Chemometrics and Intelligent Laboratory Systems 110 (2012) 81–88

Contents lists available at SciVerse ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

QSAR on aryl-piperazine derivatives with activity on malaria

Emmanuel Ibezim [a], Pablo R. Duchowicz [a,*], Erlinda V. Ortiz [b], Eduardo A. Castro [a]

[a] Instituto de Investigaciones Fisicoquímicas Teóricas y Aplicadas INIFTA (UNLP, CCT La Plata-CONICET), Diag. 113 y 64, C.C. 16, Suc.4, (1900) La Plata, Argentina
[b] Facultad de Tecnología y Ciencias Aplicadas, Universidad Nacional de Catamarca, Av. Maximio Victoria 55, (4700), Catamarca, Argentina

ARTICLE INFO

Article history:
Received 19 April 2011
Received in revised form 13 September 2011
Accepted 3 October 2011
Available online 8 October 2011

Keywords:
Aryl-piperazines
QSAR Theory
Molecular descriptors
Replacement Method
Artemisinin

ABSTRACT

In this work we offer linear regression models on a set of aryl-piperazine derivatives that are obtained by exploring a pool containing 1497 Dragon molecular descriptors, in order to establish the best relationships linking the molecular structure characteristics to their exhibited potencies against chloroquine resistant and chloroquine sensitive strains of *Plasmodium falciparum* parasite. The adjustment of the training molecular set together with the performance achieved during the internal and external validation processes leads to predictive QSAR models. In addition, we derive alternative linear models based on the Coral methodology, which lead to satisfactory results. We apply the final equations to predict the activity on some unknown compounds having non-observed activities.

1. Introduction

Malaria is a vector-borne infectious disease caused by protozoan parasites. It is widespread in tropical and subtropical regions, including parts of the Americas, Asia, and Africa. Every year, there are approximately 350–500 million cases of malaria [1], killing between one and three million people, the majority of whom are young children in Sub-Saharan Africa [2]. Ninety percent of malaria-related deaths occur in Sub-Saharan Africa. The advent of long-lasting insecticidal nets and Artemisinin-based combination therapy, plus a revival of support for indoor residual spraying of insecticide, presents a new opportunity for large-scale malaria control. Malaria is commonly associated with poverty, but is also a cause of poverty and a major hindrance to economic development [3].

People usually get malaria from the bite of *Anopheles* mosquitoes, the disease being caused by protozoan parasites of the genus *Plasmodium*. Five species of the *Plasmodium* parasite can infect humans; the most serious forms of the disease are caused by *Plasmodium falciparum*. Malaria caused by *Plasmodium vivax*, *Plasmodium ovale* and *Plasmodium malariae* causes milder disease in humans that is not generally fatal. A fifth species, *Plasmodium knowlesi*, causes malaria in macaques but can also infect humans. This group of human-pathogenic *Plasmodium* species is usually referred to as malaria parasites.

Several antimalarial drugs have been formulated for the treatment and prevention of the disease, but these have led to development of resistance by the parasites to most of the drugs in use. Specifically, there is reportedly, rapid spread of *P. falciparum* resistance to available antimalarial drugs [4]. Thus, there is a constant need for developing new antimalarial compounds. Ethnic medicine has provided two of the most efficacious drugs, Quinine and Artemisinin (and its analogs) and the ongoing screening of medicinal plants yields new lead compounds [5]. Work has been done on malaria vaccines with limited success and more exotic controls, such as genetic manipulation of mosquitoes to make them resistant to the parasite, have also been considered [6].

Although some vaccines are under development, none is currently available for malaria that provides a high level of protection [7]; preventive drugs must be taken continuously to reduce the risk of infection. These prophylactic drug treatments are often too expensive for most people living in endemic areas. Most adults from endemic areas have a degree of long-term infection, which tends to recur, and also possess partial immunity (resistance); the resistance reduces with time, and such adults may become susceptible to severe malaria if they have spent a significant amount of time in non-endemic areas.

In last decades, Quantitative Structure-Activity Relationships (QSAR) [8], have been applied in many areas enabling to prevent time consuming and cost during the analysis of biological activities of interest. The main hypothesis involved in any QSAR is the assumption that the variation of the behavior of chemical compounds, as expressed by any experimentally measured biological or physicochemical property, can be correlated with numerical entities related to some aspect of the chemical structure termed molecular descriptors [9,10]. Descriptors are generally used to describe different characteristics/attributes of the chemical structure in order to yield information about the activity/property being studied. In general, QSAR studies are effected by various factors from which the most relevant are: (a) the selection of the best molecular descriptors that should include maximum information

* Corresponding author. Fax: +54 221 425 4642.
E-mail addresses: pabloducho@gmail.com, prduchowicz@yahoo.com.ar (P.R. Duchowicz).

0169-7439/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2011.10.002

of molecular structures and a minimum overlap between them; (b) the optimal number of descriptors to be included in the model; (c) the use of suitable modeling methods; (d) the composition of the training and test sets; and (e) the employment of validation techniques to verify the predictive performance of the developed models.

We consider that the linear methodology is the best statistical technique for analyzing present dataset of aryl-piperazines, as few experimental observations are available on it and thus it is necessary to employ the lowest number of optimized parameters during the model development. In this way, we resort to the Replacement Method (RM) as variable subset selection approach applied on a pool containing more than a thousand of descriptors, as this technique has been successful for selecting relevant structural descriptors [11–15]. Finally, another main interest of present research is to apply the so derived QSAR models for estimating the antimalarial potency on some new structures, for which there still are no experimental activities.

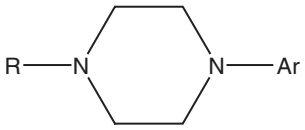
2. Materials and methods

2.1. Experimental data set

The experimental inhibitory concentrations (IC_{50}) in micromolar units of aryl-piperazine derivatives against the chloquine resistant strains W2 and FCR3 and against the chloroquine-sensitive strains D10 and NF54 are extracted from a recent publication [4]. In Table 1 we provide the experimental activities, and for modeling purposes these values are converted into logarithm units ($\log_{10} IC_{50}$).

