Journal of Fundamental Sciences Vol. 6, No. 1 (2010) 76-83.

ISSN 1823-626X

Journal of Fundamental Sciences

available online at http://jfs.ibnusina.utm.my

Quantitative structure-activity relationship for antimalarial activity of artemisinin

Rosmahaida Jamaludin and Mohamed Noor Hasan

Department of Chemistry, Faculty of Science, Universiti Teknologi Malaysia, 81310 UTM Skudai, Johor.

Received 20 April 2010, Revised 4 June 2010, Accepted 13 June 2010, Available online 25 June 2010

ABSTRACT

The increase in resistance to older drugs and the emergence of new types of infection have created an urgent need for discovery and development of new compounds with antimalarial activity. Quantitative-Structure Activity Relationship (QSAR) methodology has been performed to develop models that correlate antimalarial activity of artemisinin analogs and their molecular structures. In this study, the data set consisted of 197 compounds with their activities expressed as log RA (relative activity). These compounds were randomly divided into training set (n=157) and test set (n=40). The initial stage of the study was the generation of a series of descriptors from three-dimensional representations of the compounds in the data set. Several types of descriptors which include topological, connectivity indices, geometrical, physical properties and charge descriptors have been generated. The number of descriptors was then reduced to a set of relevant descriptors by performing a systematic variable selection procedure which includes zero test, pairwise correlation analysis and genetic algorithm (GA). Several models were developed using different combinations of modelling techniques such as multiple linear regression (MLR) and partial least square (PLS) regression. Statistical significance of the final model was characterized by correlation coefficient, r^2 and root-mean-square error calibration, *RMSEC*. The results obtained were comparable to those from previous study on the same data set with r^2 values greater than 0.8. Both internal and external validations were carried out to verify that the models have good stability, robustness and predictive ability. The cross-validated regression coefficient (r^2_{cv}) and prediction regression coefficient (r^2_{test}) for the external test set were consistently greater than 0.7. The QSAR models developed in this study should facilitate the search for new compounds with antimalarial activity.

| QSAR| Antimalarial | Artemisinin| GA-PLS | MLR |

http://dx.doi.org/10.11113/mjfas.v6n1.181

1. INTRODUCTION

Malaria is an infectious disease spread by the bite of female *Anopheles* mosquito. Generally, malaria on human is caused by four *Plasmodium* species which are *vivax*, *malariae*, *ovale* as well as the most prevalent and life threatening parasite, *falciparum* [1]. This deadly disease is a major health problem especially in the developing world, killing approximately two million people each year. Malaria medications such as chloroquine, mefloquine and fansidar have become ineffective against the disease [2]. Thus, the increases in resistance to older drugs have created an urgent and continuous need for discovery and development of new antimalarial agents to treat sensitive and drug-resistant strains of malaria. Artemisinin (depicted in *Figure 1*) also known as *qinghaosu* , isolated from Chinese medicinal herb *Artemisia annua L.* not only have significant antimalarial activity but also kill the parasites more rapidly and are eliminated quickly. Artemisinin and its derivatives such as artemether, arteether, artesunate and dihydroartemisinin are toxic to malarial parasite at nanomolar concentrations [3].

Corresponding author at: Department of Chemistry, Faculty of Science
Universiti Teknologi Malaysia, 81310 UTM Skudai, Johor
E-mail addresses: rosma@ic.utm.my (Rosmahaida Jamaludin)

This antimalarial compound is a sesquiterpene endoperoxide lactone with unique structure and mechanism of action. The presence of a peroxide bridge, a known source of oxygen-free radicals is essential for its antimalarial activity. Basically, the mode of action involves two steps. First, catalytic activation of artemisinin where the endoperoxide bridge break open to generate free radical intermediates in a reaction catalyzed by iron or heme in the malaria parasites [4]. Posner showed that the resulting free radicals are carbon-centred [5] and also suggested that ferryl ions (Fe[IV]=O) appear to form in this iron-mediated decomposition of artemisinin [6]. The next step is alkylation which involves formation of covalent adducts between the drug and malarial protein as well as heme (covalent artemisinin-heme adducts) in high yield under very mild conditions that appear to damage specific intracellular targets and could mediate the killing action of artemisinin derivatives [7-9]. Despite clinical success of artemisinin, considerable efforts have been made to develop more potent, selective, nontoxic, clinically useful newer semi-synthetic and synthetic derivatives that have similar mechanism of action yet are superior in activity in order to eradicate or control this infection throughout the world.

