9th International Electronic Conference on Synthetic Organic Chemistry. ECSOC-9
1-30 November 2005. http://www.usc.es/congresos/ecsoc/9/ECSOC9.HTM & http://www.mdpi.net/ecsoc/

[G013]

A Novel Approach for Computer-Aided "Rational" Drug Design: Theoretical and Experimental Assessment of a Promising Method for Virtual Screening and *in silico* Design of New Antimalarial Compounds.

Yovani Marrero-Ponce,[§] Maité Iyarreta-Veitía,[†] Alina Montero-Torres,[§] Carlos Romero-Zaldivar,[§] Carlos A. Brandt,[‡] Priscilla E. Ávila,[δ] Karin Kirchgatter,[δ] and Yanetsy Machado.[§]

[§]Department of Pharmacy, Faculty of Chemical-Pharmacy and Department of Drug Design, Chemical Bioactive Center. Central University of Las Villas, Santa Clara, 54830, Villa Clara, Cuba.
[†]Centre d'etudes Pharmaceutiques, CNRS Biocis UMR 8076; Laboratoire de Synthese de composes d'interet biologique, Faculté de pharmacie. Université Paris-Sud 5, rue J.B. Clément, 92296, Châtenay-Malabry Cedex, France.
[‡]Department of Organic Chemistry, Butantan Institute, Av. Vital Brasil 1500, Butantã, São Paulo, SP, 05503-900, Brazil.
[δ]Núcleo de Estudos em Malária, Superintendência de Controle de Endemias (SUCEN), Av. Dr. Eneas de Carvalho Aguiar 470, Cerqueira César, São Paulo, SP, 05403-000, Brazil.

e-mail: **yovanimp@qf.uclv.edu.cu** or **amontero@uclv.edu.cu**

Abstract

Malaria is one of the most significant public health concerns in many tropical and subtropical regions of the world, with 40% of the world population exposed to malaria-causing parasites. Increasing resistance of *Plasmodium spp.* to existing therapies has heightened alarms about malaria in the international health community. Nowadays there is a pressing need to identify and develop new drug-based antimalarial therapies. In an effort to overcome this problem, the main aim of this study was to develop simple linear discriminant-based QSAR models for the classification and prediction of antimalarial activity using some of the *TOMOCOMD-CARDD* fingerprints, so as to enable computational screening from virtual combinatorial datasets. In this sense a database of 1562 organic-chemicals having great structural variability; 597 of them antimalarial agents and 965 compounds having other clinical uses, was analyzed and presented as a helpful tool not only for theoretical chemist but also for other researchers in this area. These series of compounds were processed by a *k*-means cluster analysis in order to design training and predicting sets. Afterward, two linear classification functions were derived toward discrimination between antimalarial and non-antimalarial compounds. The models (including non-stochastic and stochastic indices) classify correctly more than 93% of compounds in both training and external prediction datasets. They showed high Matthews´ correlation coefficients; 0.889 and 0.866 for training and 0.855 and 0.857 for test set. Models predictivity were also assessed and validated by the random removal of 10% of the compounds to form a test set, for which predictions were made from the models. The overall mean of the correct classification for this process (leave-group 10% full-out cross-validation) for obtained equations with non-stochastic and stochastic quadratic fingerprints were 93.93% and 92.77%, correspondingly. The quadratic maps-based *TOMOCOMD-CARDD* approach implemented in this work was successfully compared with four of the most useful models for antimalarials selection reported to date. The models developed with non-stochastic and stochastic quadratic indices were then used in a simulation of a virtual search for Ras FTase inhibitors with antimalarial activity;

70% and 100% of the 10 inhibitors used in this virtual search were correctly classified, showing the ability of the models to identify new lead antimalarials. Finally, these two QSAR models were used in the identification of previously un-known antimalarials compounds. In this sense, three synthetic intermediaries of quinolinic compounds were evaluated as active/inactive ones using the developed models. The synthesis and biological evaluation of these chemicals against two Malaria strains, using Chloroquine as reference, was performed. An accuracy of 100% with the theoretical predictions was observed. The compound **3** shown antimalarial activity, being the first report of an arylaminomethylenemalonate having such activity. This result opens a door to a virtual study considering a higher variability of the central core already evaluated, as well as other chemicals not included in this family. We conclude that the approach described here seems to be a promising QSAR tool for molecular discovery of novel classes of antimalarial drugs which may meet the dual challenges posed by drug-resistant parasites and the rapid progression of malaria illness.

1. BACKGROUND

Malaria remains one of the most serious health threats in the world, affecting 300-400 million people and claiming ca. 3 million lives each year.[1,2] Due to the increasing prevalence of multidrug resistant of malaria parasites to standard chemotherapy, the discovery and use of nontraditional antimalarials with novel modes of action is becoming widespread.[3-5] Knowing the complexity and cost of the process of drug discovery, the use of “rational” search methodologies is recommended. Consequently, medicinal chemists are called to developing more efficient strategies for the search of novel candidates to be assayed as antimalarial drugs. In this sense, computer-aided drug design approach emerges as a promising solution to this problematic.[6-9] One of the major goals of such design strategy is the identification from large databases or libraries, of structural subsystems responsible for a specific biological activity. Using computational approaches based on discrimination functions, it is possible to classify active compounds from inactive ones and to predict, using clustering and similarity searching, the biological activity of new lead compounds.[10-14]

In this context, our research group has recently introduced a novel scheme to perform rational *–in silico-* molecular designs (or selection/identification of lead drug-like chemicals) and QSAR/QSPR studies, known as ***TOMOCOMD-CARDD*** (acronym of ***TO***pological ***MO***lecular ***COM***puter ***D***esign-***C***omputer ***A***ided “***R***ational” ***D***rug ***D***esign).[15] This method has been developed to generate molecular fingerprints based on the application of the discrete mathematics and linear algebra theory to chemistry. In this sense, atom, atom-type and total quadratic and linear molecular fingerprints have been defined in analogy to the quadratic and linear mathematical maps.[16,17] This *-in silico-* method has been successfully applied to the prediction of several physical, physicochemical and chemical properties of organic compounds.[16-19] In addition, *TOMOCOMD-CARDD* has been extended to consider three-dimensional features of small/medium-sized molecules based on the trigonometric 3D-chirality correction factor approach.[20]

A later paper allowed the description of the significance-interpretation and the comparison to other molecular descriptors.[17,18] The approach describes changes in the electronic distribution with the time throughout the molecular backbone. Specifically, the features of the k^{th} total and local quadratic and linear indices were illustrated by examples of various types of molecular structures, including chain length and branching as well as content of heteroatoms, and multiple bonds.[17,18] Additionally, the linear independence of the atom-type quadratic and linear fingerprints to other 229 0D-3D "DRAGON" molecular descriptors was demonstrated. In this sense was concluded that local *TOMOCOMD-CARDD* fingerprints are independent indices which contain important structural information to be used in QSPR/QSAR and drug design studies.[17,18]

The prediction of the pharmacokinetical properties of organic compounds is a problem that can also be addressed using this approach. In this sense, this method has been used to estimate the intestinal–epithelial transport of drug in human adenocarcinoma of colon cell line type 2 (Caco-2) culture of a heterogeneous series of drug-like compounds.[21-23] The obtained results suggested that the *TOMOCOMD-CARDD* method was able of predicting the permeability values and it proved to be a good tool for studying the oral absorption of drug candidates during the drug development process.

The *TOMOCOMD-CARDD* strategy has also been useful for the selection of novel subsystems of compounds having a desired property/activity. In this sense, it was successfully applied to the virtual (computational) screening of novel anthelmintic compounds, which were then synthesized and *in vivo* evaluated on *F. Hepatica*.[24,25]

Studies for the fast-track discovery of novel paramphistomicides, antimalarial and antibacterial compounds were also conducted with this theoretical approach.[26-29]

Later, promising results have been found in the modeling of the interaction between drugs and HIV Ψ-RNA packaging-region in the field of bioinformatics using the *TOMOCOMD-**CANAR*** (***C**omputed-**A**ided **N**ucleic **A**cid **R**esearch) approach.[30,31] Finally, an alternative formulation of our approach for structural characterization of proteins was carried out recently.[32,33] This extended method [*TOMOCOMD-**CAMPS*** (***C**omputed-**A**ided **M**odelling in **P**rotein **S**cience)] was used to encompass protein stability studies – specifically how alanine substitution mutation on Arc repressor wild-type protein affects protein stability– by means of a combination of protein linear or quadratic indices (macromolecular fingerprints) and statistical (linear and non-linear model) methods.[32,33]

In the present work, *TOMOCOMD-CARDD* strategy is used to find quantitative models which allow the discrimination of antimalarial compounds from inactive ones in a rational way using non-stochastic and stochastic quadratic indices. A virtual screening for the search of new leads compounds with a novel action mechanism is performed for the case of Ras FTPase inhibitors with antimalarial activity. Finally, we present the design, synthesis and *in vitro* evaluation against two Plasmodium falciparum strains of synthetic intermediates of quinolinic compounds, as starting point for the development of new non-expensive antimalarials.

2. THEORETICAL FRAMEWORK

The theoretical scaffold of the *TOMOCOMD-CARDD's* molecular descriptors family was split into two parts; one for describing the mathematical features of non-stochastic fingerprints and the other one related with the stochastic quadratic indices.

2.1. Non-Stochastic Quadratic Fingerprints

Implemented in the subprogram *CARDD* of the *TOMOCOMD* software, the atom, atom-type and total non-stochastic quadratic fingerprints can be calculated from both, molecular pseudograph's atom adjacency matrix and molecular vector of small-to-medium-sized organic compounds. The general principles of these quadratic indices have been explained in some detail elsewhere.[14,18,20-23,25,26] However; an overview of this approach will be given.

For a given molecule composed of n atoms, the "molecular vector" (X) is constructed and the k^{th} total quadratic indices, $\boldsymbol{q_k}(x)$ are calculated as quadratic forms as shown in Eq. 1,

$$q_k(x) = \sum_{i=1}^{n} \sum_{j=1}^{n} {}^k a_{ij} x_i x_j \qquad \textbf{(1)}$$

where, n is the number of atoms in the molecule and $x_1,\ldots,x_n$ are the coordinates or components of the "molecular vector" (X) in a system of canonical basis vectors of $\Re^n$. The components of the molecular vector are numerical values, which can be considered as weights (atom-labels) for the vertices of the pseudograph. Certain atomic properties (electronegativity, atomic radii, etc) can be used with this propose. In this work, the Pauling electronegativities are selected as atom weights.[34]

The coefficients ${}^k a_{ij}$ are the elements of the k^{th} power of the symmetrical square matrix **M**(G) of the molecular pseudograph (G), and are defined as follows:

$$\begin{aligned} a_{ij} &= P_{ij} \text{ if } i \neq j \text{ and } \exists\ e_k \in \text{E (G)} \\ &= L_{ii} \text{ if } i = j \\ &= 0 \text{ otherwise} \end{aligned} \qquad \textbf{(2)}$$

where E(G) represents the set of edges of G. P_{ij} is the number of edges (bonds) between vertices (atoms) v_i and v_j, and L_{ii} is the number of loops in v_i.

Equation (**1**) for $q_k(x)$ can be written as the single matrix equation:

$$\boldsymbol{q_k}(\mathbf{x}) = \mathbf{X}^t \mathbf{M}^{\boldsymbol{k}} \mathbf{X} \qquad \textbf{(3)}$$

where **X** is a column vector (a $nx1$ matrix), $\mathbf{X}^t$ the transpose of **X** (a $1xn$ matrix) and $\mathbf{M}^{\boldsymbol{k}}$ the k^{th} power of the matrix **M** of the molecular pseudograph G (mathematical quadratic form's matrix).

In addition to total quadratic indices, computed for the whole-molecule, a local-fragment (atom and atom-type) formalisms can be developed. These descriptors are termed local quadratic indices, $\boldsymbol{q_{kL}}(x)$.[14,18,20-23,25,26] The definition of these descriptors is as follows:

$$q_{kL}(x) = \sum_{i=1}^{m} \sum_{j=1}^{m} {}^k a_{ijL} x_i x_j \qquad \textbf{(4)}$$

where, m is the number of atoms of the fragment of interest and ${}^k\boldsymbol{a}_{ijL}$ is the element of the row "i" and column "j" of the matrix $\mathbf{M}^{\boldsymbol{k}}_{\text{L}}$. This matrix is extracted from the $\mathbf{M}^{\boldsymbol{k}}$ matrix and contains information referred to the vertices (atoms) of the specific molecular fragments and also of the molecular environment. The matrix $\mathbf{M}^{\boldsymbol{k}}_{\text{L}} = [{}^k a_{ijL}]$ with elements ${}^k a_{ijL}$ is defined as follows:

$$\begin{aligned} {}^k a_{ijL} &= {}^k a_{ij} \textit{ if } \text{both } v_i \text{ and } v_j \text{ are atoms contained within the molecular fragment} \\ &= {}^1/_2\, {}^k a_{ij} \textit{ if } v_i \text{ or } v_j \text{ is an atom contained within the molecular fragment but not both} \\ &= 0 \text{ otherwise} \end{aligned} \qquad \textbf{(5)}$$

These local analogues can also be expressed in matrix form by the expression:

$$\boldsymbol{q_{kL}}(x) = \mathbf{X}^t\, \mathbf{M}^k_{\mathbf{L}}\, \mathbf{X} \qquad \textbf{(6)}$$

Notice that the above scheme follows the spirit of a Mulliken population analysis.[35] Also note that for every partitioning of a molecule into Z molecular fragment there will be Z local molecular fragment matrices. In this case, if a molecule is partitioned into Z molecular fragments, the matrix $\mathbf{M}^{\boldsymbol{k}}$ can be partitioned into Z local matrices $\mathbf{M}^{\boldsymbol{k}}_{\mathbf{L}}$**,** L = 1,... Z, and the k^{th} power of matrix **M** is exactly the sum of the k^{th} power of the local Z matrices. In this way, the total quadratic indices are the sum of the quadratic indices of the Z molecular fragments:

$$\boldsymbol{q}_{\boldsymbol{k}}(x) = \sum_{L=1}^{Z} q_{kL}(x) \qquad \textbf{(7)}$$

Atom and atom-type quadratic fingerprints are specific cases of local quadratic indices. In this sense, the k^{th} atom-type quadratic indices are calculated by adding the k^{th} atom quadratic indices for all atoms of the same type in the molecule.