A previous SAR study [4] has established that aryl-piperazines in which the terminal secondary amino group is unsubstituted are found to display a mefloquine-type antimalarial behavior in being significantly more potent against the chloroquine-resistant (W2 and FCR3) strains of *P. falciparum* than against the chloroquine-sensitive (D10 and NF54) strains. The substitutions at the secondary amino group lead to a dramatic drop in activity across all strains as well as disappearance of the preferential potency against resistant strains that are observed for the unsubstituted counterparts.

Table 1
Experimental antiplasmodial activity values (IC_{50}) for aryl-piperazine derivatives in micromolar units.



ID	R	Ar	W2	FCR3	NF54	D10
1	Hydrogen	Phenyl	24.8	13.8	92.5	152.7
2	Hydrogen	2-fluorophenyl	68.7	44.61	103.2	105.6
3	Hydrogen	4-fluorophenyl	24.9	19.8	95.2	135.0
4	Hydrogen	2-chlorophenyl	18.1	11.8	78.7	152.2
5	Hydrogen	3-chlorophenyl	4.67	4.69	59.1	67.7
6	Hydrogen	4-chlorophenyl	11.53	7.13	64.8	74
7	Hydrogen	3-trifluoromethylphenyl	11.49	9.56	92.9	112.8
8	Hydrogen	2-methoxyphenyl	35.75	30.33	83.7	151.2
9	Hydrogen	3-methoxyphenyl	26.5	32.3	85.3	228
10	Hydrogen	4-methoxyphenyl	66.2	61.6	115.7	167
11	Hydrogen	2-ethoxyphenyl	16.8	10.96	93.1	143.3
12	Hydrogen	2,8-di(trifluoromethyl) quinoline	11.2	2.03	15.0	7.8
13	Hydrogen	7-chloroquinoline	1.21	1.36	1.02	2.02
14	2-bromobenzyl	Phenyl	69.6	62.1	58	101.3
15	2-bromobenzyl	4-fluorophenyl	41.5	62.9	63.1	103.1
16	2-bromobenzyl	2-chlorophenyl	56.0	75.1	52.4	104.1
17	2-bromobenzyl	2-ethoxyphenyl	52.6	54.2	46.9	79.7
18	2-bromobenzyl	2,8-di(trifluoromethyl) quinoline	43.0	36.1	23.6	47.3
19	Ciclohexylmethyl	7-chloroquinoline	15.6	4.12	8.06	19.2
20	Benzyl	7-chloroquinoline	16.9	13.5	11.62	9.04
21	2-bromobenzyl	7-chloroquinoline	58.1	20.1	11.46	18.3
22	2-iodobenzyl	7-chloroquinoline	13.0	6.49	3.9	12.3

2.2. Geometry optimization and molecular descriptors calculation

The initial conformations of the compounds are drawn by means of the "Model Build" modulus of the HyperChem 6.03 program for Windows [16]. We pre-optimize the molecular structures with the Molecular Mechanics Force Field (MM+) procedure included in the HyperChem, and refine the resulting geometries by means of the Semiempirical Method PM3 from the Molecular Orbitals Theory using the Polak–Ribiere algorithm and a gradient norm limit of 0.01 kcal.Å^{-1}.

Afterwards, we compute 1497 molecular descriptors using the Dragon program [9], including descriptors of all types such as Constitutional, Topological, Geometrical, Charge, GETAWAY (Geometry, Topology and Atoms-Weighted AssemblY), WHIM (Weighted Holistic Invariant Molecular descriptors), 3D-MoRSE (3D-Molecular Representation of Structure based on Electron diffraction), Molecular Walk Counts, BCUT descriptors, 2D-Autocorrelations, Aromaticity Indices, Randic Molecular Profiles, Radial Distribution Functions, Functional Groups, Atom-Centred Fragments, Empirical and Properties [17]. We also include in the analysis 5 descriptors obtained from the semiempirical calculation (molecular dipole moment, total energy, energy of the HOMO and LUMO molecular orbitals, and HOMO-LUMO gap).

In addition, we calculate atomic charge density-based descriptors encoding electronic and structural information relevant to the chemistry of intermolecular interactions, by means of the Recon 5.5 software [18]. This sort of computed descriptors are not provided by Dragon software, while the robustness of Recon has previously been demonstrated elsewhere [19, 20]. Recon is an algorithm for the reconstruction of molecular charge densities and charge density-based electronic properties of molecules, using atomic charge density fragments precomputed from ab initio wavefunctions. The method is based on the Quantum Theory of Atoms in Molecules [21]. A library of atomic charge density fragments has been built in a form that allows for the rapid retrieval of the fragments and molecular assembly. In present case, the smiles chemical notation is employed as input for the generation of 248 Transferable Atom Equivalent (TAE) descriptors, developed by Breneman and co-workers [22].

In this way, the total number of calculated structural descriptors for the molecular set under analysis results in 1750 variables.

2.3. Model development

The QSAR established in this work are obtained via two different linear modeling approaches with the purpose of comparing the consistency of our results: a) the search of molecular descriptors via multivariable linear regressions; and b) the calculation of flexible descriptors with the CORAL (CORrelation And Logic) program.

2.3.1. Linear descriptors search

In recent years theoretical and experimental researchers have focused an increasing attention on finding the most efficient tools for selecting molecular descriptors in QSAR studies. There is a great number of available feature selection methods to search the best structural descriptors from a pool of variables, and the Replacement Method (RM) [23, 24], employed here, has been successfully applied elsewhere [11–14, 25]. In brief, the RM is an efficient optimization tool which generates multi-parametric linear regression QSAR models on a training (calibration) molecular set by searching the set **D** of D descriptors for an optimal subset **d** of $d \ll D$ ones with minimum model's standard deviation (S). The quality of the results achieved with this technique approaches that obtained by performing an exact (combinatorial) full search of molecular descriptors although,

of course, requires much less computational work. Finally, RM results consider the Variance Inflation Factor (*VIF*), a method of detecting the severity of multicollinearity which represents a high degree of correlation (linear dependency) among several independent variables [26, 27]. The VIF_{ij} for a given descriptor *i* can be easily calculated (Eq. (1)) if we know the correlation coefficient between that descriptor and the remaining *j* ones of the model (R_{ij}):

$$VIF_{ij} = \frac{1}{1-R_{ij}^2}. \quad (1)$$

In practice, when $VIF_{ij} > 10$ then this would indicate that there exists significant multicollinearity in the chosen subset of descriptors.