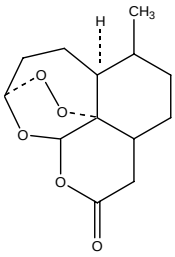

Figure 1: Structure of artemisinin

Several Quantitative-Structure Activity Relationship (QSAR) studies have been reported on the data set of artemisinin analogues for antimalarial activity. Avery [10] utilized comparative molecular field analysis (CoMFA) which was dependent on molecular conformation and structural alignment along with hologram QSAR (HQSAR). All of the models emerging from both methods were of practical quality and exhibited good predictive ability. However, models that considered racemic compounds in the training set have lower r^2 (correlation coefficient for training set) and r^2_{test} (correlation coefficient for test set) values compared to the models that exclude them. Alternatively, Guha [11] developed 2D-QSAR model which was not dependent on molecular alignments. He has developed both linear and nonlinear models to link the structures to their reported biological activity. The best linear model provided an interpretation of the Structure-Activity Relationship (SAR) trend present in the data set, while the neural network model provided superior predictive ability. In a recent work [12], Srivastava had built robust QSAR models with high predictive ability using a combination of topological, electro-topological state indices, electronic and thermodynamic descriptors of chemical structures. He had used Genetic Algorithm (GA) and Multiple Linear Regression (MLR) as tools to model the activity of artemisinin analogues.

QSAR is a modelling technique in which the observed activities or properties of chemical compounds are correlated with structural descriptors derived from the molecular structures that can be represented in mathematical equation as shown below:

$$\text{Molecular activity} = f(\text{descriptor}) = a_1d_1 + a_2d_2 + a_3d_3 + \ldots + a_nd_n \quad (1)$$

where $d_1, d_2, d_3, \ldots d_n$ are structural descriptors and $a_1, a_2, a_3 \ldots a_n$ are regression coefficients.

The models developed can be utilized to study the correlation between structural features of artemisinin with their biological activities, predict activities of compounds not included in the model development process as well as to form a basis for understanding factors affecting their activities [13, 14].

The aim of this work is to employ well-characterized data set of artemisinin analogues with *in vitro* antimalarial activity collected by Avery [10] using practical method of variable selection to develop reliable predictive QSAR models. Combination of different types of molecular descriptors ranging from 0D to 3D descriptors were employed to describe compound structural diversity and correlate them quantitatively to antimalarial activity. For this study, the best linear models provided good statistical properties and predictive ability as well as sound physical interpretation of the structure-activity trend captured by the model. Subsequently, these models are expected to perform well as rapid screening tools to uncover new and useful anti-malarial drugs of artemisinin analogues from a large library of compounds.

2. EXPERIMENTAL

Typically, the first step in a QSAR study is structure entry and molecular modeling together with generation of the 3D models of each compound in the data set. The next procedure involves descriptor generation followed by feature selection. The subsequent steps are the construction of the QSAR models using the descriptors set and finally, validation of the model by predicting the property of compounds in the external prediction set.

All structural and biological data of 211 artemisinin analogues used in this study were based on the previous research taken from the literature [10]. The molecules were either peroxides or trioxanes and were categorized into different classes as presented in *Table 1*.

Each of these compounds had associated *in vitro* bioactivity values against the drug-resistant malaria strain *P. falciparum* (W-2 clone). The dependent variable was log RA (relative activity) which fell into the range [-4, 1.47] and was defined as:

$$logRA = log\left(\frac{IC_{50}\ of\ Artemisinin}{IC_{50}\ of\ analog}\right) \times log\left(\frac{MW\ of\ analog}{MW\ of\ Artemisinin}\right) \quad (2)$$

Table 1: Different classes of artemisinin analogues used in this study

Class of compounds	Number of compounds
Artemisinin analogs	40
Deoxyartemisinin analogs	15
10-Substituted artemisinin analogs	49
Secoartemisinin analogs	7
Bicycloartemisinin analogs	5
Azaartemisinin analogs	19
Artemisinin derivatives lacking the D-ring	33
Dihydroartemisinin derivatives	7
Various derivatives of artemisinin and arteether	13
Miscellaneous artemisinin analog	12

The RA was calculated from the experimentally derived control IC_{50} (reported in ng/ml) of artemisinin divided by the IC_{50} of the analogue and corrected for molecular weight (*MW*). Essentially, the selected compounds had been tested using the same assay method and had a reported control activity of artemisinin. This was

crucial for a bioassay method that employed parasitized red blood cells because there could be interday and interlaboratory variations in the IC_{50} [10]. Since the data set contained 14 enantiomeric pairs and assuming that only one enantiomer was bioactive, the member of each pair with the lowest log RA value was removed [11].