In the atom-type quadratic indices formalism, each atom in the molecule is classified into an atom-type (fragment), such as heteroatoms, hydrogen bonding (H-bonding) to heteroatoms (O, N and S), halogen atoms, aliphatic carbon chain, aromatic atoms (aromatic rings), an so on. For all data sets, including those with a common molecular scaffold as well as those with diverse structure, the k^{th} atom-type quadratic indices provide important information.

2.2. Atom, Atom-type, and Total Stochastic Quadratic Fingerprints

Notice that the mathematical quadratic form's matrices, $\mathbf{M}^{\boldsymbol{k}}$, are graph-theoretical electronic-structure models, like the "extended Hückel" model. The $\mathbf{M}^1$ matrix considers all valence-bond electrons (σ- and π-networks) in one step, and their power k (k = 0, 1, 2, 3…) can be considered as an interacting-electronic chemical-network model in steps k. This model can be seen as an intermediate one between the quantitative quantum-mechanical Schrödinger equation and classical chemical bonding ideas.[38]

Recently, our research group has also developed a new method based on the Markov chain theory, which has been successfully employed in QSPR and QSAR studies.[13,37,39] This approach also describes changes in the electron (stochastic) distribution and vibrational decay with time throughout the molecular backbone using Markov chain formalism.

The present approach is based on a simple model for the intramolecular (stochastic) movement of all valence-bond electrons. Let us consider a hypothetical situation in which a set of atoms is free in space at an arbitrary initial time (t_0). In this time, the electrons are distributed around atomic nuclei. Alternatively, these electrons can be distributed around cores in discrete intervals of time t_k. In this sense, the electron at an arbitrary atom i can move to other atoms at different discrete time periods t_k (k = 0, 1, 2, 3…) throughout the chemical-bonding network.

The k^{th} stochastic molecular pseudograph's atom adjacency matrix [$\mathbf{S}^k$(G)] can be obtained from $\mathbf{M}^k$. Here, $\mathbf{S}^k(\text{G}) = \mathbf{S}^k = [{}^k\mathbf{s}_{ij}]$ is a squared table of order n (n = number of atoms), and the elements ${}^k\mathbf{s}_{ij}$ are defined as follows:

$${}^k s_{ij} = \frac{{}^k a_{ij}}{{}^k SUM_i} = \frac{{}^k a_{ij}}{{}^k \delta_i} \qquad \textbf{(9)}$$

where ${}^k a_{ij}$ are the elements of the k^{th} power of **M,** and the SUM of the ith row of $\mathbf{M}^k$ are named the k-order vertex degree of atom i, ${}^k\delta_i$. The k^{th} $\mathbf{s}_{ij}$ elements are the transition probabilities with which the electrons moving from atom i to j in the discrete time period t_k (step-by-step). Notice that the k^{th} elements $\boldsymbol{s}_{ij}$ takes into consideration the information of the molecular topology in step *k* throughout the chemical-bonding (σ- and π-) network. For instance, the ${}^{2}\mathbf{s}_{ij}$ values can distinguish between hybrid states of atoms in bonds. In this sense, it can clearly be seen from Table 1 that electrons will have a higher probability of returning to the sp N atom $p(N_{10}) = 0.75$ than to the sp^2 N atom $p(N_6) = 0.33$ in t_2. A similar behavior can be observed among the different hybrid states of the C atoms in the molecule of 2-formyl-6-methyl-benzonitrile (see Table 1): Csp^3 [$p(C_{11}) = 0.25$]; Csp^2 [$p(C_2) = 0.625$]; Csp$^2_{\text{arom}}$ [$p(C_3) = 0.285$, $p(C_4) = 0.3$, $p(C_5) = 0.33$, $p(C_7) = 0.33$, $p(C_8) = 0.25$]; and Csp [$p(C_9) = 0.769$]. This is a logical result as the electronegativity scale of these hybrid states is taken into account. The *k*th total and local stochastic quadratic indices, ${}^{s}\boldsymbol{q}_{k}(x)$ are calculated in the same way that the non-stochastic quadratic indices, but using the *k*th stochastic molecular pseudograph's atom adjacency matrix, $\mathbf{S}^{k}(G)$, as mathematical quadratic forms' matrices.

3. MATERIALS AND METHODS

3.1. Computational Methods: *TOMOCOMD-CARDD* Approach

TOMOCOMD is an interactive program for molecular design and bioinformatic research.[15] It consists of four subprograms: (CARDD:Computed-Aided 'Rational' Drug Design, CAMPS:Computed-Aided Modeling in Protein Science, CANAR:Computed-Aided Nucleic Acid Research and CABPD:Computed-Aided Bio-Polymers Docking). Each one of them allows drawing the structures (drawing mode) and calculating molecular 2D/3D (calculation mode) atom- and bond-based descriptors. In the present report, we outline salient features concerned with only the subprogram CARDD.

The main steps for the application of this method in QSAR/QSPR and drug design can be briefly summarized as follows:

1. Drowning the molecular pseudographs for each molecule of the data set, using the drawing mode. This procedure is performed by a selection of the active atomic symbol belonging to the different groups in the periodic table of the elements,
2. Use of appropriate weights in order to differentiate the molecular atoms,
3. Compute the total and local (atom and atom-type) quadratic indices of the molecular pseudograph's atom adjacency matrix. They can be carried out in the software calculation mode, where you can select the atomic properties and the family descriptor previously to calculate the molecular indices. This software generates a table in which the rows correspond to the compounds, and columns correspond to the total and local quadratic indices or other family of molecular descriptors implemented in this program,
4. Development of a QSPR/QSAR equation by using several multivariate analytical techniques, such as multilinear regression analysis (MRA), neural networks (NN), linear discrimination analysis (LDA), and so on. In this sense it is possible to find a quantitative relation between an activity **A** and the quadratic fingerprints having, for instance, the following appearance

   $$\mathbf{A} = a_0\boldsymbol{q}_0(x) + a_1\boldsymbol{q}_1(x) + a_2\boldsymbol{q}_2(x) + \ldots + a_k\boldsymbol{q}_k(x) + c \qquad \mathbf{(10)}$$

   where **A** is the measured activity, $\boldsymbol{q}_k(x)$ are the *k*th total quadratic indices, and the $\boldsymbol{a}_k$*'s* are the coefficients obtained by the linear regression analysis.
5. Test of the robustness and predictive power of the QSPR/QSAR equation by using internal (leave-*one*-out and leave-group-out cross-validation) and external (using a test set and an external predicting set) validation techniques.

The following descriptors were calculated in this work:

i) $\boldsymbol{q_k}(x)$ and $\boldsymbol{q_k}^{\mathrm{H}}(x)$ are the k^{th} total quadratic indices not considering and considering H-atoms in the molecular pseudograph (G), respectively.

ii) $\boldsymbol{q}_{\boldsymbol{k}\mathrm{L}}(x_{\mathrm{E}})$ and $\boldsymbol{q}_{\boldsymbol{k}\mathrm{L}}^{\mathrm{H}}(x_{\mathrm{E}})$ are the k^{th} local (atom-type = heteroatoms: S, N, O) quadratic indices not considering and considering H-atoms in the molecular pseudograph (G), correspondingly. These local descriptors are putative H-bonding acceptors.

iii) $\boldsymbol{q}_{\boldsymbol{k}\mathrm{L}}^{\mathrm{H}}(x_{\mathrm{E-H}})$ are the k^{th} local (atom-type = H-atoms bonding to heteroatoms: S, N, O) quadratic indices considering H-atoms in the molecular pseudograph (G). These local descriptors are putative H-bonding donors.

The k^{th} stochastic total [$^{\mathrm{s}}\boldsymbol{q_k}(x)$ and $^{\mathrm{s}}\boldsymbol{q_k}^{\mathrm{H}}(x)$] and local [$^{\mathrm{s}}\boldsymbol{q_k}(x_{\mathrm{E}})$, $^{\mathrm{s}}\boldsymbol{q_k}^{\mathrm{H}}(x_{\mathrm{E}})$ and $^{\mathrm{s}}\boldsymbol{q_k}^{\mathrm{H}}(x_{\mathrm{E-H}})$] quadratic indices were also computed.

3.2. Data Set

It is well known, that the quality of the classification models is highly dependent on the quality of the selected data set. The most critical aspect for constructing the training set is to warrant a great molecular diversity on it. Taking that into account, we selected a large data set of 1562 organic-chemicals having great structural variability; 597 of them are antimalarial agents[2,7-9, 40-82]and the other ones are non-antimalarials[41,82] (965 compounds having other clinical uses, such as antivirals, sedative/hypnotics, diuretics, anticonvulsivants, hemostatics, oral hypoglycemics, antihypertensives, antihelminthics, anticancer compounds and so on. It is clear that the declaration of these compounds as "inactive" antimalarial per se does not guarantee antimalarial side-effects for some of these organic-chemical drugs that have been left undetected so far. This problem can be reflected in the results of classification for the series of inactive chemicals.

On the other hand, the data set of active compounds was selected by considering representatives of most of the different structural patterns and action modes for the case of the antimalarial activity. For instance, it includes: 1) alkaloidal and synthetic quinoline-based antimalarial drugs which involve the blockage of the function of the food vacuole (4- and 8-aminoquinolines,[9,70] peptide derivatives,[52] dimeric quinolines[47,49] and other compounds such as indolo[3,2-*c*]quinolines[7] and methylene blue derivatives), 2) peptide (fluoromethyl ketone peptide derivatives) and nonpeptide (phenothiazines and chalcones) falcipain-cysteine protease inhibitors,[42,45] 3) peptide and nonpeptide inhibitors of malarials aspartyl protease plasmepsin II,[48] 4) agents interfering with *Plasmodium Falciparum* phospholipids metabolism (primary, secondary, tertiary amines and quaternary ammonium and bisammonium salts),[69] 5) antimalarials which have ability to inhibiting electron transport processes and respiratory systems by acting as ubiquinone antagonists (hydroxynaphthoquinones such as atovaquone),[40] 6) selective inhibitors of lactate dehydrogenase from malaria parasite (some derivatives of the sesquiterpene 8-deoyhemigossylic acid),[40] 7) antimalarial chemicals which act by selectively inhibiting malarial dihydrofolate reductase-thymidylate synthase (pyrimethamine and it is analogs),[8] 8) antiparasitic agents affecting DNA topoisomerases (e.g., anticancer acridines)[53] and 9) artemisinin-type antimalarials and other simple¸bicyclic and tetraciclic endoperoxides (incluiding lactone ring-open analogs the trioxane).[2-5, 44, 51, 54-65, 66-68, 72] These antimalarials endoperoxides appears to have a two-step mode of action. In the first step, the 'artesmisinin' compounds are activated by heme or molecular iron to produce free radicals and electrophilic (alkylating) intermediates. In the second step, these reactive species react with and damage specific malarial membrane-associated proteins. Other compounds for which have not been found or defined a specific mode of action, but have been reported as antimalarial agents were also included.[41,50, 82] Figure 1 shows a representative sample of such active compounds.