2.3.2. The CORAL method

CHEMPREDICT/CORAL (CORrelation And Logic) version 1.4 [28], is a freeware for Windows. Each molecular structure must be represented by SMILES (Simplified Molecular Input Line Entry System) notation, calculated with ACD/ChemSketch software [29]. CORAL approach is based on the presence of certain SMILES attributes occurring in the molecule which can be associated to the activity of the molecule under evaluation [30–33]. As SMILES attributes are used the symbols representing the chemical elements and the other symbols used in SMILES for cycles, branching of molecular skeleton, charges, etc. The CORAL modeling process considers not only the presence of individual elements SMILES attributes (s_k), but also clusters of two (ss_k) and three (sss_k) elements. For example, SMILES = Clc1ccccc1 then s_k = (Cl, c, 1, c, c, c, c, c, 1); ss_k = (Clc, c1, cc, cc, cc, cc, cc, c1); sss_k = (Clc1, c1c, ccc, ccc, ccc, ccc, cc1).

The fragment-based model is a one-variable linear correlation between the activity values and the flexible descriptor (*DCW*) that is defined as:

$$DCW(\text{threshold}) = \alpha \sum_k CW(s_k) + \beta \sum_k CW(ss_k) + \gamma \sum_k CW(sss_k). \quad (2)$$

where α, β, γ are 1 or 0, and *CW* is the correlation weight for the element/s of the SMILES. The threshold is the parameter to define rare (noise) SMILES attributes. The rare SMILES attributes can lead to overtraining: excellent correlation for the training set accompanied by poor correlation for the validation set. Thus they can bring 'noise'. The threshold can be defined as 0, 1, 2, …*N*, with *N* being the number of compounds in training set. If threshold is defined 5, all SMILES attributes that take place in less than 5 SMILES notations of the training set will be classified as rare. In present study, numerical data for *CW* can be calculated by the Monte Carlo simulation by maximizing *R* parameter, the correlation coefficient between the activity values and the *DCW* descriptor defined in Eq. (2) for the training molecular set. The quality of the prediction is dependent on the selected options/parameters in the algorithm, such as the number of epochs used during the Monte Carlo optimization procedure, D_{start}, $d_{precision}$, dR_{weight}, dC_{weight}, threshold range and others, which should be correctly specified in order to calculate the *DCW* values. More specific details on the CORAL algorithm can be found in the recent literature [30–33].

2.3.3. Analysis of the Happenstance of the model

Another simple way of proving that the structure-activity relationships established in this study do not result from happenstance involves checking their robustness by means of the so-called y-randomization [34]. This technique consists on scrambling the experimental property values in such a way that they do not correspond to the respective compounds. After analyzing 10,000 cases of Y-Randomization for each developed QSAR, the smallest standard deviation value obtained using this procedure (S^{Rand}) turned out to be a poorer (greater) value when compared to the one found when considering the true calibration (*S*). Therefore, the correlations found are not fortuitous and result in real structure-activity relationships.

2.3.4. Model validation

Every QSAR research has to fulfill the basis of the QSAR hypothesis, that is to say, the appropriate validation of the defined mathematical models in order to verify that these relationships behave predictive and are not only limited to work correlatively on the training set. The theoretical validation practiced over each linear regression developed is based on the Leave-One-Out Cross Validation procedure (loo) [35]. Statistical parameters R_{loo} and S_{loo} measure the stability of the developed QSAR upon inclusion/exclusion of compounds, and according to the specialized literature, R_{loo}^2 should be greater than 0.7 for obtaining a validated model [36]. However, from our own experience in establishing QSAR models, $R_{loo}^2 < 0.7$ could also lead to satisfactory models, as the Leave-More-Out technique provides the predictive power of the model by defect, in the sense that no-compound should be excluded and all training compounds should be present during the cross-validation evaluation. Therefore, we consider that R_{loo}^2 should be greater than the value 0.7 but this is not an exclusive rule.

We also apply a rigorous and more realistic validation that consists on omitting from the complete molecular set presented in Table 1 some compounds which constitute the 'test set', denoted here as 'test'. The main purpose of performing such a splitting is to assess whether the QSAR found have predictive capability for estimating the activity on the "fresh" test set compounds (never seen by the model). We select the molecules composing the training and test series as a previous step to the model search, and this is done in such a way that both sets share similar qualitative structure-activity characteristics. In addition, we use the same molecules as test set for each strain of the *P. falciparum* parasite.

2.3.5. Degree of contribution of selected descriptors

In order to determine the relative importance of each descriptor in the linear regression model, we calculate standardized regression coefficients (b_j^s) through the following equation:

$$b_j^s = \frac{s_j . b_j}{s_Y} \qquad j = 1,\ldots,d \quad (3)$$

where is the regression coefficient of descriptor *j*, and s_j and s_Y are the standard deviations for that descriptor and for the activity, respectively. Eq. (3) allows one to assign a greater importance to those molecular descriptors that exhibit larger absolute standardized coefficients [37].

3. Results and discussion

We use the Matlab 7.0 program in all our calculations [38]. In every reported QSAR, *N* is the number of training set molecules, *range* is the experimental range of activities covered by the model, *d* the number of descriptors of the model, *R* is the correlation coefficient, *S* the model's standard deviation, *F* is the Fisher parameter, *res* the residual for a given molecule (difference between the experimental and predicted activity), outliers > x.*S* indicates the number of molecules predicted to have *res* greater than x times *S*, $Corr^{max}$ represents the maximum intercorrelation coefficient between two given descriptors of the model, VIF_{ij} is the variance inflation factor, *loo* subindex belong to the Leave-One-Out Cross Validation result, test subindex applies to the test set, and *Rand* supraindex stands for Y-Randomization.

The following linear QSAR are obtained on each strain, for which we discuss the relative contributions of the structural descriptors to the predicted antiplasmodial activities:

W2 Strain:

$$\log_{10}IC_{50} = -1.628(\pm0.8) - 0.107(\pm0.02)\cdot DISPv + 1.982(\pm0.5)\cdot R3e \quad (4)$$

$N=16$, $range=0.0828–1.843$, $d=2$, $N/d=8$, $R=0.889$, $S=0.22$, $F=24.3$, outliers $>2.5S=0$, $Corr^{max}=0.573$, $R_{loo}=0.833$, $S_{loo}=0.27$, $S^{Rand}=0.24$, $R_{test}=0.365$.