The division of data set into training set for model development and test set for model validation was performed by random method. Around 25% of the objects were placed in the test set where 40 molecules were removed from the original 197 compounds that evenly spanned the antimalarial activity range, as well as the structure diversity of the database. In other words, they were chosen in such a way that they were as representative as possible of the global data set to determine the external predictive ability of the resulting model.

All the software packages used in this study were run on Microsoft Window XP on a Pentium IV system. ChemDraw Ultra version 6.0 (Cambridge Soft) was used to draw 2D model molecular structure of the compounds. Next, Chem3D Ultra version 6.0 was utilized to convert the molecular structure to 3D structure and afterward, the structures of the compounds were energy minimized using MOE version 2009.10 (Chemical Computing Group Inc.) software.

The next step in the process was to characterize each molecule in the data set with an appropriate set of computer generated molecular descriptors that are derived from the 3D models. The molecular descriptors for all the compounds were solely calculated using DRAGON software package version 5.4 [15] that encoded topological, geometric, structural, and physical properties of the molecules [16]. All descriptors were auto-scaled to zero mean and unit standard deviation where same variance was given to the informative and uninformative variables.

After numerical descriptors had been calculated for each compound, the number of descriptors was reduced to a set of descriptors that were information-rich but as small as possible. Prior to analysis, objective feature selection was performed where highly correlated and redundant descriptors were removed from the pool. This included constant and near-constant variables. In addition, one from each correlated descriptor pair having pair-wise correlation coefficient greater than 0.95 was randomly removed. Then, an identical test was carried out manually in which a descriptor was rejected if the values of the descriptors for more than 90% of the molecules were identical. This resulted in a reduced pool of 488 descriptors for further analysis.

Next, subjective feature selection was performed to identify a descriptor subset that best map an accurate link between structure and property of interest. Techniques for selecting the best subset of variables included GA and forward stepwise multiple regression. GA very often led to a significant improvement of the predictive ability when correctly applied and could also be used as assistance during interpretation in order to understand which variables were correlated with a specific activity of the compounds.

Feature selection and models generation were achieved using routines in PLS Toolbox 5.2 (Eigenvector Research Inc.) in Matlab 7.5.0 (2007) (The Mathworks Inc., Natick, MA). In PLS Toolbox 5.2, GA variable selection was performed using the Genetic Algorithm *GUI* (genalg) function instead of the command-line version (gaselectr) which used the minimum root-mean-square error cross-validation (*RMSECV*) as objective function. The form of the objective function favoured sets that had *RMSECV* that was as low as possible, while minimizing the number of parameters being used as descriptors.

The optimal values for the GA parameter were determined after several GA-PLS and GA-MLR runs with different settings of initial populations as well as based on the GA settings used by other researchers [17-22]. Several factors had to be considered when choosing the appropriate value for the GA parameter. High variables to objects ratio where the critical point was 5, would lead to senseless model where it would model noise instead of information. Therefore, in the study, window width was set to 1 because the number of variables and molecules were 488 and 157 respectively, thus the ratio was approximately 3 which was quite reasonable. Another point to consider when setting the parameters was to perform high number of different runs and try to extract some information from all of them to increase reliability. However, the runs must be stopped very early to avoid modeling noise. Since small part of the domain was explored and different runs yielded different final results, global information should be obtained in several runs. Thus, the parameters used for the GA included maximum generations of 100 with replicate runs of 30. Other settings that differed from the default settings were percent initial terms of 10 and number of latent variable (LV) was 5. The combination of variables producing the best response was taken as the final solution where variables were entered according to the frequency of selections. Therefore, instead of directly utilizing the best data set selected by GA, the final model was obtained following a stepwise approach where the frequency with which each variable was selected in the top chromosome of each run was used [17, 22].

Several models were built by employing combinations of variable selection and statistical methods. For quantitative modeling, QSAR models were developed using partial least square (PLS) and MLR methods. The hybrid approach that integrated GA and PLS or GA and MLR was applied to variable selection and modeling.

Statistical significance of the final model was characterized with correlation coefficient, r^2 and root mean square error of calibration, *RMSEC*. The high value of r^2 and the low value of *RMSEC* indicated a more stable model. Furthermore, QSAR models were presented as QSAR equations. The regression coefficients that were reported in brackets after the descriptor abbreviation indicated the significance of an individual descriptor presented in the regression model. The plot of predicted vs. experimental activity displayed the activity predicted by a QSAR model against the experimentally measured or observed activity.