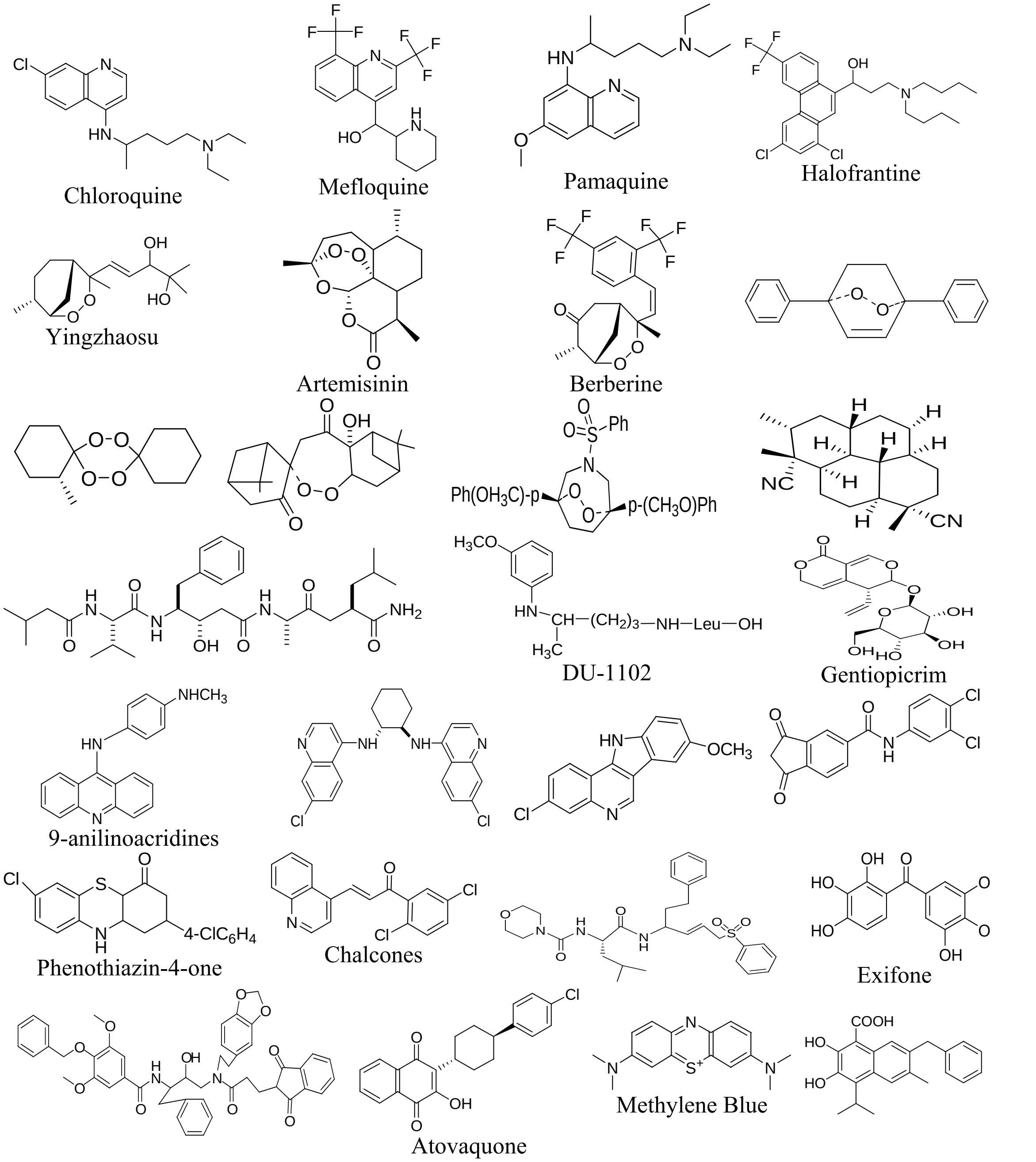


Figure 1. Random, but not exhaustive, sample of the molecular families of antimalarial agents studied here.

Later, two k-means cluster analyses (k-MCA) were performed for active and inactive series of compounds, which permitted us to split the dataset (1562 organic-chemicals) into training and predicting series.[83,84] That is, all cases were processed using k-MCA in order to design training and predicting data series in a "rational" way. The main idea consists in carrying out a partition of either active or inactive series of chemicals in several statistically representative classes of compounds. Thence, one may select from the members of all these classes of training and predicting series. This procedure ensures that any chemical class (as determined by the clusters derived from k-MCA) will be represented in both compounds' series.

3.3. Chemometric Methods

k-means cluster analysis (k-MCA). The statistical software package STATISTICA was used to develop the k-MCA.[85] The number of members in each cluster and the standard deviation of the variables in the cluster (kept as low as possible) were taking into account, to have an acceptable statistical quality of data partition in clusters. We also made an inspection of the standard deviation (SS) between and within clusters, of the respective Fisher ratio and their *p*-level of significance, which was considered to be lower than 0.05.[83,84]

Linear Discriminant Analysis. In spite of several chemometric techniques to find good discriminant functions exist, such as SIMCA or neural networks, we select the linear discriminant analysis (LDA) in order to generate the classifier function on the basis of the simplicity of the method. The use of this statistical analysis will permit to classify new compounds as active or inactive ones from molecular descriptors.

LDA was carried out with the STATISTICA software.[85] The considered tolerance parameter (proportion of variance that is unique to the respective variable) was the default value for minimum acceptable tolerance, which is 0.01. Forward stepwise was fixed as the strategy for variable selection. The principle of parsimony (Occam's razor) was taken into account as strategy for model selection. In connection, we selected the model with a high statistical signification but having as few parameters (a_k) as possible and maximizes the degrees of freedom. In the equation **10**, a_k are the coefficients of the classification function, determined by the least square method as implemented in LDA modulus of STATISTICA.[85]

The quality of the models were determined by examining Wilks' λ parameter (*U*-statistic), square Mahalanobis distance (D^2), Fisher ratio (F) and the corresponding *p*-level (p(F)) as well as the percentage of good classification in the training and test sets. Models with a proportion between the number of cases and variables in the equation lower than 5 were rejected.

The Wilks' λ statistics is helpful to evaluating the total discrimination, and can take values between zero (perfect discrimination) and one (no discrimination). The D^2 indicates the separation of the respective groups.

The biological activity (antibacterial in this case) was codified by a dummy variable "***Class***". This variable indicates the presence of either an active compound (***Class*** = 1) or an inactive compound (***Class*** = –1). The classification of cases was performed by means of the posterior classification probabilities. This is the probability to which the respective case belongs to a particular group (active or inactive) and it is proportional to the

Mahalanobis distance. On completion, the posterior probability is the probability, based on our knowledge of the values of others variables, to which the respective case belongs to a particular group. By using the models, one compound can then be classified as active, if $\Delta P\% > 0$, being $\Delta P\% = [P(\text{Active}) - P(\text{Inactive})]x100$ or as inactive otherwise. P(Active) and P(Inactive) are the probabilities with which the equations classify a compound as active and inactive, respectively.

On the other hand, validation is a crucial aspect of any QSAR/QSPR modeling.[86,87] One of the most popular validation criteria is the leave-*one*-out (LOO) cross-validation method (internal validation). This method systematically removes one data point at a time from the data set. A QSAR/QSPR model is then constructed based on this reduced data set and subsequently used to predict the removed data point. This procedure is repeated until a complete predictions set is obtained. Good results in this experiment can be considered as a proof of the high predictive ability of the models. However, this assumption is generally incorrect and it can be that it exists lack of correlation between the good LOO results and the high predictive ability of QSAR/QSPR models.[86,87] Thus, the good behavior of models in an LOO procedure appears to be the necessary but not the sufficient condition for the models, to have a high predictive power. In this sense, Golbraikh and Tropsha[87] emphasized that the predictive ability of a QSAR/QSPR model can be estimated by using only a test set (external validation) of compounds that were not used for building the model. For this reason, in order to assess the predictability of the obtained model, external validation procedures were carried out. In this sense, the statistical robustness and predictive power of the obtained model was assessed using a prediction (test) set.

In the present work leave-group-out (LGO) cross-validation strategy was carried out.[86] In this case, 10% of the data set was used as group size, i.e. groups including 10% of the training data set are left out and predicted for the model based on the remaining 90%. This process was carried out 10 times on 10 unique subsets. In this way, every observation was predicted once (in its group of left-out observations). The overall mean for this process (10% full leave-out cross-validation) was used as a good indication of robustness and stability of the obtained models.

Finally, the calculation of percentages of global good classification (accuracy), sensibility, specificity (also known as 'hit rate'), false positive rate (also known as 'false alarm rate') and Matthews correlation coefficient (MCC) in the training and test sets permits carrying out the assessment of the model.[88] While the sensitivity is the probability of correctly predicting a positive example, the specificity is the probability that a positive prediction is correct. On the other hand, MCC quantifies the strength of the linear relation between the molecular descriptors and the classifications, and it may often provide a much more balanced evaluation of the prediction than, for instance, the percentages.[88]

Orthogonalization of Descriptors. The orthogonalization process of molecular descriptors was introduced by Randić several years ago as a way to improve the statistical interpretation of the models by using interrelated indices.[89-95] This process is an approach in which molecular descriptors are transformed in such a way that they do not mutually correlate. The main philosophy of this approach is to avoid the exclusion of descriptors on the basis of its collinearity with other variables previously included in the model. Both, the non-orthogonal descriptors and derived orthogonal descriptors, contain the same information. In this sense the same statistical parameters of the QSAR models are obtained.[89-95] It is known that the interrelatedness among the different descriptors can result in highly unstable regression coefficients, which makes it impossible to knowing the relative importance of an index and underestimates the utility of the regression coefficients in a model. However, in some cases strongly interrelated descriptors can enhance the quality of a model because the small fraction of a descriptor which is not reproduced by its strongly interrelated pair can provide positive contributions to the modeling. On the other hand, the coefficient of the QSAR model based on orthogonal descriptors are stable to the inclusion of novel descriptors, which permits to interpret the regression coefficients and evaluated the role of individual fingerprints to the QSAR model.

The Randić method of orthogonalization has been described in detail in several publications.[89-95] Thus, we will give only a general overview here. The first step in orthogonalizing the molecular descriptors included in models is to select the appropriate order of orthogonalization, which in this case is the order in which the variables were selected in the forward stepwise search procedure of the statistical analysis.[95] The first variable (V_1) is taken as the first orthogonal descriptors $^1O(V_1)$, and the second one (V_2) is orthogonalized with respect to it [$^2O(V_2)$] by taking the residual of its correlation with $^1O(V_1)$, which is that part of the descriptors V_2 not reproduced by $^1O(V_1)$. Similarly, from the regression of V_3 versus $^1O(V_1)$, the residual is the part of V_3 that is not reproduced by $^1O(V_1)$ and it is labeled $^1O(V_3)$. The orthogonal descriptor $^3O(V_3)$ is obtained by repeating this process in order to also make it orthogonal to $^2O(V_2)$. The process is repeated until all variables are completely orthogonalized, and the orthogonal variables are then used to obtain the new model.

3.4. Chemistry

IR spectra were recorded with a FTIR-BOMEM spectrometer using KBr disks for solid or NaCl cell for liquids (υ in cm^{-1}). ^1H NMR and ^{13}C NMR spectra were recorded on a Bruker ADPX-300 (300 mHz) using $CDCl_3$ as solvent. The calibration of spectra was carried out on TMS (internal ^1H) and $CDCl_3$ (^{13}C) signals δ ^1H (TMS) = 0; δ ^{13}C ($CDCl_3$) = 77.0. Chloroquine diphosphate was supplied by “Fundação para o Remédio Popular” (Brazil). All solvent were previously dried and purified before use, according to standards established in the literature.[96, 97]

3.5. Determination of *in vitro* Antiplasmodial Activity

In vitro antiplasmodial evaluation was performed by using the susceptibility microtechnique.[98] Two strains of *Plasmodium falciparum*, K1-chloroquine resistant, and Palo Alto-chloroquine sensitive, kindly provided by the WHO Registry of Standard Strains of Malaria Parasites at the University of Edinburgh, were continuously maintained in culture and used in these assays.[99] The parasites freezing and thawing procedures were based on that described.[100] The parasites were cultivated to 5% hematocrit in RPMI 1640 medium with 25 mM HEPES, 21 mM sodium bicarbonate, 370 μM hypoxantine, 40 μg/ml gentamycin, and 10% human A^+ or O^+ serum provided by Fundação Pró-Sangue/Hemocentro de São Paulo. Washed human O^+ erythrocytes were added to the culture as necessary. Synchronization was obtained by treatment with D-sorbitol when the parasites were predominantly in the young trophozoite stage.[101] Stock solutions of the compounds (1 000 pmol/100 μL of ethanol) were used to prepare different concentrations (1, 2, 4, 6, 8, 16, 32 and 100 pmol/well) in aqueous solution. A stock solution of chloroquine diphosphate (1 000 pmol/100 μL in water) was used to prepare a series of concentrations (1, 2, 4, 6, 8, 16 and 32 pmol/well) to check the sensitivity of the isolates. Flat bottomed microtitre plates were dosed adding 100 μl of each concentration/well. The plates were dried at 37°C and stored at 4°C. An aliquot of 100 μl of culture with a parasitemia between 0.5-1.0% and parasites in young trophozoite stage was added to each well of the microtitre plates. A control without compound and a sensitivity test to chloroquine were performed in parallel. Microplates were incubated in a candle jar with a gas mixture of 3% CO_2, 5% O_2, 92% N_2, and maintained at 37°C for 24-36 h. Giemsa-stained thick blood smears were prepared from each well when controls showed presence of schizonts by optical microscopy. The number of schizonts was counted per 200 asexual parasites and the tests were considered valid when this number was equal or superior to 10%. The minimum inhibitory concentration (MIC) of each compound was defined by the lowest concentration that completely inhibited the schizont maturation.

4. RESULTS AND DISCUSSION

4.1. Training and test sets design through k-means cluster analysis

The first step in this study was the design of the training and predicting series to prevent non-random distribution of chemicals between the two sets. This was achieved using k-MCA.[83,84] This "rational" design of training and predicting series allowed us to design both sets that are representative of the entire "experimental universe".

We carried out first a k-MCA with active compounds and afterwards with inactive ones. A first k-MCA (I) split antimalarials in 20 clusters with 33, 18, 29, 29, 21, 59, 46, 57, 37, 16, 9, 35, 24, 55, 17, 22, 25, 34, 13, and 18 members. On other hand, the inactive compound series was also partitioned into 20 clusters (k-MCA II) with 58, 26, 78, 26, 48, 64, 60, 53, 80, 72, 46, 64, 41, 68, 58, 25, 4, 22, 23, and 49 members.