For the case of Eq. (4), both *DISPv* and *R3e* descriptors take positive numerical values. Their absolute standardized regression coefficients are 1.083 and 0.661, respectively, and thus *DISPv* is the most important variable of this model. The sign of the regression coefficients in Eq. (4) suggests that increasing values of *DISPv* and decreasing values of *R3e* would lead to lower predicted IC_{50} values.

FCR3 Strain:

$$\log_{10}IC_{50} = 1.743(\pm0.1) - 0.059(\pm0.02)\cdot DISPv + 1.694(\pm0.4)\cdot Mor29m \quad (5)$$

$N=16$, $range=0.134–1.876$, $d=2$, $N/d=8$, $R=0.902$, $S=0.25$, $F=28.5$, outliers $>2.S=0$, $Corr^{max}=0.399$, $R_{loo}=0.826$, $S_{100}=0.34$, $S^{Rand}=0.28$, $R_{test}=0.906$.

This equation has similar contributions to the modeled activity for descriptors *DISPv* ($b^s_{DISPv}=0.493$) and *Mor29m* ($b^s_{Mor29m}=0.584$): the second variable can either take positive or negative numerical values. The sign of the regression coefficients in Eq. (5) indicates that increasing values of *DISPv* and decreasing values of *Mor29m* would tend to generate lower predicted IC_{50} values.

D10 Strain:

$$\log_{10}IC_{50} = -13.389(\pm1) + 12.577(\pm0.9)\cdot AROM + 1.349(\pm0.2)\cdot R2e \quad (6)$$

$N=16$, $range=0.00860–2.063$, $d=2$, $N/d=8$, $R=0.968$, $S=0.16$, $F=96.8$, outliers $>2.5.S=0$, $Corr^{max}=0.323$, $R_{loo}=0.934$, $S_{loo}=0.23$, $S^{Rand}=0.25$, $R_{test}=0.987$

Eq. (6) has *AROM* as the most important descriptor ($b^s_{AROM}=1.019$), while *R2e* has $b^s_{R2e}=0.412$. As both parameters are positive quantities and the sign of the regression coefficients are positive, increasing values of descriptors in Eq. (6) would lead to higher predicted activities.

NF54 Strain:

$$\log_{10}IC_{50} = -13.117(\pm1) + 14.835(\pm1)\cdot AROM + 0.0551(\pm0.01)\cdot RDF030e \quad (7)$$

$N=16$, $range=0.305–2.358$, $d=2$, $N/d=8$, $R=0.969$, $S=0.15$, $F=99.3$, outliers $>2.5.S=0$, $Corr^{max}=0.730$, $R_{loo}=0.955$, $S_{loo}=0.18$, $S^{Rand}=0.21$, $R_{test}=0.948$.

As happen in the previous equation, here *AROM* is the most important descriptor ($b^s_{AROM}=1.251$), while *RDF030e* has $b^s_{RDF030e}=0.459$. The increment of these two positive descriptors leads to higher predicted antiplasmodial activities.

All the above models obey the semiempirical "Rule of Thumb", stating that at least five or six data points should be present per descriptor [39]: a single descriptor does not achieve enough accuracy for predicting the activities, while two-descriptors models are acceptable for the number of training molecules involved ($N=16$). Dispersion plots of residuals (residuals as function of predicted activities) for each QSAR are provided in Figs. S1–S8 of Supplementary figures with the purpose of demonstrating the validity of these multivariable linear regressions. Although some outliers are detected in these plots having residuals exceeding the 2.*S* value, we decide to derive a general model having applicability to any biomolecule without restrictions, and so we do not remove such molecules from the training set.

Correlation matrices together with the numerical values for each descriptor appearing in the established models are also provided as part of the supplementary material in Tables S1 and S2, respectively. The predicted activities for each QSAR are supplied in Table 2, from which it is appreciated that Eqs. (4)–(7) predict the experimental activities of the test set compounds reasonably well (test data denoted with ^) and thus the models are predictive and properly validated. A straight line trend is observed for the predicted log $_{10}IC_{50}$ as function of experimental values in Figs. 1–8.

The Variance Influence Factor (VIF_{ij}) parameter for each chosen descriptor in each model (Table S1) suggests that the descriptors are non-collinear and include non-redundant structural information content. All the models given in Eqs. (4)–(7) require conformational-dependent molecular descriptors. The geometrical variable *DISPv* is a 3D-descriptor obtained from moment expansions that do not require molecular superposition or alignment for the assignment of molecular similarity, incorporating information about the magnitude of the displacement between the molecular centroid (center of mass) and the polarizability-field center (center of charge) [40]. This kind of parameters has been found valuable for the prediction of the electrophoretic mobilities of peptides.

GETAWAY (GEometry, Topology, and Atom-Weights AssemblY) [41], type of descriptors *R2e* and *R3e*, the R autocorrelation of lag 2 and lag 3, respectively, are weighted by atomic Sanderson electronegativities. GETAWAY have been designed with the main purpose of matching the 3D-molecular geometry and are derived from the elements h_{ij} of the Molecular Influence matrix (**H**), obtained through the values of atomic Cartesian coordinates. The diagonal elements of **H** (h_{ii}) are called leverages, and represent the influence of each atom on the shape of the molecule. For instance, the mantle atoms always have higher h_{ii} values than atoms near the molecule center, while each off-diagonal element h_{ij} represents the degree of accessibility of the j^{th} atom to interactions with the i^{th} atom. The influence/distance

Table 2
Experimental and predicted antimalarial activities on different strains of *Plasmodium falciparum* obtained with linear models based on RM technique.