Besides, the plot of studentized residual vs. predicted value was used to detect any outliers.

Finally, both internal and external validations were carried out in this study to verify that the models have good stability, robustness and predictivity. Cross-validation by leave-one-out provided rigorous internal check on the models while external validation involved predicting activity of compounds in the external test set. Therefore, statistical results would be reported in r^2_{CV} and r^2_{test} respectively.

Y-randomization test or Y scrambling consisted of rebuilding models using shuffled or randomized activities of the training sets followed by evaluation of predictive accuracy of the resultant models in comparison to the original model. Often it was used along with cross-validation and the calculation procedure was repeated in the same manner. The goals of this widely used method were to establish model robustness, to ensure that models did not merely capture noise and to assess if models are the result of chance correlations. If a true QSAR relationship existed with the real dependent variable, results for the scrambling runs should be very poor.

3. RESULTS & DISCUSSION

Generally, PLS can simply treat large data matrices, extract relevant part of the information, produce reliable but very complex models and almost insensitive to noise [14] whereas Genetic Algorithm (GA) is a feature selection technique used to select the most informative variables. However, combination of PLS and GA to find the best QSAR model is more beneficial because it improves the predictive ability of the model and at the same time enhance its simplicity.

In this work, the performance of GA was measured by comparing the *RMSECV* or r^2 of the model proposed by GA with the model containing all the variables. The results of the two final models are as summarized in *Table 2*.

Table 2: Statistics of the PLS model

Parameter	PLS (without GA)	PLS and GA
r^2	0.735	0.815
r^2_{CV}	0.622	0.780
r^2_{test}	0.690	0.766
RMSECV	0.939	0.706
Number of descriptors used	488	11

Apparently, integration of GA in building PLS model offered significant improvements over the model developed using PLS alone in terms of higher predictive power as well as substantially reducing the number of descriptors. It was a common practice in QSAR studies to obtain a model containing as few variables as possible to ensure there was no overfitting and hence led to a simple and interpretable model.

The best subset of descriptors to build the above GA-PLS model was based on the frequency of variables selected in GA as shown in *Figure 2* where 11 peaks representing descriptors had been identified as the best combination that yielded high r^2 in the training set and linear fit predictive r^2 for test sets with fewer variables compared to other models. The most frequently selected variables were concentrated in the region of atom-centered fragments particularly O-063 descriptor, corresponding to the R-O-O-R functional group.

In order to investigate the presence of outliers, a plot of studentized residual vs. predicted activity was employed. Once the outliers were detected, they were removed from the training set and regeneration of the linear model was carried out mainly to enhance the quality of the linear model for subsequent interpretation using PLS.

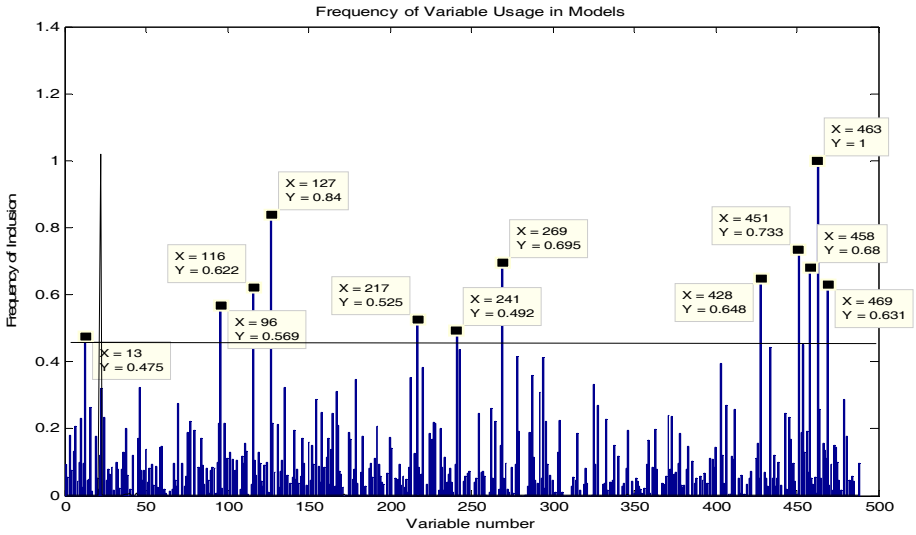

Figure 2: Frequency of variables selected in models by GA.