Then, selection of the training and prediction sets was performed by taking, in a random way, compounds belonging to each cluster. From these 1562 compounds, 1120 were chosen at random to forming the training set, being 437 of them actives and 683 inactive ones. The great structural variability of the selected training data set makes it possible, not only the discovery of lead compounds with determined mechanisms of antimalarial activity, but also with novel modes of action. It will be well-illustrated in this paper in a virtual experiment for lead generation.

The remaining subseries composed of 160 antimalarials and 282 compounds with different biological properties were prepared as test sets for the external cross-validation of the models. These compounds were never used in the development of the classification models. Figure 2 graphically illustrates the above-described procedure where two independent cluster analyses (one for active and the other for inactive chemiclas) were performed, to select a representative sample for the training and test sets.

The *k*th total and atom-type non-stochastic quadratic indices were used, with all variables showing *p*-levels of <0.05 for the Fisher test. From the k-MCA, it can be concluded that the structural diversity of several up-to-date known antimalarials (as codified by

TOMOCOMD-CARDD descriptors) may be described at least by 20 statistically homogeneous clusters of chemicals.

4.2. Developing Classification Functions

The use of linear discriminant analysis (LDA) in rational drug design has been extensively used by different authors.[10-14] Being the key of the present study, we developed two classification functions using topological descriptors computed with the *TOMOCOMD-CARDD* software.[15] These linear models are given below together with statistical parameters:

$$\textbf{\textit{Class}} = -10.059 -0.08844\boldsymbol{q}_0(x) +0.07085\boldsymbol{q}_1(x) +0.18907\boldsymbol{q}_0^{\mathrm{H}}(x) -0.0256\boldsymbol{q}_2^{\mathrm{H}}(x) \\ +0.0528\boldsymbol{q}_{2\mathrm{L}}(x_{\mathrm{E}}) +0.19849\boldsymbol{q}_{1\mathrm{L}}^{\mathrm{H}}(x_{\mathrm{E}}) -0.09913\boldsymbol{q}_{2\mathrm{L}}^{\mathrm{H}}(x_{\mathrm{E}}) -0.19816\,\boldsymbol{q}_{1\mathrm{L}}(x_{\mathrm{E\text{-}H}}) \\ +2.658x10^{-8}\boldsymbol{q}_{15\mathrm{L}}(x_{\mathrm{E\text{-}H}}) \qquad (\mathbf{11})$$

$N = 1120 \quad \lambda = 0.32 \quad D^2 = 8.8 \quad F(9, 1110) = 258.32 \quad p<0.0001$

$$\textbf{\textit{Class}} = -8.7734 +0.7734{}^{\mathbf{s}}\boldsymbol{q}_0^{\mathrm{H}}(x) +0.84022{}^{\mathbf{s}}\boldsymbol{q}_1^{\mathrm{H}}(x) -1.20567{}^{\mathbf{s}}\boldsymbol{q}_2^{\mathrm{H}}(x) \\ +0.29627{}^{\mathbf{s}}\boldsymbol{q}_{1\mathrm{L}}^{\mathrm{H}}(x_{\mathrm{E}}) -0.3805{}^{\mathbf{s}}\boldsymbol{q}_3^{\mathrm{H}}(x) -0.1833{}^{\mathbf{s}}\boldsymbol{q}_{1\mathrm{L}}(x_{\mathrm{E}}) +2.3858{}^{\mathbf{s}}\boldsymbol{q}_0(x_{\mathrm{E\text{-}H}}) -1.0558{}^{\mathbf{s}}\boldsymbol{q}_{1\mathrm{L}}(x_{\mathrm{E\text{-}H}}) \\ -1.1887\,{}^{\mathbf{s}}\boldsymbol{q}_{2\mathrm{L}}(x_{\mathrm{E\text{-}H}}) -0.7662{}^{\mathbf{s}}\boldsymbol{q}_{3\mathrm{L}}(x_{\mathrm{E\text{-}H}}) \qquad (\mathbf{12})$$

$N = 1120 \quad \lambda = 0.35 \quad D^2 = 7.7 \quad F(10,1109) = 203.11 \quad p<0.0001$

where N is the number of compounds, λ is Wilks' statistics, D^2 is the squares of Mahalanobis distances, F is the Fisher ratio and *p* is the signification level.

Model **11**, which includes non-stochastic indices, classified correctly 94.73% of the compounds in the training dataset, misclassifying only 59 compounds of a total of 1120. The percentage of false actives in this data set was only 3.66%, i.e. 25 inactive compounds were classified as actives from 683 cases. Conversely, 34 compounds from the group of 437 actives were misclassified as inactive ones (7.78% of misclassification).

The statistical analysis of model **12** showed similar results. In this case, the overall accuracy of the model was 93.13%. Only 4.98 % of misclassification for the inactive group was observed (34 inactive compounds were classified as active ones from a total of 683). In this case 43 compounds from 437 (9.84%) were false inactives.

The classification of all compounds in the complete training dataset provides some assessment of the goodness of fit of the models, but it does not provide a thorough criterion of how the models can predict the biological properties of new compounds. To assess such predictive power, the use of an external test set is essential. In this sense, the activity of the compounds in such set was predicted with the two obtained discrimination functions. The overall accuracy for this group was 93.89% (27/442) and 93.44% (29/442) using model **11** and **12**, respectively. Taking into account the number of compounds used in the external test set, we can see that the model **11** classifies correctly 97.16 % (274/282) of the inactives and 88.13% (141/160) of the actives, while model **12** classifies correctly 95.74 % (270/282) of the inactives and 89.38% (143/160) of the considered antimalarials. It can be seen that the number of misclassified inactive compounds is relative low for both models. This is a desirable condition to consider a model as adequate, taking into account that this number represents inactive compounds that will be sent to biological assays and in this way, loss of time and resources.[12]

The results of global classification of compounds, in both training and external prediction sets, are shown in Table 1. This table also lists most parameters commonly parameters used in medical statistics (accuracy, sensitivity, specificity and false positive rate) and the Matthews correlation coefficient (MCC) for both obtained models.[88] These models, Eqs.

11 and **12**, showed a high MCC of 0.89 (0.87) and 0.86 (0.86) in training (test) sets, correspondingly.

Table 1. Global Results of the Classification of Compounds in the Training and Test Sets.

	Matthews Corr. Coefficient	**Accuracy 'Q_{Total}' (%)**	**Sensitivity 'hit rate' (%)**	**Specificity (%)**	**False positive rate 'false alarm rate' (%)**
non-stochastic descriptors [Eq. (11)]					
Training set	0.889	94.73	92.2	94.2	3.6
Test set	0.866	93.89	88.1	94.6	2.8
stochastic descriptors [Eq. (12)]					
Training set	0.855	93.13	90.2	92.1	4.9
Test set	0.857	93.44	89.4	92.3	4.2

A second experiment, considering a leave-group-out (LGO) strategy, was carried out for both models as internal validation procedure.[86] The overall mean of the correct classification for this process for Eq. **11** and Eq. **12** were 93.93% and 92.77%, respectively. For a 10% full leave-out cross-validation procedure, this level of cross-validated classification is a good indication of robustness and stability of the obtained models. The results of the LGO procedure are shown in Table 2.

Table 2. Predictivity based on the Use of Ten Randomly Selected Subsets (LGO cross-validation) of LDA Models.

Group	**% Global Good Classification**	
	Eq. 11	**Eq. 12**
1	96.43	91.97
2	95.54	94.64
3	83.93	83.93
4	91.07	92.86
5	96.43	97.32
6	93.75	92.86
7	97.32	95.54
8	98.21	99.11
9	96.43	92.86
10	90.18	86.60
Overall mean	**93.93**	**92.77**
Standard Deviation	**4.39**	**4.58**

In summary, the calculation of percentages of good classification in the training and external data sets, and an internal cross-validation procedure permitted us to carry out the assessment of the models.

A close inspection of the molecular descriptors included in both LDA-based QSAR models showed that several of these fingerprints are strongly interrelated to each other.

In Table 3 we resume the results of the orthogonalization of molecular descriptors included in both models. In this case, the equations **11a** and **12a** correspond to the final models with the orthogonalized molecular indices (see Table 8). Here, we used the symbols ${}^{m}O(\boldsymbol{q}_k(x))$, where the superscript m expresses the order of importance of the variable $(\boldsymbol{q}_k(x))$ after a preliminary forward stepwise analysis and O means orthogonal.

Table 3. Results of Randić's Orthogonalization Analysis.

		Orthogonal atom, atom-type and total non-stochastic quadratic indices						
${}^{1}O(\boldsymbol{q}_0{}^{H}(x))$	${}^{2}O(\boldsymbol{q}_2{}^{H}(x))$	${}^{3}O(\boldsymbol{q}_{15L}(x_{E-H}))$	${}^{4}O(\boldsymbol{q}_{1L}(x_{E-H}))$	${}^{5}O(\boldsymbol{q}_1(x))$	${}^{6}O(\boldsymbol{q}_0(x))$	${}^{7}O(\boldsymbol{q}_{2L}{}^{H}(x_E))$	${}^{8}O(\boldsymbol{q}_{2L}(x_E))$	${}^{9}O(\boldsymbol{q}_{1L}{}^{H}(x_E))$
1	0	0	0	0	0	0	0	0
	1	0	0	0	0	0	0	0
		1	0	0	0	0	0	0
			1	0	0	0	0	0
				1	0	0	0	0
					1	0	0	0
						1	0	0
							1	0
								1

LDA-based model derived with orthogonal atom, atom-type and total non-stochastic quadratic indices

$$\textbf{\textit{Class}} = -0.15069 +4.7535\ {}^{1}O(\boldsymbol{q}_0{}^{H}(x)) -3.80426\ {}^{2}O(\boldsymbol{q}_2{}^{H}(x)) +1.17955\ {}^{3}O(\boldsymbol{q}_{15L}(x_{E-H})) -2.36650{}^{4}O(\boldsymbol{q}_{1L}(x_{E-H})) +6.22277\ {}^{5}O(\boldsymbol{q}_1(x)) -15.73721\ {}^{6}O(\boldsymbol{q}_0(x)) -0.97037\ {}^{7}O(\boldsymbol{q}_{2L}{}^{H}(x_E)) +11.35210\ {}^{8}O(\boldsymbol{q}_{2L}(x_E)) +12.44961\ {}^{9}O(\boldsymbol{q}_{1L}{}^{H}(x_E)) \qquad \textbf{(11.a)}$$

N = 1120 $\lambda = 0.32$ $D^2 = 3.9$ F = 258.32   MCC = 0.89   Accuracy (%) = 94.73   %(+) = 96.34   %(-) = 92.22 $p<0.0001$

		Orthogonal atom, atom-type and total stochastic quadratic indices							
${}^{1}O({}^{s}\boldsymbol{q}_0{}^{H}(x))$	${}^{2}O({}^{s}\boldsymbol{q}_2{}^{H}(x))$	${}^{3}O({}^{s}\boldsymbol{q}_1{}^{H}(x))$	${}^{4}O({}^{s}\boldsymbol{q}_{1L}(x_{E-H}))$	${}^{5}O({}^{s}\boldsymbol{q}_{1L}{}^{H}(x_E))$	${}^{6}O({}^{s}\boldsymbol{q}_{1L}(x_E))$	${}^{7}O({}^{s}\boldsymbol{q}_3{}^{H}(x_{E-H}))$	${}^{8}O({}^{s}\boldsymbol{q}_0(x_{E-H}))$	${}^{9}O({}^{s}\boldsymbol{q}_{3l}(x_{E-H}))$	${}^{10}O({}^{s}\boldsymbol{q}_{2L}(x_{E-H}))$
1	0	0	0	0	0	0	0	0	0
	1	0	0	0	0	0	0	0	0
		1	0	0	0	0	0	0	0
			1	0	0	0	0	0	0
				1	0	0	0	0	0
					1	0	0	0	0
						1	0	0	0
							1	0	0
								1	0
									1

LDA-based model derived with orthogonal atom, atom-type and total stochastic quadratic indices

$$\textbf{\textit{Class}} = -0.5589 +3.9445\ {}^{1}O({}^{s}\boldsymbol{q}_0{}^{H}(x)) -54.2074\ {}^{2}O({}^{s}\boldsymbol{q}_2{}^{H}(x)) +40.3252\ {}^{3}O({}^{s}\boldsymbol{q}_1{}^{H}(x)) -0.8304\ {}^{4}O({}^{s}\boldsymbol{q}_{1L}(x_{E-H})) +1.7579\ {}^{5}O({}^{s}\boldsymbol{q}_{1L}{}^{H}(x_E)) -4.7052\ {}^{6}O({}^{s}\boldsymbol{q}_{1L}(x_E)) -34.7293\ {}^{7}O({}^{s}\boldsymbol{q}_3{}^{H}(x)) +3.9482\ {}^{8}O({}^{s}\boldsymbol{q}_0(x_{E-H}) -5.7690\ {}^{9}O({}^{s}\boldsymbol{q}_{3L}(x_{E-H}) -8.4606\ {}^{10}O({}^{s}\boldsymbol{q}_{2L}(x_{E-H}) \qquad \textbf{(12.a)}$$

N = 1120 $\lambda = 0.35$ $D^2 = 7.68$ F= 203.11   MCC= 0.86   Accuracy (%) = 93.13   %(+) = 90.16   %(-)= 95.02 $p<0.0001$

It must be highlighted here that the orthogonal descriptor-based models coincides with the collinear (i.e. ordinary) *TOMOCOMD-CARDD* descriptors-based models in all statistical parameter. That is to say, the statistical coefficients of LDA-QSARs λ, F, MCC, accuracy, %(+) [good classifications in the active group] and %(-) [good classifications in the inactive group] are the same whether we use a set of non-orthogonal descriptors or the corresponding set of orthogonal indices. This is not surprising because the latter are derived as a combination of the former and cannot have more information content than the former.[89-91] Only the D^2 values were different in both equation sets. This is because before carrying out the orthogonalization process, all the variables were standardized. In standardization, all values of selected variables (molecular descriptors) were replaced by standardized values, which are computed as follows: Std. score = (raw score - mean)/Std. deviation. LDA algorithms at one point need to assess the distances between group's centroids (or between cases and centroids), and obviously, when computing D^2 distances, LDA need to decide on a scale. Because the different molecular fingerprints included here used entirely "different types of scales", the data were standardized so that each variable has a mean 0 and a standard deviation of 1. This fact also makes interpretation of the coefficients, in the LDA-QSAR equations, possible. Therefore, ${}^{m}O(\boldsymbol{q}_k(x))$ may be classified according to the distance k into short- (0-5), mid- (6-10), and long-range non-stochastic and stochastic quadratic indices. The information in Table 8 clearly shows that the major contribution to antimalarial activity is providing by short-range *TOMOCOMD-CARDD* descriptors.