Strain	log $_{10}IC_{50}$(W2)		log $_{10}IC_{50}$ (FCR3)		log $_{10}IC_{50}$(D10)		log $_{10}IC_{50}$ (NF54)	
ID	Exp.	Eq. (4)	Exp.	Eq. (5)	Exp.	Eq. (6)	Exp.	Eq. (7)
1^	1.394	1.623	1.140	1.268	1.966	1.858	2.184	2.057
2	1.837	1.711	1.649	1.339	2.014	1.826	2.024	1.995
3	1.396	1.616	1.297	1.170	1.979	1.909	2.130	2.064
4	1.258	1.337	1.072	1.042	1.896	1.820	2.182	2.110
5^	0.669	1.055	0.671	0.627	1.772	1.836	1.831	2.051
6	1.062	0.882	0.853	0.645	1.812	1.855	1.869	2.063
7	1.060	1.436	0.980	1.212	1.968	2.318	2.052	2.229
8^	1.553	1.049	1.482	1.432	1.923	1.799	2.180	2.103
9	1.423	1.477	1.509	1.558	1.931	2.020	2.358	2.047
10	1.821	1.848	1.790	1.666	2.063	2.029	2.223	2.045
11	1.225	1.247	1.040	1.143	1.969	1.877	2.156	2.072
12	1.049	0.904	0.307	0.686	1.176	1.078	0.892	0.818
13	0.083	0.255	0.134	0.512	0.009	0.272	0.305	0.378
14	1.843	1.713	1.793	2.091	1.763	1.759	2.006	2.086
15^	1.618	1.739	1.799	1.793	1.800	1.668	2.013	2.161
16	1.748	1.672	1.876	1.930	1.719	1.621	2.017	2.212
17	1.721	1.275	1.734	1.427	1.671	1.747	1.901	2.010
18	1.633	1.502	1.558	1.357	1.373	1.324	1.675	1.564
19	1.193	1.156	0.615	0.337	0.906	0.702	1.283	1.295
20^	1.228	0.837	1.130	0.693	1.065	0.707	0.956	1.164
21^	1.764	1.162	1.303	1.178	1.059	0.676	1.262	1.166
22	1.114	1.437	0.812	0.904	0.591	0.684	1.090	1.177

^Test set compound.

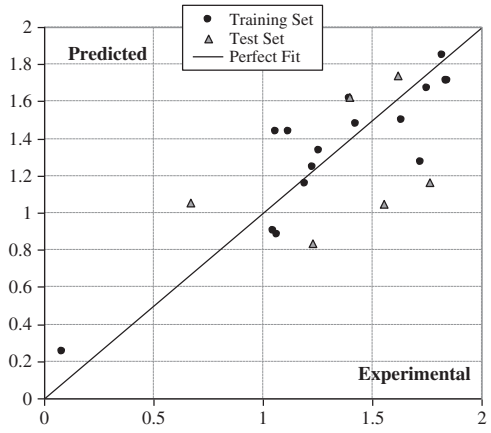


Fig. 1. Predicted log $_{10}IC_{50}[mM]$ for W2 strain according to Eq. (4) as function of experimental values.

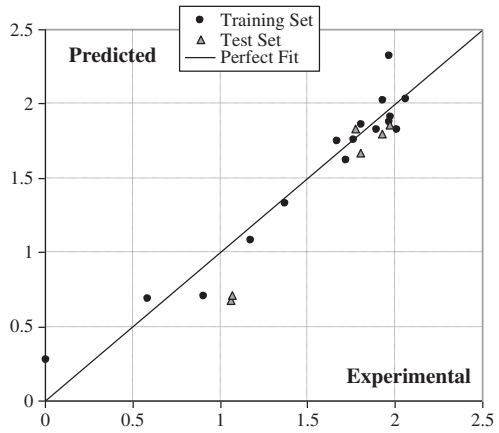


Fig. 3. Predicted log $_{10}IC_{50}[mM]$ for D10 strain according to Eq. (6) as function of experimental values.

matrix (**R**) involves a combination of the elements of the **H** matrix with those of the Geometric Matrix (**G**).

The 3D-MoRSE (3D Molecule Representation of Structure based on Electron diffraction) descriptor *Mor29m* is the signal 29/weighted by atomic masses. Such kind of descriptors is obtained through the molecular transform generally employed in electron diffraction studies [42]. Electron diffraction does not directly yield atomic coordinates, but provides diffraction patterns from which the atomic coordinates are derived by mathematical transformations. These descriptors are defined in order to reflect the contribution to the biological activity under investigation, at a prescribed scattering angle, of a given atomic property, and also employ weights for distinguishing atoms.

The aromaticity index *AROM* measures the degree of aromaticity of a compound [17]. It can only be calculated once the molecular geometry is properly optimized with an accurate electronic structure method. Another descriptor is *RDF*030*e*, the Radial Distribution Function 3.0/weighted by atomic Sanderson electronegativities [41]. The 3D-Radial Distribution Functions descriptors defined for an ensemble of atoms may be interpreted as the probability distribution of finding an atom in a spherical volume of certain radius, also incorporating different atomic properties in order to differentiate the contribution of atoms to the activities. For the case of *RDF*030*e*, the sphere radius is of 3.0 Å and atomic Sanderson electronegativities are employed for atoms.

Now, it is feasible to improve the statistical performance of Eqs. (4)–(7) by using linear models established via flexible descriptor definitions calculated with the CORAL program. We run a Monte Carlo simulation for obtaining the correlation weights of Eq. (2), achieving the following QSAR models:

W2 Strain:

$$\log_{10}IC_{50} = -0.116(\pm 0.1) + 0.0709(\pm 0.005)\cdot DCW(W2) \quad (8)$$

$N = 16$, $range = 0.0828–1.843$, $d = 1$, $N/d = 16$, $R = 0.964$, $S = 0.12$, $F = 186.6$, outliers $> 3.S = 1$, $R_{loo} = 0.958$, $S_{loo} = 0.13$, $S^{Rand} = 0.21$, $R_{test} = 0.887$.

FCR3 Strain:

$$\log_{10}IC_{50} = -1.890(\pm 0.2) + 0.0960(\pm 0.006)\cdot DCW(FCR3) \quad (9)$$

$N = 16$, $range = 0.134–1.876$, $d = 1$, $N/d = 16$, $R = 0.975$, $S = 0.13$, $F = 270.9$, outliers $> 2.5.S = 0$, $R_{loo} = 0.970$, $S_{loo} = 0.14$, $S^{Rand} = 0.34$, $R_{test} = 0.932$.