Based on the residual plot depicted in *Figure 3*, six compounds (029, 096, 107, 173, 175 and 188, refer to reference [10] for the structures) appeared as outliers and were eliminated. The quality of the above QSAR model had been further improved after the removal of these compounds. The plots for quality of the prediction models for the training compounds before and after the removal of outliers are as shown in *Figures 4* and *5*. Those compounds were found to be outliers most probably because of their very low activity.

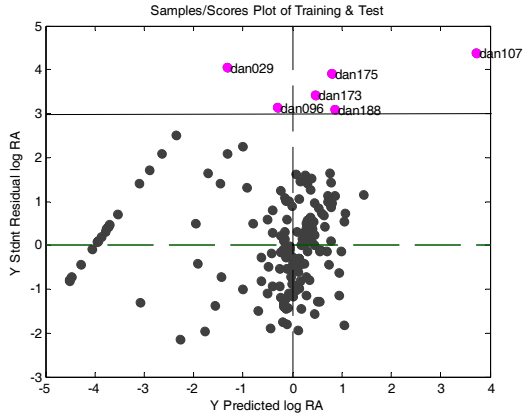

Figure 3: Plot of studentized residual vs. predicted value for PLS model.

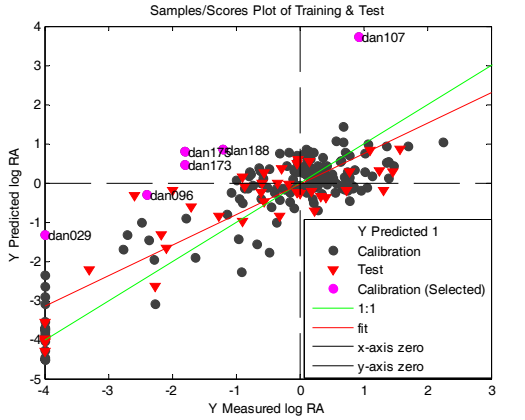

Figure 4: Plot of predicted vs. measured log RA before outlier removal

Figure 5: Plot of predicted vs. measured log RA after outlier removal

The best significant relationship between the molecular descriptors and antimalarial activity for GA-PLS model has been deduced to:

$$logRA = -0.967\,(GATS6v) + 1.11\,(EEig05r) + 0.799\,(BELm6) + 0.0376\,(Mor03u) + 0.757\,(Mor29u) + 1.30\,(Mor26m) - 24.2\,(R6e+) - 0.589\,(C\text{-}011) + 0.147\,(H\text{-}051) + 1.76\,(O\text{-}063) + 2.55\,(Q\text{-}mean) + 7.34 \quad (3)$$

(r^2 = 0.815, r^2_{CV} = 0.768, *RMSEC* = 0.649, *RMSECV* = 0.706, r^2_{test} = 0.766, *RMSEP* = 0.794)

The QSAR model developed in this study was statistically best fitted and consequently used for the prediction of antimalarial activity of test set molecules as reported in *Table 3*. Log RA predicted by GA-PLS model was consistent with the experimental data and in good agreement with each other. The high value of r^2_{CV} suggested a more stable and suitable model for predicting compounds not included in the training set as evident from the reasonable r^2_{test} value. Thus, PLS model gave excellent predictions with reliable statistical properties.

However, it should be noted that a PLS analysis only provided guideline regarding the interpretation of the descriptors in the model and it did not provide exact quantitative descriptions of descriptor contribution. Clearly, a major portion of the above PLS model was 3D-MoRSE (descriptors calculated by summing atom weights viewed by a different angular scattering function) and atom-centered fragments descriptors. The positive sign of descriptor coefficient such as mean absolute charge denoted as Q-mean showed that increasing the charge polarization caused the log RA activity to increase. Meanwhile, GATS6v and R6e$^+$ descriptors had negative effects on the activity.

Lastly, the activity values for the training set without the outliers were scrambled several times and linear models were reconstructed with the randomized dependent variables to ensure that the linear models were not due to chance correlation. Most QSAR models generated in the Y-randomization test exhibited relatively low values of the statistical parameters for both training and test sets with the r^2 values ranging from 0.223 to 0.351 while the r^2_{test} values ranging from 0.002 to 0.498. Hence, these results implied that chance correlation was negligible and the QSAR model obtained for the given data set was reliable.