4.3. Comparative Analysis of the Obtained Structure-Based Classification Models for Describing the Antimalarial Activity of a Heterogeneous Series of Compounds

In a previous paper, some of the present authors reported two classification models of antimalarial activity using the same training data set, but including non-stochastic and stochastic linear indices.[27] With the aim to evaluate comparatively the ability of the non-stochastic and stochastic quadratic indices to encode chemical information and the quality of the obtained LDA-based classification models, we performed an examination of some statistical parameters. Table 4 summarizes the mains results achieved with both *TOMOCOMD-CARDD* descriptors (based on both quadratic and linear maps).

Table 4. Comparative Analysis of the Obtained Structure-Based Classification Models for Describing the Antimalarial Activity of a Heterogeneous Series of Compounds.

Models' features to be compared[a]	***Structure-Based Classification Models of Antimalarial Activity***					
	Eq. 11	**Eq. 12**	**Eq. 13**	**Eq. 14**	**Eq. 15**	**Eq. 16**
N total	1562	1562	1562	1562	59	60
N antimalarials	597	597	597	597	25	25
Technique [b]	LDA	LDA	LDA	LDA	LDA	LDA
Wilks'λ (U-statistics)	0.32	0.35	0.35	0.38	0.55	0.35
F	258.32	203.11	261.61	202.73	9.83	8.88
D^2	8.8	7.7	7.92	6.90	-	-
p-level	<0.0001	<0.0001	<0.0001	<0.0001	-	-
Training set						
N total	1120	1120	1120	1120	41	45
N antimalarials	437	437	437	437	17	19
Accuracy (%)	94.73	93.13	94.02	91.52	82.92	91.11
MCC[c]	0.89	0.86	0.87	0.82	0.65	0.82
Families of drugs [d]	broader range	broader range	broader range	broader range	low range	low range
Test set						
N total	442	442	442	442	18	15
N antimalarials	160	160	160	160	8	6
Predictability (%)	93.89	93.44	93.42	90.50	88.88	60.00
MCC[c]	0.87	0.86	0.86	0.79	0.92	0.22
Families of drugs [d]	broader range	broader range	broader range	broader range	low range	low range

[a] Equations **11** and **12** are reported in this work and models **13** and **14** were obtained previously by the present authors using non-stochastic and stochastic linear indices.[27] Equations **15** and **16** were reported by Gozalbez et al. [6] for two different studies: Eq. **15** was performed for the classification of antimalarial drugs and non-antiprotozoan drugs and, Eq. **16** for the discrimination between antimalarials and antiprotozoan drugs without antimalarial activity. [b] LDA refers to Linear discriminant analysis. [c] Matthews correlation coefficient. [d] Only largely represented families were considered.

Making use of the models obtained here (Eqs. **11** and **12**) which includes non-stochastic and non-stochastic quadratic indices, 94.73% and 93.13% of compounds in the training dataset were correctly classified. As can be observed in Table 9, the models **13** and **14**, obtained considering non-stochastic and stochastic linear indices,[27] shows lower values for such parameters (accuracy of 94.02% (93.42%) and 91.52% (90.50%) in training (test) set, correspondingly. Also the models reported in this work shown a higher MCC than models obtained in our previous study. As can be seen, models develop with quadratic maps-based *TOMOCOMD-CARDD* descriptors (Eqs. **11** and **12**) shows better parameters in all cases that models development with linear maps-based *TOMOCOMD-CARDD* indices (Eqs. **13** and **14**; see also equations 10 and 11 in reference 27). In this sense, we can conclude, that with the use of quadratic indices it is possible to codify useful chemical information and to obtain classification models comparable or even better than those obtained using analogous descriptors already reported.

On the other hand, in the last decade other two *–in silico-* method have also been used to develop two structure-based classification models (Eqs. **15** and **16** in Table 9) of antimalarial activity, which give rise to a good discrimination of this activity in large and heterogeneous series of organic compounds.[6] We also pretend to compare both approaches in order of showing the potentialities of our method. In this case, due to differences in the composition of experimental data used in carrying out the QSAR, it is not feasible to perform a "strict" comparison between the method reported previously[6] and the current approach. However, a relative comparison could be based on the kind of method used for deriving the QSAR and their statistical parameters, the number and diversity of chemical structural patterns contained in the data, the overall accuracy (%), Matthews correlation coefficient and the method which was used for the validation of the models. Table 9 also shows these chemometric coefficients for all approaches.

The global good classification in the training set of quadratic maps-based *TOMOCOMD-CARDD* models was higher than the two reported LDA equations (see Table 9). It is remarkable that the *TOMOCOMD-CARDD* models were derived from training series 27.3(1120/41), and 24.8(1120/45) times bigger than the series used by Gozalbes et. al.[6] In this sense, the overall accuracy in test sets of quadratic maps-based *TOMOCOMD-CARDD* models was higher than the rest of two reported LDA equations (see Table 9).

Another remarkable aspect is refereed to the spectrum of structural patterns considered in the studies under comparison. Without doubts, for the development of the *TOMOCOMD-CARDD* models reported here, a broader diversity of antimalarial was considered.

4.4. Virtual Screening of Ras FTase Inhibitors: An Experiment of Lead Generation

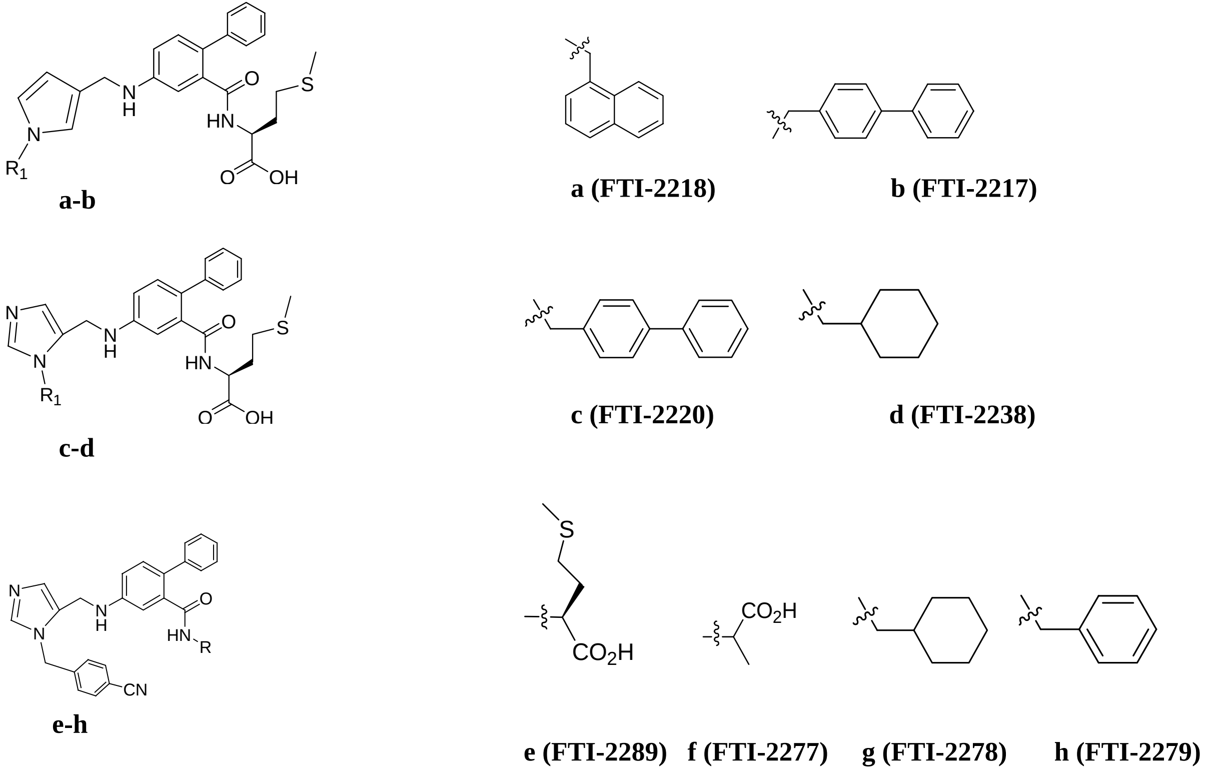
One of the most important aspects of any quantitative structure-activity relationship model is its ability to predict the desired activity for new compounds not included in the training data set. Virtual screening of large databases considering the use of such models has emerged as an interesting alternative to high-throughput screening and an important drug-design tool.[102-104] With the aim of testing the ability of our models to detecting new lead compounds with "unknown" structures, we carried out a simulated virtual screening of inhibitors of Farnesyltransferase (FTAse) that showed potent antimalarial activity in cell assays.[105] No one compound with this kind of structure was included in the training data set, and in this sense this evaluation is equivalent to the discovery of new lead compounds using the developed models. In this simulation, 10 previously reported FTase inhibitors with potent antimalarial activity were evaluated with models **11** and **12** as active/inactive ones. The results of the classification are shown in Table 5 and the molecular structures are illustrated in Scheme 1.

Table 5. Results of the Virtual Screening Simulation of Peptidomimetic Inhibitors of Protein Farnesyltransferase (FTAse) that Showed Potent Antimalarial Activity in Cell Assays.

Compound[a]	***P. falciparum*** **Infected RBC** **ED_{50} (μg/mL)**[b]	**Eq. 11**		**Eq. 12**	
		(ΔP%)[c]	**class**	**(ΔP%)**[c]	**class**
a (FTI-2218)	10	88.40	+	88.05	+
b (FTI-2217)	12	93.44	+	90.88	+
c (FTI-2220)	5	97.13	+	92.63	+
d (FTI-2238)	10	27.23	+	74.48	+
e (FTI-2289)	13	-39.22	-	19.46	+
f (FTI-2277)	3	69.39	+	90.44	+
g (FTI-2278)	3	-25.08	-	26.14	+
h (FTI-2279)	4	-61.68	-	38.75	+
i (FTI-2291)	10	18.90	+	36.45	+
j (FTI-2153)	2	21.28	+	58.45	+

[a] Compounds a-j were taken from Ohkanda et al., 2001 (Ref. 105). [b] Inhibition at 20μM, RBC = Red Blood Cell. [c] Results of the classification of compounds obtained from Eqs. **11** and **12**, respectively.

Scheme 1. Molecular Structure of Peptidomimetic Inhibitors of Protein Farnesyltransferase (FTAse) that Showed Potent Antimalarial Activity in Cell Assays.



a-b **a (FTI-2218)** **b (FTI-2217)**

c-d **c (FTI-2220)** **d (FTI-2238)**

e-h **e (FTI-2289)** **f (FTI-2277)** **g (FTI-2278)** **h (FTI-2279)**

As can be seen, both models classify correctly most of the 10 selected compounds. In the first case only 3 FTase inhibitors were classified as false inactives (70% of correct classification), while with model **12** the prediction has an overall accuracy of 100%.

This result is in accordance with the character of the *TOMOCOMD-CARDD* approach, which permits to consider implicitly, through the calculation of non-stochastic and stochastic quadratic molecular descriptors, substructural and global features responsible for a specific activity. In this way, new lead compounds could be designed using the *TOMOCOMD-CARDD* method described in this paper.

4.5. Experimental Results: Discovery of Novel Quinolinic Intermediaries as Antimalarial Compounds

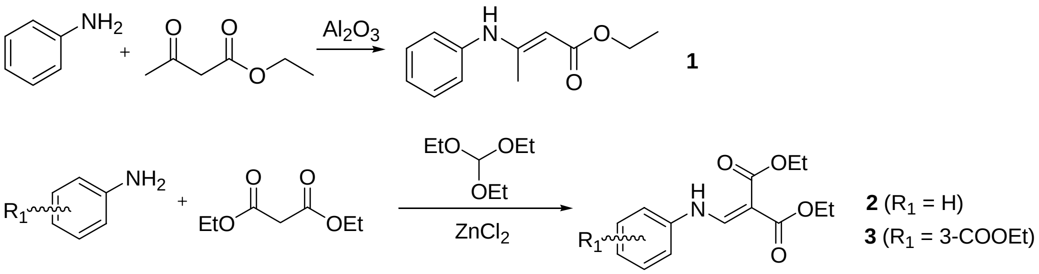
The aim of the present work is the development of discriminant functions for the rational design (or selection/identification) of new antimalarial compounds. As shown, we explored the ability of our classification models to find new active compounds carrying out an experiment of lead generation for the case of Ras FTPase inhibitors. These results encouraged us to developing a search of novel active compounds not described yet as antimalarials in the literature.