D10 Strain:

$$\log_{10}IC_{50} = -0.160(\pm 0.03) + 0.0782(\pm 0.001)\cdot DCW(D10) \quad (10)$$

$N = 16$, $range = 0.00860–2.063$, $d = 1$, $N/d = 16$, $R = 0.998$, $S = 0.040$, $F = 3310.2$, outliers $> 3.S = 0$, $R_{loo} = 0.997$, $S_{loo} = 0.044$, $S^{Rand} = 0.28$, $R_{test} = 0.940$.

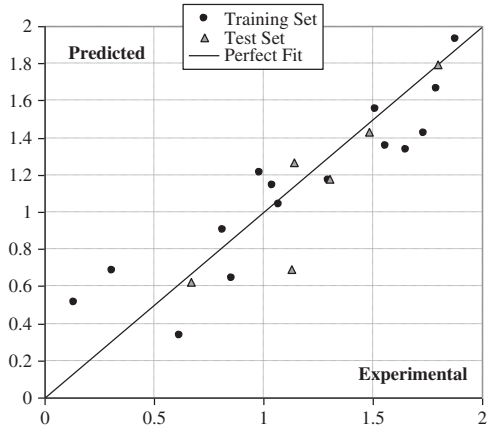


Fig. 2. Predicted log $_{10}IC_{50}[mM]$ for FCR3 strain according to Eq. (5) as function of experimental values.

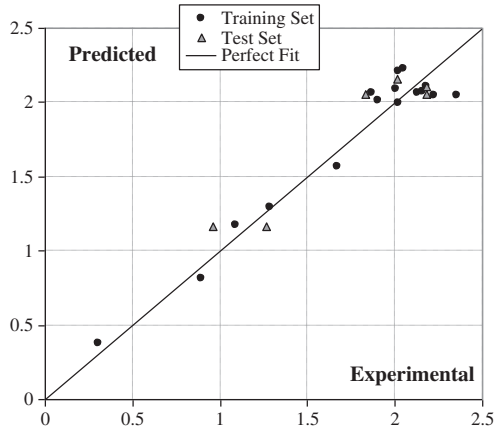


Fig. 4. Predicted log $_{10}IC_{50}[mM]$ for NF54 strain according to Eq. (7) as function of experimental values.

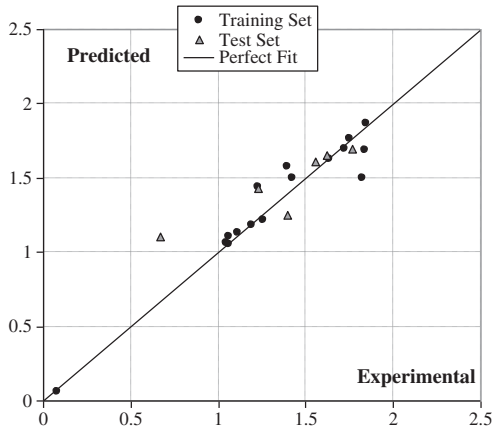


Fig. 5. Predicted log $_{10}IC_{50}[mM]$ for W2 strain according to Eq. (8) as function of experimental values.

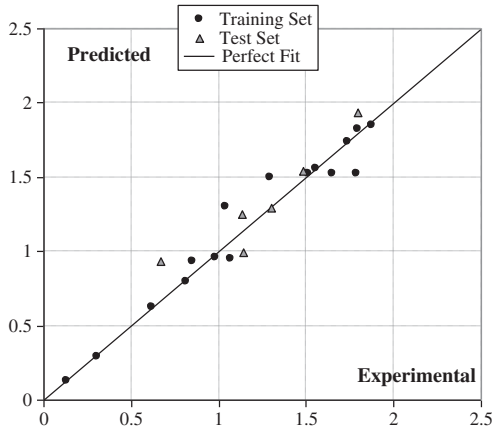


Fig. 6. Predicted log $_{10}IC_{50}[mM]$ for FCR3 strain according to Eq. (9) as function of experimental values.

NF54 Strain:

$$\log_{10}IC_{50} = 0.544(\pm 0.05) + 0.0539(\pm 0.002)\cdot DCW(NF54) \quad (11)$$

$N=16$, $range=0.305$–2.358, $d=1$, $N/d=16$, $R=0.991$, $S=0.078$, $F=789.4$, outliers $>2.5.S=0$, $R_{loo}=0.989$, $S_{loo}=0.086$, $S^{Rand}=0.19$, $R_{test}=0.950$.

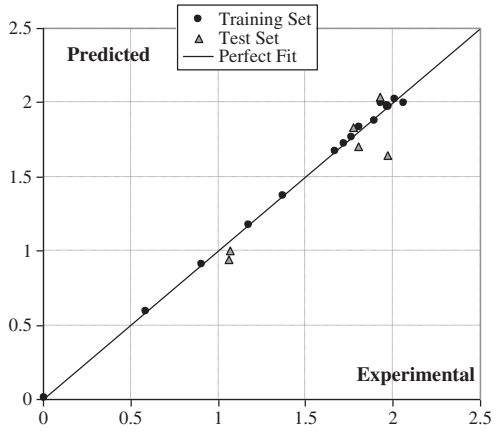


Fig. 7. Predicted log $_{10}IC_{50}[mM]$ for D10 strain according to Eq. (10) as function of experimental values.

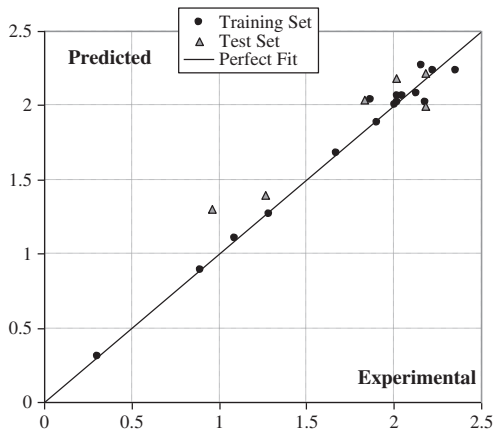


Fig. 8. Predicted log $_{10}IC_{50}[mM]$ for NF54 strain according to Eq. (11) as function of experimental values.

The numerical parameters used in the CORAL calculation are: number of epochs: 20, number of probes: 3, range of threshold values: 0–2, $D_{start}=0.5$, $d_{precision}=0.01$, $dR_{weight}=0$, $dC_{weight}=0$, threshold range = 0–2, and $\alpha=\beta=0$ (refer to Eq. (2)). Figs. 5–8 plot the predicted activities as function of the experimental data. The predictions achieved by models 8–11 are included in Table 3.