Multiple Linear Regression Analysis (MLRA) technique had been used to build the best QSAR model using two different methods of variable selection which were GA and forward stepwise in order to select the most informative variables. The equation for the final model obtained by combining the GA and MLR is as shown below:

$$logRA = -0.445\ (nR05) + 5.83\ (BELm6) - 0.0171\ (RDF070u) - 0.398\ (Mor04v) + 1.02\ (H8m) - 0.556\ (C011) + 1.64\ (O\text{-}063) - 10.1 \quad (4)$$

(r^2 = 0.804, r^2_{CV} = 0.779, *RMSEC* = 0.629, *RMSECV* = 0.669, r^2_{test} = 0.740, *RMSEP* = 0.823)

As evident from the studentized residual vs. predicted plot, eight observations appeared to be distinct outliers (29, 196, 186, 096, 107, 082, 116 and 189). A plot of predicted vs. experimental log RA after outliers removal is as shown in *Figure 6*. The QSAR model developed was statistically reliable and predictive with high r^2 and r^2_{test} values as well as low error.

Table 3: Observed and predicted activities of artemisinin derivatives in the test set.

Compound	Relative activity			Compound	Relative activity		
	Observed	Predicted	Residual		Observed	Predicted	Residual
5	-0.17	-0.04	0.13	118	-2.17	-1.34	0.83
10	-0.28	0.35	0.63	120	-2.26	-2.64	-0.38
15	-0.55	0.24	0.79	125	-0.04	0.48	0.52
20	-2.00	-0.21	1.79	130	0.00	-0.27	-0.27
25	-0.35	-0.01	0.34	135	1.46	0.28	-1.18
30	-2.09	-1.67	0.42	140	-4.00	-4.01	-0.01
45	-4.00	-3.99	0.01	143	-0.32	-0.85	-0.53
48	-4.00	-4.06	-0.06	150	1.10	0.83	-0.28
52	-4.00	-4.29	-0.29	155	0.33	-0.35	-0.68
56	-4.00	-3.55	0.45	160	-0.59	-0.02	0.57
70	1.32	-0.20	-1.52	165	-0.04	-0.26	-0.22
75	0.78	0.29	-0.49	170	-0.90	0.15	1.05
80	-0.03	0.59	0.62	178	0.16	0.55	0.39
87	-0.55	-0.48	0.07	181	0.41	-0.39	-0.79
93	-1.70	-0.61	1.09	187	0.23	-0.71	-0.94
97	-0.89	-0.99	-0.11	191	-2.59	-0.32	2.27
100	0.74	-0.19	-0.92	195	-3.30	-2.22	1.08
110	-0.60	-0.26	0.34	200	1.25	0.29	-0.96
112	-1.27	-0.85	0.42	206	1.57	0.85	-0.72
115	-0.86	-0.09	0.77				

The best linear model consisted of seven descriptors is as represented in Equation (4). The first descriptor with the largest coefficient value was Burden eigenvalues descriptors denoted by BELm6.

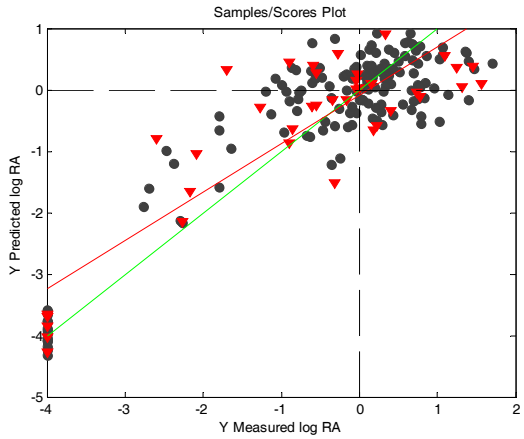

Figure 6: Plot of predicted vs. measured log RA as per Equation (4) after removal of outliers.

In the third analysis, MLR was coupled with forward stepwise and the best equation relating antimalarial activity and molecular descriptors is displayed below:

$$logRA = 31.3\ (PW4) - 1.21\ (GATS5m) + 5.15\ (BELm6) + 1.99\ (PJI3) - 0.929\ (E1m) + 11.9\ (G1e) - 10.3\ (Gs) + 7.50\ (H8m) - 0.697\ (H8e) + 2.16\ (O\text{-}063) - 17.9 \quad (5)$$

(r^2 = 0.821, r^2_{CV} = 0.794, *RMSEC* = 0.639, *RMSECV* = 0.685, r^2_{test} = 0.784, *RMSEP* = 0.778)

It was found that compounds numbered 29, 33, 94, 107 and 95 were outliers and the quality of the QSAR model had improved after the elimination of these compounds from the data set. A plot of observed versus predicted relative activity values from the final best linear model is as represented in *Figure 7*. The statistical output of stepwise MLR model mentioned above had confirmed the robustness and excellent external predictivity of the constructed model.