In this sense, we also explore a large dataset of organic-chemicals through virtual screening in order to discover novel candidates for antimalarial drug-like compounds. A great number of the candidates to be assayed as antimalarial, detected with our models, were sent to biological assays and their presentation will be the objective of a forthcoming paper. Nevertheless, in this work we want to show some promissory outcomes of this computational screening, which can represent an important starting point to the design of novel antimalarials.

It is well known, that the major of compounds used in the treatment of malaria are quinolinic derivatives such as quinine, chloroquine, mefloquine, halofantrine and primaquine. Acyclic β-enaminoesters and arylaminomethylenemalonates are synthetic intermediates of quinolinic compounds and can be achieved by economic and simple synthetic routes.[106,107] On the other hand, there are not many researches related to the biological activity of enamine compounds. Taking that into account, we explored in our search the behavior of some acyclic β-enamino esters and arylaminomethylenemalonates. Three of these compounds were initially evaluated with models **11** and **12** and in order to corroborate the predictions, prepared with excellent yields by very economic and simple methods, and evaluated against two strains of *Plasmodium falciparum.*

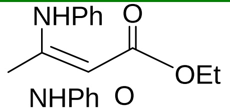
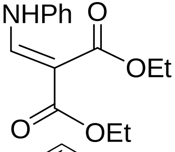
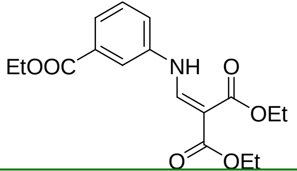
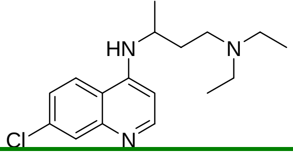
The acyclic β-enamino ester **1** were prepared by means of a nucleofilic addition of the aromatic amine to the keto group of the corresponding β-keto ester, using a previously described methodology.[108] Arylaminometilenemalonates were synthesized by means of "one pot" process, starting from equimolar quantities of the corresponding aniline, ethyl malonate and ethyl orthoformate in the presence of catalytic amount of $ZnCl_2$.[109] Both general procedures are shown in Scheme 2. All the structures were confirmed by spectroscopic data analysis which is given as Supporting Information.

Scheme 2. Synthetic Procedure for the Synthesis of Quinolinic Intermediaries.



The results of the prediction process using models **11** and **12**, as well as the minimum inhibitory concentration (MIC) for the three assayed compounds against K1 and Palo Alto strains are shown in Table 6.

Table 6. Synthetic Intermediates of Quinolinic Compounds Evaluated in the Present Study, their Classification (ΔP%) According to the *TOMOCOMD-CARDD* Approach, their Antimalarial Activity against two Malarial Strain and Antimalarial Activity of Chloroquine.

Compound	**Structure**	**ΔP%**[a]	**class**	**ΔP%**[b]	**class**	**MIC(pmol/well)**	
						K1	Palo Alto
1		-93.09	-	-78.83	-	100	100
2		-63.02	-	-1.25	-	>100	>100
3		**31.21**	+	**89.02**	+	**32**	**16**
Chloroquine		**77.93**	+	**77.96**	+	**8**	**4**

[a, b]Results of the classification of compounds obtained from Eqs. **11** and **12**, correspondingly.

The sensitivity control of each strain was carried out with chloroquine diphosphate. The MIC of chloroquine for sensitive strains is 5.7 pmol/well, i.e. strains with MIC above of this value are resistant to this compound.[110] In our study, the determined value of the MIC for K1 strain was 8 pmol/well (µmol/L) and for Palo Alto strain 4 pmol/well (0.8 µmol/L) confirming the sensitivity of the used strains.

As expected, compound **1** did not show activity against K1 and Palo Alto strains. The inhibition of the schizont maturation was observed at 100 pmol/well. Compound **2** did not inhibit the growth of parasites at any of the assayed concentrations (MIC > 100 pmol/well). Conversely, and in accordance with the predictions, the best results were observed for compound **3**, which showed a MIC = 32 pmol/well against K1 strain and a MIC = 16 pmol/well) for the case of Palo Alto strain.

Taking into account that this is the first report of an arylaminomethylenemalonate with antimalarial activity, the result can be considered as a very promissory starting point for the future design and refinement of novel compounds with higher antimalarial activity. That is to say, compound 3 was tested at higher doses than chloroquine diphosphate (reference or control antimalarial drug), but this result leaves a door open to a virtual variation study of the structure of these compounds in order to improve their antimalarial activity. Other chemicals in the same family as compound **3**, as well as other chemicals not in this family, were also predicted as antimalarials. The synthesis, characterization, and biological evaluation of these compounds are, however, beyond the scope of the present paper and will be discussed elsewhere. It is important to recall that the aim of this study is not to validate the model but to provide an experimental example of how to use the model for potential drug discovery.

5. CONCLUDING REMARKS

The introduction and use of graph theoretical descriptors for rational drug design has become an attractive tool for medicinal chemists. In this sense, the fusion of high throughput screening and classification-based QSAR models in an attempt to minimize the costs in terms of time, financial, human, and animal resources is becoming a viable alternative to massive screening. In this work, we have shown that *TOMOCOMD-CARDD* approach can be applied to generate useful quantitative models for the classification of antimalarials. In flexible way, this method permits a quick *in silico* discovery of new candidates to lead compounds making use of a minimum of resources. Considering a training data set of compounds with a considerable structural variability, we reduce the degree of uncertainly for this process. The simulated virtual screening of Ras FTase inhibitors with antimalarial properties has proved the ability of our models for an adequate discrimination of new active compounds from inactive ones. The collected data of active compounds used in this study, results an important tool not only for the theoretical research, but for the general scientific work in this area.

Using the developed models, a new lead candidate has been identified as a promising starting point for the design of new arylaminomethylenemalonates with potent antimalarial activity. Some works in this direction are at the moment in progress and will be published in a forthcoming paper.

The interactive character of the *TOMOCOMD-CARDD* approach permits the future inclusion of new antimalarial drugs in the training data set and the generation of each time more “intelligent” models. In this sense, the new considered structural patterns will recognized for the models and a better discrimination of such kind of compounds will be obtained. However, this point is out of the general scope of the present work.

Supporting Information Available: The complete list of compounds used in training and prediction sets, as well as their structures, posterior classification according to model **11** and **12**, chemistry and data analysis of the obtained chemicals is available free of charge via Internet at http://pubs.acs.org.

ACKNOWLEDGEMENTS: We thank FAPESP, CAPES and CNPq for the financial support.

6. REFERENCES AND NOTES

(1) Walsh, J. A. Disease Problems in the World. *Ann. N.Y. Acad. Sci.* **1989**, *569*, 1-16.

(2) Torok, D. S.; Ziffer, H. Synthesis and Antimalarial Activities of N-Substituded 11-Azaartemisinins. *J. Med. Chem*. **1995**, *38*, 5045-5050.

(3) Posner, G. H.; O'Dowd, H.; Ploypradith, P.; Cumming, J. N.; Xie, S.; Shapiro, T. A. Antimalarial Cyclic Peroxy Ketals. *J. Med. Chem*. **1998**, *41*, 2164-2167.

(4) Posner, G. H.; Cumming, J. N.; Woo, S. H.; Ploypradith, P.; Xie, S.; Shapiro, T. A. Orally Active Antimalarial 3-Substituted Trioxanes: New Synthetic Methodology and Biological Evaluation. *J. Med. Chem*. **1998**, *41*, 940-951.

(5) Lin, A. J.; Zikry, A. B.; Kyle, D. E. Antimalarial Activity of New Dihydroartemisinin Derivatives. 7. 4- (p-Substituted phenyl)-4 (R or S)-[10 (alpha or beta)-hydroartemisininoxy]butyric Acids. *J. Med. Chem*. **1997**, *40*, 1396-1400.

(6) Gonzalbes, R.; Gálvez, J.; Moreno, A.; García-Domenech, R. Discovery of New Antimalarial Compounds by Use of Molecular Connectivity Techniques. *J. Pharm. Pharmacol.* **1999**, *52*, 111-117.

(7) Go, M. L.; Ngiam, T. L.; Tan, A. L. C.; Kuaha, K.; Wilairat, P. Structure-Activity Relationships of some indolo(3,2-c)quinolines with Antimalarial Activity. *Eur. J. Pharm. Sci.* **1998**, *6*, 19-26.

(8) McKie, J. H.; Douglas, K. T.; Chan, C.; Roser, S. A.; Yates, R.; Read, M.; Hyde, J. E.; Dascombe, M. J.; Yuthavong, Y.; Sirawaraporn, W. Rational Drug Desing Approach for Overcoming Drug Resistance: Application to Pyrimethamine Resistance in Malaria. *J. Med. Chem*. **1998**, *41*, 1367-1370.

(9) De Dibyendu; Krogstad, F. M.; Byers, L. D.; Krogstad, D. J. Structure-Activity Relationships for Antiplasmodial Activity Among 7-Substituted 4-Aminoquinolines. *J. Med. Chem*. **1998**, *41*, 4918-4926.

(10) Estrada, E.; Peña, A.; García-Domenech, R. Designing Sedative/Hypnotic Compounds from a Novel Substructural Graph-theoretical Approach. *J. Comput.–Aided Mol. Design.* **1998**, *12*, 583-595.

(11) Estrada, E.; Peña, A. In Silico Studies for the Rational Discovery of Anticonvulsant Compounds. *Bioorg. Med. Chem*. **2000**, *8,* 2755-2770.

(12) Estrada, E.; Uriarte, E.; Montero, A.; Teijeira, M.; Santana, L.; De Clercq, E. A Novel Approach for Virtual Screening and Rational Design of Anticancer Compounds. *J. Med. Chem*. **2000**, *43*, 1975-1985**.**

(13) González-Díaz, H; Marrero-Ponce, Y.; Hernández, I; Bastida, I; Tenorio, E; Nasco, O; Uriarte, U; Castañedo, N.; Cabrera, M.A.; Aguila, E.; Marrero, O.; Morales, A.; Pérez, M. 3D-MEDNEs: An Alternative "In Silico" Technique for Chemical Research in Toxicology. 1. Prediction of Chemically Induced Agranulocytosis. *Chem. Res. Toxicol.* **2003,** *16,* 1318-1327.

(14) de Julián-Ortiz, J. V.; de Alapont, C. G.; Ríos-Santamarina, I.; García-Doménech, R.; Gálvez, J. Prediction of Properties of Chiral Compounds by Molecular Topology. *J. Mol. Graphics Mod.* **1998**, *16*, 14-18.

(15) Marrero-Ponce Y, Romero V (2002) **TOMOCOMD** software. Central University of Las Villas. **TOMOCOMD** (**TO**pological **MO**lecular **COM**puter **D**esign) for

Windows, version 1.0 is a preliminary experimental version; in future a professional version will be obtained upon request to Y. Marrero: yovanimp@qf.uclv.edu.cu or ymarrero77@yahoo.es

(16) Marrero-Ponce, Y. Total and Local Quadratic Indices of the Molecular Pseudograph's Atom Adjacency Matrix: Applications to the Prediction of Physical Properties of Organic Compounds. *Molecules.* **2003**, *8*, 687-726.

(17) Marrero-Ponce, Y. Linear Indices of the "Molecular Pseudograph's Atom Adjacency Matrix": Definition, Significance-Interpretation and Application to QSAR Analysis of Flavone Derivatives as HIV-1 Integrase Inhibitors. *J. Chem. Inf. Comput. Sci*. **2004**, *44*, 2010-2026.

(18) Marrero-Ponce, Y. Total and Local (Atom and Atom-Type) Molecular Quadratic Indices: Significance-Interpretation, Comparison to Other Molecular Descriptors and QSPR/QSAR Applications. *Bioorg. Med. Chem*. **2004,** *12,* 6351-6369.

(19) Marrero-Ponce, Y.; Castillo-Garit, J. A.; Torrens, F.; Romero-Zaldivar, V.; Castro E. Atom, Atom-Type and Total Linear Indices of the "Molecular Pseudograph's Atom Adjacency Matrix": Application to QSPR/QSAR Studies of Organic Compounds. *Molecules.* **2004**, *9,* 1100-1123.

(20) Marrero-Ponce, Y.; González-Díaz, H.; Romero-Zaldivar, V.; Torrens, F.; Castro, E. A. 3D-Chiral Quadratic Indices of the "Molecular Pseudograph's Atom Adjacency Matrix" and their Application to Central Chirality Codification: Classification of ACE Inhibitors and Prediction of σ-Receptor Antagonist Activities. *Bioorg. Med. Chem.* **2004,** *12,* 5331-5342.

(21) Marrero-Ponce, Y.; Cabrera, M., A.; Romero, V.; Ofori, E.; Montero, L. A. Total and Local Quadratic Indices of the "Molecular Pseudograph's Atom Adjacency Matrix". Application to Prediction of Caco-2 Permeability of Drugs. *Int. J. Mol. Sci.* **2003***, 4,* 512-536.