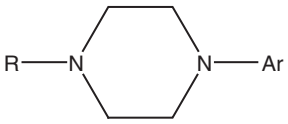
It is easily appreciated from the statistical parameters of calibration and leave-one-out validation that the quality of Eqs. (8)–(11) outperforms that of Eqs. (4)–(7). However, an important remark has to be made upon establishing Eqs. (8)–(11): as a linear optimization problem having a greater number of parameters is involved in the CORAL method, it results not possible to establish a predictive model by minimizing only S of the training set, and the test set statistics has also to be monitored. Therefore, it is not possible to establish flexible descriptors-based models without considering the independent test set data, as it is the case of Eqs. (4)–(7). We decide to include the CORAL models in order to compare their predictions on unknown compounds.

Table 3
Experimental and predicted antimalarial activities on different strains of *Plasmodium falciparum* obtained with linear models based on the Coral method.

Strain	log $_{10}IC_{50}$(W2)		log $_{10}IC_{50}$ (FCR3)		log $_{10}IC_{50}$(D10)		log $_{10}IC_{50}$(NF54)	
ID	Exp.	Eq. (8)	Exp.	Eq. (9)	Exp.	Eq. (10)	Exp.	Eq. (11)
1 ^	1.394	1.247	1.140	0.996	1.966	1.643	2.184	1.992
2	1.837	1.689	1.649	1.520	2.014	2.021	2.024	2.067
3	1.396	1.578	1.297	1.500	1.979	1.972	2.130	2.079
4	1.258	1.216	1.072	0.954	1.896	1.879	2.182	2.025
5 ^	0.669	1.105	0.671	0.934	1.772	1.829	1.831	2.037
6	1.062	1.105	0.853	0.934	1.812	1.829	1.869	2.037
7	1.060	1.052	0.980	0.961	1.968	1.968	2.052	2.059
8 ^	1.553	1.613	1.482	1.541	1.923	2.042	2.180	2.221
9	1.423	1.502	1.509	1.521	1.931	1.992	2.358	2.234
10	1.821	1.502	1.790	1.521	2.063	1.992	2.223	2.234
11	1.225	1.435	1.040	1.301	1.969	1.979	2.156	2.265
12	1.049	1.062	0.307	0.290	1.176	1.176	0.892	0.890
13	0.083	0.061	0.134	0.128	0.009	0.009	0.305	0.308
14	1.843	1.864	1.793	1.820	1.763	1.763	2.006	2.003
15 ^	1.618	1.654	1.799	1.935	1.800	1.702	2.013	2.183
16	1.748	1.765	1.876	1.851	1.719	1.719	2.017	2.024
17	1.721	1.698	1.734	1.740	1.671	1.671	1.901	1.887
18	1.633	1.625	1.558	1.555	1.373	1.373	1.675	1.678
19	1.193	1.186	0.615	0.627	0.906	0.906	1.283	1.271
20 ^	1.228	1.434	1.130	1.248	1.065	0.999	0.956	1.299
21 ^	1.764	1.693	1.303	1.291	1.059	0.946	1.262	1.391
22	1.114	1.128	0.812	0.796	0.591	0.591	1.090	1.103

^Test set compound.

Table 4
Predicted antiplasmodial activities for aryl-piperazine derivatives according to Eqs. (4)–(11).



ID[a]	R	Ar	log $_{10}IC_{50}$(W2)		log $_{10}IC_{50}$(FCR3)		log $_{10}IC_{50}$(NF54)		log $_{10}IC_{50}$(D10)	
			Eq. (4)	Eq. (8)	Eq. (5)	Eq. (9)	Eq. (6)	Eq. (10)	Eq. (7)	Eq. (11)
53	Hydrogen	8-chloro quinoline	0.510	0.446	0.761	0.255	0.268	0.135	0.431	0.528
54	Hydrogen	6-chloro quinoline	0.490	0.146	0.610	0.233	0.267	0.013	0.372	0.281
61	Hydrogen	3-trifluoromethyl 7-chloro quinoline	0.363	0.733	0.360	0.660	0.486	1.451	0.482	1.422
62	Hydrogen	3-trifluoromethyl 6-chloro quinoline	0.606	0.685	0.384	0.559	0.483	1.398	0.503	1.459
66	Hydrogen	2-chloro 6-trifluoromethyl quinoline	0.554	0.499	0.914	0.067	0.839	0.978	0.673	0.707
67	Hydrogen	2-chloro 5-trifluoromethyl quinoline	0.575	0.415	0.513	−0.038	0.482	0.971	0.658	0.734

[a] Numbering of compounds refer to Table S3.

Eqs. (4)–(11) are applied to predict unknown aryl-piperazine compounds, with the purpose of extracting new biochemical information from the QSAR models designed in this work. We draw and optimize 145 structurally-related compounds which do not exhibit experimentally assigned antiplasmodial activities (refer to Table S3). The top-six most active predicted compounds are included in Table 4. In accordance with a previously established structure-activity relationships (SAR) study [4], the most active unknown compounds provided in this table appear unsubstituted at the secondary amino group. Eqs. (4)–(7) derived with Dragon theoretical descriptors and Eqs. (8)–(11) obtained with the Coral methodology offer similar trends for the predictions of these unknown molecules although, however, one may pay more attention to the results obtained with Eqs. (4)–(7) that are supposed to be more truthful for obeying to linear equations involving a fewer number of optimized parameters in their formulations.

4. Conclusions

We hope that the QSAR established in this work may serve as a guide for providing the structural requirements affecting the antimalarial activities of aryl-piperazine derivatives, through the identification of the most relevant selected molecular descriptors in the models. In this line, we applied the developed QSAR to predict some unknown structurally-related molecules. We are especially careful in validating the relationships with the Leave-One-Out Cross Validation method and by leaving some of the molecules as part of an external test set. Finally, the results presented in this work resort to two different methodologies: (a) application of linear models for selecting the most relevant structural parameters, and (b) employment of flexible (property dependent) molecular descriptors.

Supplementary materials related to this article can be found online at doi:10.1016/j.chemolab.2011.10.002.