The standardized regression coefficients revealed the significance of an individual descriptor presented in the regression model. The order of significance of the descriptors was PW4 > G1e > Gs > H8m > BELm6 > O-063 > PJI3 > GATS5m > E1m > H8m. Obviously, the effect of molecular shape indices obtained by considering the number of paths and the number of walks within a graph for all atoms followed by WHIM descriptors which were related to symmetry index, on the activity of the artemisinin

derivatives were more significant than that of the other descriptors [23].

The properties of the molecules in this study that were responsible for antimalarial activity were determined on the basis of the information derived from the QSAR models. Description for each descriptor included in all the QSAR models described earlier is as tabulated in *Table 4*.

It is interesting to highlight that the descriptor O-063 emerged in all the three final models and was also the most frequently selected variables in models by GA. This descriptor indicated the presence of the R-O-O-R fragment of atom-centered fragments where the two adjacent oxygen atoms formed a peroxide functional group. Hence, this confirmed the previous finding that endoperoxide bridge seemed to be responsible for the antimalarial activity. Artemisinin derivatives lacking the endoperoxide bridge were devoid of antimalarial activity.

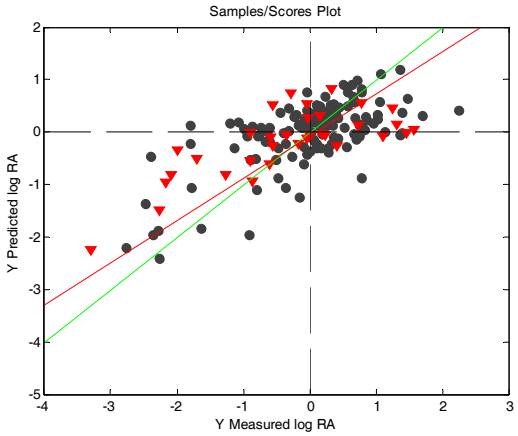

Figure 7: Predicted vs. experimental activity of compounds after removal of outliers

Table 4: Desciptors which were included in the QSAR models

Symbol	Type of descriptors	Desciption
nR05	Constitutional	Number of 5 membered rings
PW4	Topological	Path/walk 4-Randic shape index
GATS5m	2D-autocorrelations	Geary autocorrelation – lag 5 / weighted by atomic masses
GATS6v	2D-autocorrelations	Geary autocorrelation – lag 6 / weighted by atomic van der Waals
EEig05r	Edge adjacency indices	Eigenvalue 05 from edge adjacency matrix weighted by resonance integrals
BELm6	Burden eigenvalues	Lowest eigenvalue n.6 of Burden matrix / weighted by atomic masses
PJ13	Geometrical	Petitjean shape index
Mor03u	3D MoRSE	3D-MoRSE-signal 03 / unweighted
Mor29u	3D MoRSE	3D-MoRSE-signal 29 / unweighted
Mor26m	3D MoRSE	3D-MoRSE-signal 26 / weighted by atomic masses
Mor04v	3D MoRSE	3D-MoRSE-signal 04 / weighted by atomic van der Waals volumes.
RDF070u	RDF descriptors	Radial distribution function – 7.0 unweighted
E1m	WHIM descriptors	First compound accessibility directional WHIM index / weighted by atomic masses
G1e	WHIM descriptors	First component symmetry directional WHIM index / weighted by atomic Sanderson electronegativities
Gs	WHIM descriptors	Total symmetry index / weighted by atomic electrotopological states
H8e	GETAWAY descriptors	H autocorrelation of lag 8 / weighted by atomic Sanderson electronegativities
R6e$^-$	GETAWAY descriptors	R maximal autocorrelation of lag 6 / weighted by atomic Sanderson electronegativities
H8m	GETAWAY	H autocorrelation of lag 8 / weighted by atomic masses
C-011	Atom-centered fragments	CR3X
H-051	Atom-centered fragments	H attached to alpha-C
O-063	Atom-centered fragments	R-O-O-R
Qmean	Charge descriptors	mean absolute charge (charge polarization)

The linear models in this work did exhibit significant predictive ability and provide interpretability. However, the Forward Stepwise-MLR model appeared to be the best. Based on the results of artemisinin analogues, good QSAR model could be developed using the available QSAR methods and was comparable to the original study performed by other researchers [10-12].

