statistical Analysis*, Pearsons, Delhi, 2005, pp. 668–730.
[22] *SPSS is statistical software of SPSS Inc.*, USA.
[23] D. Rogers and A.J. Hopfinger, *Application of genetic function approximation to quantitative structure–activity relationships and quantitative structure–property relationships*, J. Chem. Inf. Comput. Sci. 34 (1994), pp. 854–866.
[24] H. Wold, *Research Papers in Statistics*, Wiley, New York, 1966.
[25] H. Jores-Kong and H. Wold, *Systems under Indirect Observation: Causality, Structure, Prediction*, North-Holland, Amsterdam, 1982.
[26] S. Wold and L. Eriksson, *Validation tools*, in *Chemometric Methods in Molecular Design*, H. van de Waterbeemd, ed., VCH, Weinheim, 1995, pp. 312–317.
[27] Y. Fan, L.M. Shi, K.W. Kohn, Y. Pommier, and J.N. Weinstein, *Quantitative structure–antitumor activity relationships of camptothecin analogues: Cluster analysis and genetic algorithm-based studies*, J. Med. Chem. 44 (2001), pp. 3254–3263.
[28] P.P. Roy and K. Roy, *On some aspects of variable selection for partial least squares regression models*, QSAR Comb. Sci. 27 (2008), pp. 302–313.
[29] S.S. Kulkarni and V.M. Kulkarni, *3D quantitative structure–activity relationship of interleukin 1-beta converting enzyme inhibitors: A comparative molecular field analysis study*, J. Med. Chem. 42 (1999), pp. 373–380.
[30] P.P. Roy, J.T. Leonard and K. Roy, *Exploring the impact of the size of training sets for the development of predictive QSAR models*, Chemom. Intel. Lab. Sys. 90 (2008), pp. 31–42.

[31] H. Kubinyi, F.A. Hamprecht and T. Mietzner, *3D quantitative similarity–activity relationships (3D QSiAR) from SEAL similarity matrices*, J. Med. Chem. 41 (1998), pp. 2553–2564.
[32] G.R. Marshall, In *3D QSAR in drug design – Theory, methods and applications*, H. Kubinyi, ed., ESCOM, Leiden, 1994, pp. 117–133.
[33] P.P. Roy, S. Paul, I. Mitra and K. Roy, *On two novel parameters for validation of predictive QSAR models*, Molecules 14 (2009), pp. 1660–1701.
[34] I. Mitra, P.P. Roy, S. Kar, P.K. Ojha, and K. Roy, *On further application of r^2_m as a metric for validation of QSAR models*, J. Chemom. 24 (2010), pp. 22–33.
[35] K. Roy and S. Paul, *Exploring 2D and 3D QSARs of 2,4-diphenyl-1,3-oxazolines for ovicidal activity against tetranychus urticae*, QSAR Comb. Sci. 28 (2009), pp. 406–425.
[36] K. Roy and S. Paul, *Docking and 3D QSAR studies of protoporphyrinogen oxidase inhibitor 3H-pyrazolo [3,4-d] [1,2,3] triazin-4-one derivatives*, J. Mol. Model 16 (2010), pp. 137–153.
[37] *MINITAB is a statistical software of Minitab Inc.*, USA.
[38] *STATISTICA is a statistical software of STATSOFT Inc.*, USA.
[39] L.B. Kier, *An index of molecular flexibility from kappa shape attributes*, Quant. Struct. Act. Relat. 8 (1989), pp. 221–224.