Acknowledgments

E.I. would like to thank the Third World Academy of Sciences (TWAS) for a fellowship that made possible this work. P.R.D. and E.A.C. thank to the Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET) and the Universidad Nacional de La Plata for supporting this work.

References

[1] Malaria Facts. Centers for Disease Control and Prevention, 2011. http://www.cdc.gov/malaria/about/facts.html.
[2] R.W. Snow, C.A. Guerra, A.M. Noor, H.Y. Myint, S.I. Hay, Nature 434 (2005) 214.
[3] Malaria: Disease Impacts and Long-Run Income Differences (PDF). Institute for the Study of Labor, 2007. http://ftp.iza.org/dp2997.pdf
[4] C.A. Molyneaux, M. Krugliak, H. Ginsburg, K. Chibale, Biochemical Pharmacology 71 (2005) 61.
[5] M.L. Willcox, G. Bodeker, BMJ 329 (2004) 1156.
[6] S. Yoshida, Y. Shimada, D. Kondoh, Y. Kouzuma, A.K. Ghosh, M. Jacobs-Lorena, R.E. Sinden, PLoS Pathogens 3 (2007) 1962.
[7] [7] RTS,S vaccine protection rate.
[8] C. Hansch, A. Leo, Exploring QSAR, Fundamentals and Applications in Chemistry and Biology, American Chemical Society, Washington, D. C, 1995.
[9] Milano Chemometrics and QSAR Research Group, http://michem.disat.unimib.it/chm.
[10] M. Karelson, Molecular Descriptors in QSAR/QSPR, Wiley-Interscience, New York, 2000.
[11] P.R. Duchowicz, M. Fernández, J. Caballero, E.A. Castro, F.M. Fernández, Bioorganic & Medicinal Chemistry 14 (2006) 5876.
[12] P.R. Duchowicz, M.G. Vitale, E.A. Castro, M. Fernandez, J. Caballero, Bioorganic & Medicinal Chemistry 15 (2007) 2680.
[13] P.R. Duchowicz, A. Talevi, L.E. Bruno-Blanch, E.A. Castro, Bioorganic & Medicinal Chemistry 16 (2008) 7944.
[14] M. Goodarzi, P.R. Duchowicz, C.H. Wu, F.M. Fernández, E.A. Castro, Journal of Chemical Information and Modeling 49 (2009) 1475.
[15] E. Vicente, P.R. Duchowicz, E.A. Castro, A. Monge, Journal of Molecular Graphics & Modelling 28 (2009) 28.
[16] (Hypercube, Inc.). Hyperchem 7, , 2007http://www.hyper.com Gainesville.
[17] R. Todeschini, V. Consonni, Molecular Descriptors for Chemoinformatics, Wiley-VCH, Weinheim, 2009.
[18] Recon, Version 5.5. Rensselaer Polytechnic Institute, Troy, New York, USA, http://www.drugmining.com.
[19] B.K. Lavine, C.E. Davidson, C. Breneman, W. Katt, Journal of Chemical Information and Computer Sciences 43 (2003) 1890.
[20] A. Worachartcheewan, C. Nantasenamat, T. Naenna, C. Isarankura-Na-Ayudhya, V. Prachayasittikul, European Journal of Medicinal Chemistry 44 (2009) 1664.
[21] R.F.W. Bader, Atoms in Molecules — A Quantum Theory, Clarendon Press, Oxford, 1990.
[22] C.M. Breneman, L.W. Weber, The Application of Charge Density Research to Chemistry and Drug Design, Plenum, New York, in: G.A. Jeffrey, J.F. Piniella (Eds.), , 1991.
[23] P.R. Duchowicz, E.A. Castro, F.M. Fernández, M.P. González, Chemical Physics Letters 412 (2005) 376.
[24] A.G. Mercader, P.R. Duchowicz, F.M. Fernández, E.A. Castro, Chemometrics and Intelligent Laborary Systems 92 (2008) 138.
[25] P.R. Duchowicz, M. Goodarzi, M.A. Ocsachoque, G.P. Romanelli, E.V. Ortiz, J.C. Autino, D.O. Bennardi, D. Ruiz, E.A. Castro, Science of the Total Environment 408 (2009) 277.
[26] Multicollinearity.doc © 2002 Jeeshim and KUCC625. (2003-05-09).
[27] J.D. Curto, J.C. Pinto, International Statistics Reviews 75 (2007) 114.
[28] Coral 1.4, http://www.insilico.eu/coral.
[29] ACD/ChemSketch Freeware, version 12.01, Advanced Chemistry Development, Inc., Toronto, ON, Canada, 2009www.acdlabs.com.
[30] A.A. Toropov, E. Benfenati, Current Drug Discovery Technology 4 (2007) 77.
[31] A.A. Toropov, E. Benfenati, Bioorganic & Medicinal Chemistry 26 (2008) 4801.
[32] A.A. Toropov, A.P. Toropova, E. Benfenati, Chemical Biology & Drug Design 73 (2009) 515.
[33] A.A. Toropov, A.P. Toropova, E. Benfenati, D. Leszczynska, J. Leszczynski, European Journal of Medicinal Chemistry 45 (2010) 1387.
[34] S. Wold, L. Eriksson, in: H. van de Waterbeemd (Ed.), Chemometrics Methods in Molecular Design, Weinheim, VCH, 1995, pp. 309–318.
[35] D.M. Hawkins, S.C. Basak, D. Mills, Journal of Chemical Information and Modeling 43 (2003) 579.
[36] A. Golbraikh, A. Tropsha, Journal of Molecular Graphics and Modelling 20 (2002) 269.
[37] N.R. Draper, H. Smith, Applied Regression Analysis, John Wiley & Sons, New York, 1981.

[38] Matlab 5.0, The MathWorks, Inc., Natick, Massachusetts, USA, 2011. http://www.mathworks.com.
[39] P. Bultinck, Computational Medicinal Chemistry for Drug Discovery, CRC Press, Boca Raton, FL, 2004.
[40] B.D. Silverman, Journal of Chemical Information and Modeling 40 (2000) 1470.
[41] V. Consonni, R. Todeschini, M. Pavan, P. Gramatica, Journal of Chemical Information and Modeling 42 (2002) 693.
[42] J. Schuur, P. Selzer, J. Gasteiger, Journal of Chemical Information and Modeling 36 (1996) 334.