4. CONCLUSION

The main objective of this work is to develop robust and predictive QSAR models of artemisinin and its derivatives that possess several different ring systems with antimalarial activity. The combination of descriptors generated by the DRAGON software was able to capture all the relevant structural features pertaining to antimalarial activity that reflected different aspects of molecular structure and potential intermolecular interactions. Hence, robust QSAR models with high internal and external prediction accuracy had been successfully developed in the current work. Based from the results, the best model (in terms of fitting and predictive ability) was generated by using Forward Stepwise and MLR. The final results could be further improved by using additional descriptors and other alternative modeling techniques.

Next, the rigorously validated models will subsequently be utilized in database mining. The discovery of a novel structural class of anti-malarial agents will then be confirmed experimentally. As such, these models shall be used as a basis to facilitate the design of new natural products as well as the search for new structures with anti-malarial activity from the large databases. Hopefully, these efforts are able to provide some contributions to global fights against malaria disease if not eliminate it.

ACKNOWLEDGEMENT

We thank Universiti Teknologi Malaysia for granting financial support and study leave for Rosmahaida Jamaludin.

REFERENCES

[1] Dhingra, V., Rao, V. K., & Narasu, L. M. (2000). Life Sciences. , 66(4), 279-300.
[2] Kaur, K., Jain, M., Kaur, T., & Jain, R. (2009). Bioorganic & Medicinal Chemistry, 17, 3229–3256.Avery 2002.
[3] Ploypradith, P. (2004). Acta Tropica 89, 329–342.
[4] Meshnick, S. R. (2002). International Journal for Parasitology 32, 1655–1660.
[5] Posner, G. H., Cumming, J. N., Ploypradith, P., & Oh, C. H. (1995). J. Am. Chem. Soc, 117, 5885-5886.
[6] Posner, G. H., Park, S. B., Gonza´lez, L. s., Wang, D., Cumming, J. N., Klinedinst, D., et al. (1996). J. Am. Chem. Soc. , 118, 3537-3538.
[7] Kamchonwongpaisan, S., & Meshnick, S. R. (1996). Gen. Pharmac. 27(4), 587-592.
[8] Olliaro, P. L., Haynes, R. K., Meunier, B., & Yuthavong, Y. (2001). TRENDS in Parasitology 17(3), 122-126.
[9] Cazelles, J., Robert, A., & Meunier, B. (2001). C. R. Acad. Sci. Paris, Chimie / Chemistry () 4 85–89.
[10] Avery, M. A., Alvim-Gaston, M., Rodrigues, C. R., Barreiro, E. J., Cohen, F. E., Sabnis, Y. A., et al. (2002). J. Med. Chem. , 45, 292-303.
[11] Guha, R., & Jurs, P. C. (2004). J. Chem. Inf. Comput. Sci., 44(4), 1440-1449.
[12] Srivastava, M. S., Singh, H., & Naik, P. K. (2009). Journal of Chemometrics, 23, 618–635.
[13] Beebe, K. R., Pell, R. J., & Seasholtz, M. B. (1998). Chemometrics: A practical Guide. New York: John Wiley & Sons, Inc.
[14] Leach, A. R., & Gillet, V. J. (2003). An Introduction to Chemometrics. Dordrecht: Kluwer Academic Publishers.
[15] Todeschini, R., Consonni, V., Mauri, A., & Pavan, M. (2006). DRAGON - Software for Molecular Descriptor Calculations (Version 5.4 for Windows). Milan, Italy: Talete srl.
[16] Todeschini, R., & Consonni, V. (2000). Handbook of Molecular Descriptors. Weinheim (Germany): Wiley-VCH.
[17] Leardi, R., & Gonzalez, A. L. (1998). Chemometrics and Intelligent Laboratory Systems, 41, 195-207.
[18] Hasegawa, K., & Funatsu, K. (1998). Journal of Molecular Structure (Theochem), 425, 255-262.
[19] Sutherland, J. J., & Weaver, D. F. (2003). J. Chem. Inf. Comput. Sci., 43(3), 1028-1036.
[20] Cho, S. J., & Hermsmeier, M. A. (2002). J. Chem. Inf. Comput. Sci., 42, 927-936.
[21] Sagrado, S., & Cronin, M. T. D. (2008). Analytica Chimica Acta 609 169–174.
[22] Leardi, R. (2007). Journal of Chromatography A, 1158, 226–233.
[23] Randic, M. (2001). J. Chem. Inf. Comput. Sci. , 41, 607-613.