(22) Marrero-Ponce, Y.; Cabrera, M. A.; Romero, V.; González, D. H.; Torrens, F. A New Topological Descriptors Based Model for Predicting Intestinal Epithelial Transport of Drugs in Caco-2 Cell Culture. *J. Pharm. Pharm. Sci.* **2004**, *7*, 186-199.

(23) Marrero-Ponce, Y.; Cabrera, M. A.; Romero-Zaldivar, V.; Bermejo, M.; Siverio, D.; Torrens, F. Prediction of Intestinal Epithelial Transport of Drug in (Caco-2) Cell Culture from Molecular Structure using *'in silico'* Approaches During Early Drug Discovery. *Internet Electronic J. Mol. Des.* **2005**, *4*, 124-150.

(24) Marrero-Ponce, Y.; Castillo-Garit, J. A.; Olazabal, E.; Serrano, H. S.; Morales, A.; Castañedo, N.; Ibarra-Velarde, F.; Huesca-Guillen, A.; Jorge, E.; Sánchez, A. M.; Torrens, F.; Castro, E. A. Atom, Atom-Type and Total Molecular Linear Indices as a Promising Approach for Bioorganic & Medicinal Chemistry: Theoretical and Experimental Assessment of a Novel Method for Virtual Screening and Rational Design of New Lead Anthelmintic. *Bioorg. Med. Chem*. **2005**, *13*, 1005-1020.

(25) Marrero-Ponce, Y.; Castillo-Garit, J. A.; Olazabal, E.; Serrano, H. S.; Morales, A.; Castañedo, N.; Ibarra-Velarde, F.; Huesca-Guillen, A.; Jorge, E.; del Valle, A.; Torrens, F.; Castro, E. A. *TOMOCOMD-CARDD*, a Novel Approach for Computer-Aided "Rational" Drug Design: I. Theoretical and Experimental Assessment of a Promising Method for Computational Screening and *in silico* Design of New Anthelmintic Compounds. *J. Comput. Aided Mol. Des.* **2004**, *18*, 615-633.

(26) Marrero-Ponce, Y.; Huesca-Guillen, A.; Ibarra-Velarde, F. Quadratic Indices of the "Molecular Pseudograph's Atom Adjacency Matrix" and Their Stochastic Forms: A Novel Approach for Virtual Screening and *in silico* Discovery of New Lead Paramphistomicide Drugs-like Compounds. *J. Theor. Chem.* (*THEOCHEM*). **2005**, *717*, 67-79.

(27) Marrero-Ponce, Y.; Montero-Torres, A.; Romero-Zaldivar, C.; Iyarreta-Veitía, I.; Mayón Peréz, M.; García Sánchez, R. Non-Stochastic and Stochastic Linear Indices of the "Molecular Pseudograph's Atom Adjacency Matrix": Application to "*in silico"* Studies for the Rational Discovery of New Antimalarial Compounds. *Bioorg. Med. Chem.* **2005**, *13*, 1293-1304.

(28) Marrero-Ponce, Y.; Medina-Marrero, R.; Torrens, F.; Martinez, Y.; Romero-Zaldivar, V.; Castro, E. A. Non-Stochastic and Stochastic Quadratic Indices of the Molecular Pseudograph's Atom Adjacency Matrix: A Promising Approach for Chemical Information and Modeling of Antibacterial Activity. *Bioorg. Med. Chem.* In Press. DOI: 10.1016/j.bmc.2005.02.015.

(29) Marrero-Ponce, Y.; Medina-Marrero, R.; Martinez, Y.; Torrens, F.; Romero-Zaldivar, V.; Castro, E. A. Non-Stochastic and Stochastic Linear Indices of the Molecular Pseudograph's Atom Adjacency Matrix: A Novel Approach for Computational *–in silico-* Screening and "Rational" Selection of New Lead Antibacterial Agents. *J. Mol. Mod.* Accepted for publication.

(30) Marrero-Ponce, Y.; Nodarse, D.; González-Díaz, H.; Ramos de Armas, R.; Romero-Zaldivar, V.; Torrens, F.; Castro, E. Nucleic Acid Quadratic Indices of the "Macromolecular Graph's Nucleotides Adjacency Matrix". Modeling of Footprints after the Interaction of Paromomycin with the HIV-1 Ψ-RNA Packaging Region. *Int. J. Mol. Sci.* **2004**, *5*, 276-293.

(31) Marrero-Ponce, Y.; Castillo-Garit, J.A.; Nodarse, D. Linear Indices of the "Macromolecular Graph's Nucleotides Adjacency Matrix" as a Promising Approach for Bioinformatics Studies. 1. Prediction of Paromomycin's Affinity Constant with HIV-1 Ψ-RNA Packaging Region. *Bioorg. Med. Chem.* Accepted for publication.

(32) Marrero-Ponce, Y.; Medina, R.; Castro, E. A.; de Armas, R.; González, H.; Romero, V.; Torrens, F. Protein Quadratic Indices of the "Macromolecular Pseudograph's α-Carbon Atom Adjacency Matrix". 1. Prediction of Arc Repressor Alanine-mutant's Stability. *Molecules.* **2004,** *9*, 1124–1147.

(33) Marrero-Ponce, Y.; Medina-Marrero, R.; Castillo-Garit, J. A.; Romero-Zaldivar, V.; Torrens, F.; Castro, E. A. Protein Linear Indices of the "Macromolecular Pseudograph's α-Carbon Atom Adjacency Matrix" in Bioinformatics. 1. Prediction of Protein Stability Effects of a Complete Set of Alanine Substitutions in Arc Repressor. *Bioorg. Med. Chem.* In Press. DOI: 10.1016/j.bmc.2005.01.062.

(34) Pauling, L. *The Nature of Chemical Bond*; Cornell University Press: New York, **1939**; 2-60.

(35) Walker, P. D.; Mezey, P. G. Molecular Electron Density Lego Approach to Molecule Building. *J. Am. Chem. Soc.* **1993**, *115*, 12423-12430.

(36) Golbraikh, A.; Bonchev, D.; Tropsha, A. Novel Chirality Descriptors Derived from Molecular Topology. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 147-158.

(37) González-Díaz, H.; Hernández-Sánchez, I.; Uriarte, E.; Santana, L. Symmetry Considerations in Markovian Chemicals ‘in silico’ Design (MARCH-INSIDE) I.: Central Chirality Codification, Classification of ACE Inhibitors and Prediction of σ-Receptor Antagonist Activities. *Comput. Biol. Chem*. **2003**, *27*, 217-227.

(38) Klein, D. J. Graph Theoretically Formulated Electronic–Structure Theory, *Internet Electron. J.Mol. Des.* **2003**, *2*, 814–834, http://www.biochempress.com.

(39) González-Díaz, H.; Bastida, I.; Castañedo, N., Nasco, O.; Olazabal, E., Morales, A., Serrano, H. S., Ramos de A. R. Simple Stochastic Fingerprints Towards Mathematical Modelling in Biology and Medicine. 1. The Treatment of Coccidiosis. *Bull. Math. Biol.* **2004**, *66*, 1285-1311.

(40) Ann, D. C. *Antimalarial Agents. In: Burger’s Medicinal Chemistry and Drug Discovery, Fith Edition, Volume 5: Therapeutic Agents*; Wiley-Interscience Publication: New York, **1997**.

(41) Negwer, M. *Organic-Chemical Drugs and their Synonyms*; Akademie-Verlag: Berlin, **1987**.

(42) Domínguez, J. N.; López, S.; Charris, J.; Iarroso, L.; Lobo, G.; Semenov, A.; Olson, J. E.; Rosenthal, P. J. Synthesis and Antimalarial Effects of Phenothiazine Inhibitors of a Plasmodium Falciparum Cystein Protease. *J. Med. Chem.* **1997**, *40*, 2726-2732.

(43) Hawley, S. R.; Bray, P. G.; Mungthin, M.; Atkinson, J. D.; O’Neill, P. M.; Ward, S. A. Relationship Between Antimalarial Drug Activity, Accumulation, and Inhibition of Heme Polymerization in *Plasmodium falciparum* In Vitro. *Antimicrob. Agents Chemother.* **1998**, *42*, 682-686.

(44) Rucker, G.; Schenkel, E. P.; Manns, D.; Mayer, R.; Heiden, K.; Heinzmann, B. M. Sesquiterpene Peroxides from *Senecio Selloi* and *Eupatorium rufescens*. *Planta Med*. **1996**, *62*, 565-566.

(45) Ring, C. S.; Sun, E.; Mekerrow, J. H.; Lee, G. K.; Rosenthal, P. J.; Kurtz, I. D.; Cohen, F. E. Structure-Based Inhibitor Design by Using Protein Models for the Development of Antiparasitic Agents. *Proc. Natl. Acad. Sci*. **1993**, *90*, 3583-3587.

(46) Ryley, J. F.; Peters, W. The Antimalarial of some Quinolone Esters. *Ann. Trop. Med. Parasitol.* **1970**, *64*, 209-222.

(47) Basco, L. K.; Dechy-Cabaret, O.; Ndounga, M.; Meche, F. S.; Robert, A.; Meunier, B. In vitro Activities of DU-1102, a New Trioxaquine DerivativeAgainst Plasmodium falciparum Isolates. *Antimicrob. Agents Chemother.* **2001**, *45*, 1886-1888.

(48) Haque, T.S.; Skillman, A.G; Lee, C. E.; Habashita, H.; Gluzman, I. Y.; Swing, T. J. A.; Goldberg, D. E.; Knuts, I. D.; Ellman, J. A. Potent, Low-Molecular-Weight Non-peptide Inhibitors of Malarial Aspartyl Protease Plasmepsin II. *J. Med. Chem*. **1999**, *42*, 1428-1440.

(49) Raynes, K. Bisquinoline Antimalarials: Their Role in Malaria Chemotherapy. *Int. J. Parasitol.* **1999**, *29*, 367-379.

(50) Tsai, C. S.; Shen, A. Y. Synthesis and Bilogical Evaluation of Some Potential Antimalarials. *Arch. Pharm*. **1994**, *327*, 677-679.

(51) Posner, G. H.; Tao, X.; Cumming, J. N.; Klinedinst, D.; Shapiro, T. A. Antimalarially Potent, Easily Prepared, Fluorinated Endoperoxides. *Tetrahedron Lett.* **1996**, *37*, 7225-7228.

(52) Philipp, A.; Kepler, J. A.; Johnson, B. H.; Carroll, F. I. Peptide Derivatives of Primaquine as Potential Antimalarial Agents. *J. Med. Chem*. **1998**, *31*, 870-874.

(53) Figgitt, D.; Denny, W.; Chvalitshewinkoon, P.; Wilairat, P.; Ralph, R. In vitro Study of Anticancer Acridines as Potential Antitrypanosomal and Antimalarial Agents. *Antimicrob. Agents Chemother***.** **1992**, *36*, 1644-1647.

(54) Nga, T. T. T.; Menage, C.; Begue, J. P.; Delpon, D. B.; Gantier, J. C. Synthesis and Antimalarial Activities of Fluoroalkyl Derivatives of Dihydroartemisinin. *J. Med. Chem*. **1998**, *41*, 4101-4108.

(55) Avery, M. A.; Bonk, J. D.; Chong, W. K. M.; Mehrota, S.; Miller, R.; Milhous, W.; Goins, D. K.; Venkatesan, S.; Wyandt, C.; Khan, I.; Avery, B. A. Structure-Activity Relationships of the Antimalarial Agent Artemisinin. 2. Effect of Heteroatom Substitution at O-11: Synthesis and Bioassay of N-Alkyl-11-aza-9-desmethylartemisinins. *J. Med. Chem*. **1995**, *38*, 5038-5044.

(56) Posner, G. H.; Wang, D.; González, L.; Tao, X.; Cumming, J. N.; Klinedints, D.; Shapiro, T. A. Mechanism-Based Design of Simple, Symmetrical, Easily Prepared, Potent Antimalarial Endoperoxides. *Tetrahedron Lett.* **1996**, *37*, 815-818.

(57) Pu, Y. M.; Torok, D. S.; Ziffer, H.; Pan, X. Q.; Meshnick, S. R. Synthesis and Antimalarial Activities of Several Fluorinated Artemisinin Derivatives. *J. Med. Chem*. **1995**, *38*, 4120-4124.

(58) Posner, G. H.; O'Dowd, H.; Caferro, T.; Cumming, J. N.; Ploypradith, P.; Xie, S.; and Shapiro, T. A. Antimalarial Sulfone Trioxanes. *Tetrahedron Lett.* **1998**, *39*, 2273-2276.

(59) Posner, G. H.; McGarvey, D. J.; Oh, C. H.; Kumar, N.; Meshnick, S. R.; Asawamahasadka, W. Structure-Activity Relationships of Lactone Ring-Opened Analogs of the Antimalarial 1,2,4-Trioxane Artemisinin. *J. Med. Chem*. **1995**, *38*, 607-612.

(60) Posner, G. H.; González, L.; Cumming, J. N.; Klinedints, D.; Shapiro, T. A. Synthesis and Antimalarial Activity of Heteroatom-Containing Bicyclic Endoperoxides. *Tetrahedron* **1997**, *53*, 37-50.

(61) Venugopalan, B.; Bapat, C. P.; Karnik, P. J. Synthesis of A Nonel Ring Contracted Artemisinin Derivative. *Bioorg. Med. Chem. Lett.* **1994**, *4*, 751-752.

(62) Avery, M. A.; Gao, F.; Chong, W. K. M.; Hendrickson, T. F.; Inman, W. D.; Crews P. Synthesis, Conformational Analysis, and Antimalarial Activity of Tricyclic Analogs of Artemisinin. *Tetrahedron.* **1994**, *50*, 957-972.

(63) Avery, M. A.; Mehrotra, S.; Johnson, T. L.; Bonk, J. D.; Vroman, J. A.; Miller, R. Structure-Activity Relationships of the Antimalarial Agent Artemisinin. 5. Analogs of 10-Deoxoartemisinin Substituted at C-3 and C-9. *J. Med. Chem*. **1996**, *39*, 4149-4155.

(64) Venugopalan, B.; Bapat, C. P.; Karnik, P. J.; Chatterjee, D. K.; Iyer, N.; Lepcha, D. Antimalarial Activity of Novel Ring–Contracted Artemisinin Derivatives. *J. Med. Chem*. **1995**, *38*, 1992-1927.

(65) Zouhiri, F.; Desmaele, D.; d 'Angelo, J.; Riche, C.; Gay, F.; Cicéron, L. Artemisinin Tricyclic Analogs: Role of a Methyl Group at C-5a. *Tetrahedron Lett.* **1998**, *39*, 2969-2972.

(66) Posner, G. H.; Parker, M. H.; Northrop, J.; Elias, J. S.; Ploypradith, P.; Xie, S.; Shapiro, T. A. Orally Active, Hydrolytically Stable, Semisynthetic, Antimalarial Trioxanes in The Artemisinin Family. *J. Med. Chem*. **1999**, *42*, 300-304.

(67) Cumming, J. N.; Wang, D.; Park, S. B.; Shapiro, T. A.; Posner, G. H. Design, Synthesis, Derivatization, and Structure-Activity Relationships of Simplified, Tricyclic, 1,2, 4-Trioxane Alcohol Analogues of the Antimalarial Artemisinin. *J. Med. Chem*. **1998**, *41*, 952-964.

(68) Posner, G. H.; Oh, C. H.; Gerena, L.; Milhous, W. K. Extraordinarily Potent Antimalarial Compounds: New, Structurally Simple, easily Synthesized, Tricyclic 1,2,4-Trioxanes. *J. Med. Chem*. **1992**, *35*, 2459-2467.

(69) Calas, M.; Cordina, G.; Bompart, J.; Bari, M. B.; Jei, T.; Ancelin, M. L.; Vial, H. Antimalarial Activity of Molecules Interfering with *plasmodium falciparum* Phospholipid Metabolism. Structure-Activity Relationships Analysis. *J. Med. Chem.* **1997**, *40*, 3557-3566.

(70) Ismail. F. M. D.; Dascombe, M. J.; Carr, P.; North, S. E. An Exploration of the Structure-activity Relationships of 4-Aminoquinolines: Novel Antimalarials with Activity. *J. Pharm. Pharmacol*. **1996**, *48*, 841-850.

(71) Ram, V. J.; Saxena, A. S.; Srivastavab, S.; Chandrab, S. Oxygenated Chalcones and Bischalcones as Potential Antimalarial Agents. *Bioorg. Med. Chem. Lett.* **2000**, *10*, 2159-2161.

(72) Posner, G. H.; Northrop, J.; Paik, I. H.; Borstnik, K.; Dolan, P.; Kensler, T. W.; Xiec, S.; Shapiroc, T. A. New Chemical and Biological Aspects of Artemisinin-Derived Trioxane Dimers. *Bioorg. Med. Chem.* **2000**, *10*, 227–232.

(73) Gironés, X.; Gallegos, A.; Carbó-Dorca, R. Antimalarial Activity of Synthetic 1,2,4-Trioxanes and Cyclic Peroxy Ketals, a Quantum Similarity Study. *J. Comput.–Aided Mol. Design.* **2001**, *15*, 1053–1063.

(74) Cheng, F.; Shen, J.; Luo, X.; Zhu, W.; Gu, J.; Ji, R.; Jiang, H.; Chen, K. Molecular Docking and 3-D-QSAR Studies on the Possible Antimalarial Mechanism of Artemisinin Analogues. *Bioorg. Med. Chem.* **2002**, *10*, 2883–2891.

(75) Santos-Filho, O. A.; Hopfinger, A. J. A Search for Sources of Drug Resistance by the 4D-QSAR Analysis of a Set of Antimalarial Dihydrofolate Reductase Inhibitors. *J. Comput.–Aided Mol. Design.* **2001**, *15*, 1–12.

(76) Jain, R.; Vangapandu, S.; Jain, M.; Kaur, N.; Singhb, S.; Singhb, P. P. Antimalarial Activities of Ring-Substituted Bioimidazoles. *Bioorg. Med. Chem. Lett.* **2002**, *12*, 1701–1704.

(77) Itoh, T.; Shirakami, S.; Ishida, N.; Yamashita, Y.; Yoshida, T.; Kimb, H. S.; Watayab, Y. Synthesis of Novel Ferrocenyl Sugars and their Antimalarial Activities. *Bioorg. Med. Chem. Lett.* **2000**, *10*, 1657-1659.

(78) Reichenberg, A.; Wiesner, J.; Weidemeyer, C.; Dreiseidler, E.; Sanderbrand, S.; Altincicek, B.; Beck, E.; Schlitzerc, M.; Jomaa , H. Diaryl Ester Prodrugs of FR900098 with Improved In Vivo Antimalarial Activity. *Bioorg. Med. Chem. Lett.* **2001***,* *11*, 833–835.

(79) Ryckebusch, A.; Deprez-Poulain, R.; Maes, L.; Debreu-Fontaine, M. A.; Mouray, E.; Grellier, P.; Sergheraert, C. Synthesis and in Vitro and in Vivo Antimalarial Activity of N1-(7-Chloro-4-quinolyl)-1,4-bis(3-aminopropyl)piperazine Derivatives. *J. Med. Chem.* **2003**, *46*, 542-557.

(80) Nöteberg, D.; Hamelink, E.; Hulten, J.; Wahlgren, M.; Vrang, L.; Samuelsson, B.; Hallberg, A. Design and Synthesis of Plasmepsin I and Plasmepsin II Inhibitors with Activity in *Plasmodium falciparum*-Infected Cultured Human Erythrocytes. *J. Med. Chem.* **2003**, *46,* 734-746.

(81) Murray, P. J.; Kranz, M.; Ladlow, M.; Taylor, S.; Berst, F.; Holmes, A. B.; Keavey, K. N.; Laxa-chamiec, A.; Seale, P. W.; Stead, P.; Upton, R. J.; Croft, S. L.; Clegg, W.; Elsegood, M. R. J. The Synthesis of Cyclic Tetrapeptoid Analogues of the Antiprotozoal Natural Product Apicidin. *Bioorg. Med. Chem. Lett.* **2001**, *11*, 773-776.

(82) Chapman & Hall. The Merck Index. Twelfth Edition. **1996**

(83) Mc Farland, J. W.; Gans, D. J. *Cluster Significance Analysis. In Chemometric Methods in Molecular Design*; van Waterbeemd, H., Ed.; VCH Publishers: New York, **1995**; pp 295–307.

(84) Johnson, R. A.; Wichern, D. W.; Applied Multivariate Statistical Analysis. Prentice-Hall, N.J, **1988**.

(85) STADISTICA, version 5.5; Statsoft Inc., **1999**.

(86) Wold, S; Erikson, L. *Statistical Validation of QSAR Results. Validation Tools. In Chemometric Methods in Molecular Design*; van de Waterbeemd, H., Ed.; VCH Publishers: New York, **1995**; 309-318.

(87) Golbraikh, A.; Tropsha, A. Predictive QSAR Modelling Based on Diversity Sampling of Experimental Datasets for the Training and Test Set Selection. *J. Mol. Graphic Modell.* **2002***, 20*, 269-276.

(88) Baldi, P.; Brunak, S.; Chauvin, Y.; Andersen, C. A.; Nielsen, H. Assessing the Accuracy of Prediction Algorithms for Classification: an Overview. *Bioinformatics*. **2000**, *16*, 412–424.

(89) Randić, M. Resolution of Ambiguities in Structure-Property Studies by Us of Orthogonal Descriptors. *J. Chem. Inf. Comput. Sci*. **1991**, *31*, 311-320.

(90) Randić, M. Orthogonal Molecular Descriptors. *New J. Chem*. **1991**, *15*, 517-525.

(91) Randić, M. Correlation of Enthalpy of Octanes with Orthogonal Connectivity Indices. *J. Mol. Struct.* (*Theochem*) **1991**, *233*, 45-59.

(92) Lučić, B.; Nikolić, S.; Trinajstić, N.; Jurić, D. The Structure-Property Models can be Improved Using the Orthogonalized Descriptors. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 532-538.

(93) Klein, D. J.; Randić, M.; Babić, D.; Lučić, B.; Nikolić, S.; Trinajstić, N. Hierarchical Orthogonalization of Descriptors. *Int. J. Quantum Chem.* **1997**, *63*, 215-222.

(94) Estrada, E.; Vilar, S.; Uriarte, E.; Gutierrez, Y. In Silico Studies Toward the Discovery of New Anti-HIV Nucleoside Compounds with the Use of TOPS-MODE and 2D/3D Connectivity Indices. 1. Pyrimidyl Derivatives. *J. Chem. Inf. Comput. Sci.* **2002**, *42,* 1194-1203.

(95) Estrada, E.; Uriarte, E. Recent Advances on the Role of Topological Indices in Drug Discovery Research. *Curr. Med. Chem.* **2001**, *8,* 1573-1588.

(96) Antonioletti, R.; Bonadies, F.; Orelli, L. O.; Scettri, A. Selective *C*-Alkylation of 1,3-dicarbonyl Compounds. *Gazz. Chim. Ital*. **1992**, *122*, 237-238.

(97) Perrin, D. D.; Armarego, W. L. F.; Perrin, D. R. *Purification of Laboratory Chemicals*; Pergamon Press: New York, **1980**.

(98) Rieckmann, K. H.; Campbell, G. H.; Sax, L. J.; Mrema, J. E. Drug Sensitivity of *Plasmodium falciparum*. An *in vitro* Microtechnique. *Lancet* **1978**, *1*, 22-23.

(99) Trager, W.; Jensen, J. B. Human Malaria in Continuos Culture. *Science* **1976**, *193*, 673-675.

(100) Diggs, C.; Joseph, K.; Flemmings, B.; Snodgrass, R.; Hines, F. Protein Synthesis in Vitro by Cryopreserved *Plasmodium falciparium*. *Am. J. Trop. Med. Hyg.* **1975**, *24*, 760-763.

(101) Lambros, C.; Vanderberg, J. P.; Synchronization of *Plasmodium falciparum* Erythrocytic Stages in Culture. *J. Parasitol*. **1979**, *65*, 418-420.

(102) Julián-Ortiz, J. V.; Gálvez, J.; Muñoz-Collado, C.; García-Domenech, R.; Gimeneo-Cardona, C. Virtual Combinatorial Synthesis and Computational Screening of New Potential Anti-Herpes Compounds. *J. Med. Chem*. **1999**, *42*, 3308-3314.

(103) Drie, J. H. V.; Lajiness, M. S. Approaches to Virtual Library Design. *Drug Disc. Today.* **1998**, *3*, 274-283.

(104) Lajiness, M. *Molecular Similarity-Based Methods for Selecting Compounds for Screening. In: Rouvray DH (ed) Computacional Chemical Graph Theory*; Nova Science: New York, **1990**.

(105) Ohkanda, J.; Lockman, J. W.; Yokoyama, K.; Gelb, M. H.; Croft, S. L.; Kendrich, H.; Harrell, M. I.; Feagin, J. E.; Blaskovich, M. A.; Sebti, S. M.; Hamilton, A. D. Peptidomimetic Inhibitors of Protein Farnesyltransferase Show Potent Antimalarial Activity. *Bioorg. Med. Chem. Lett.* **2001**, *11*, 761-764.

(106) Marvel, C.S.; Hager, C.S. Ethyl n-butylacetoacetate (Caproic and, α-acetyl-, ethyl ester). *Org. Synth. Coll*. **1941**, *1*, 248.

(107) Huckin, S.N.; Weiler, L.J. Alkylation of Dianions of β-keto Esters. *J. Am. Chem. Soc.* **1974**, *96*, 1082–1087.

(108) Ferraz, H. M. C.; Payret-Arrua, M. E.; De Oliveira, E. O.; Brandt C. A. A New and Efficient Approach to Cyclic β-enamino-esters and β-enamino-ketones by Iodine-Promoted Cyclization. *J. Org. Chem*. **1995**, *60*, 7357-7359.

(109) Hermecz, I.; Keresztúri, G.; Vasvári-Debreczy, L. Aminometylenemalonates and Their use in Heterocyclic Synthesis. *Adv. Heterocycl. Chem*. **1992**, *54*, 1-429.

(110) Smrkovski, L. L.; Buck, R. L.; Alcántara, A. K.; Rodríguez, C.S.; Uylangco, C.V. Studies of Resistance to Chloroquine, Quinine, Amodiaquine and Mefloquine Among Philippine Strains of Plasmodium falciparum. *Trans. R. Soc. Trop. Med. Hyg*. **1985**, *7*, 37-41.