

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

Yovani Marrero-Ponce,^{a,b,*} Alina Montero-Torres,^b Carlos Romero Zaldivar,^a Maité Iyarreta Veitia,^c Mariuchy Mayón Pérez^{a,b} and Rory N. García Sánchez^d

^aDepartment of Pharmacy, Faculty of Chemical-Pharmacy, Central University of Las Villas, Santa Clara, 54830 Villa Clara, Cuba

^bDepartment of Drug Design, Chemical Bioactive Center, Central University of Las Villas, Santa Clara, 54830 Villa Clara, Cuba

^cCentre d’études Pharmaceutiques, CNRS Biocis UMR 8076, Laboratoire de Synthèse de composés d’intérêt biologique,

Faculté de pharmacie, Université Paris-Sud 11, rue J.B. Clément, 92296 Châtenay-Malabry Cedex, France

^dLaboratorio de Investigación de Productos Naturales Antiparasitarios de la Amazonía (LIPNAA),

Universidad Nacional de la Amazonía Peruana (UNAP), Iquitos, 496 Peru

Received 8 October 2004; revised 5 November 2004; accepted 5 November 2004

Available online 8 December 2004

Abstract—Malaria is one of the most deadly diseases, affecting million of people especially in developing countries. Because of the rapidly increasing threat worldwide of malaria epidemics multidrugs resistant to therapies, there is an urgent global need to discover new classes of antimalarial compounds. In an effort to overcome this problem, we have investigated the use of structure-based classification models for the ‘rational’ selection/identification or design/optimization of new lead antimalarials from virtual combinatorial data sets. In this sense, *TOPOlogical MOlecular COmputer Design* strategy (**TOMOCOMD** approach) has been introduced in order to obtain two quantitative models for the discrimination of antimalarials. A collected data set containing 597 antimalarial compounds is presented as a helpful tool not only for theoretical chemist but for other researchers in this area. The validated models (including non-stochastic and stochastic indices) classify correctly more than 90% of compounds in both training and external prediction data sets. They showed high Matthews’ correlation coefficients; 0.87 and 0.82 for training and 0.86 and 0.79 for test set. The **TOMOCOMD**–**CARDD** approach implemented in this work was successfully compared with two of the most useful models for antimalarials selection reported so far. Thus we expect that these two QSAR models can be used in the identification of previously unknown antimalarials compounds.

© 2004 Elsevier Ltd. All rights reserved.

“If we take as our standard of importance the greatest harm to the greatest number, then there is no question that malaria is the most important of all infectious diseases.”
Sir Macfarlane Burnet as quoted in D. J. Wyler. *N. Engl. J. Med.* **1983**, 308, 875.

1. Introduction

The increasing resistance of malaria parasites to antimalarial drugs is a major contributor to the re-emergence

of this disease as a serious health problem in the world. Hundreds of millions of malaria cases occur annually, and ca. 3 million deaths take place per year.^{1,2} There is, therefore, an urgent necessity to identify new antimalarial drugs using ‘rational’ search methodologies, taking into account the complexity and cost of the process of drug discovery.^{3–5} In this sense, computer-aided drug design approach emerges as a promising tool.^{6–9} One of the major goals of the design strategy is the identification of structural subsystems responsible for a specific biological activity from large databases or libraries. Making use of 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}

Keywords: **TOMOCOMD**–**CARDD** software; Non-stochastic and stochastic linear indices; LDA; QSAR; Antimalarial compounds.

*Corresponding author. Tel.: +53 42 281192/281473; fax: +53 42 281130/281455; e-mail addresses: yovanimp@qf.uclv.edu.cu; ymarrero77@yahoo.es

In this context, our research group has recently introduced the novel computer-aided molecular design scheme TOMOCOMD-CARDD (acronym of *TO*pological *MO*lecular *CO*mputer *DE*sign-*CO*mputer *AI*ded ‘*R*ational’ *D*rug *D*esign).¹⁵ It calculates several new 2D/3D families of total and local (atom and atom type) topologic molecular descriptors, such as quadratic and linear indices, defined by analogy with the quadratic and linear mathematical maps.^{16,17} The flexibility of this ‘in silico’ method permits the study of small molecules as well as macromolecules such as nucleic acids.¹⁸ The TOMOCOMD approach has been successfully applied to the prediction of physicochemical and biological properties of chemicals and drug-like compounds.^{16–23}

The main objectives of this paper are, first, to find quantitative models, which allow the discrimination of antimalarial compounds from inactive ones and to carry out the validation of them; second, to introduce the stochastic molecular linear indices as a novel component of the TOMOCOMD-CARDD scheme and to compare the quality of the classification model obtained with this descriptors family with the model that includes non-stochastic indices. Finally, we pretend to compare the TOMOCOMD-CARDD approach for the selection or design of antimalarials with two of the most useful models reported in the literature.

2. Results and discussion

2.1. TOMOCOMD approach

TOMOCOMD¹⁵ is an interactive program for molecular design and bioinformatics research, which contains 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). In this paper, we outline salient features concerned with only the module CARDD and the calculation of non-stochastic and stochastic 2D linear indices.

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

1. Drawing the molecular pseudographs for each molecule of the data set, using the software drawing mode.
2. Selection of appropriate weights in order to differentiate the molecular atoms.
3. Compute total and local (atom and atom type) linear indices of the molecular pseudograph’s atom adjacency matrix and automatic generation of a file with the values of the selected descriptors (columns) for each molecule (rows).
4. By using mathematical techniques, such as multilinear regression analysis (MRA), Neural Networks (NN), Linear Discrimination Analysis (LDA), and so on, search of a QSPR/QSAR equation having the following appearance:

$$A = a_0 f_0(x) + a_1 f_1(x) + a_2 f_2(x) + \dots + a_k f_k(x) + c \quad (1)$$

where A is the measurement of the activity, $f_k(x)$ is the k th total linear indices, and the 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 and/or external cross-validation techniques.

Considering this methodology, the following descriptors were calculated:

- (i) $f_k(x)$ and $f_k^H(x)$ (k th total linear indices not-considering and considering H-atoms in the molecular pseudograph (G)).
- (ii) $f_{kL}(x_E)$ and $f_{kL}^H(x_E)$ (k th local (atom-type = heteroatoms: S, N, O) linear indices not-considering and considering H-atoms in the molecular pseudograph (G)).
- (iii) $f_{kL}^H(x_{E-H})$ (k th local (atom type = H-atoms bonding to heteroatoms: S, N, O) linear indices considering H-atoms in the molecular pseudograph (G)).

The k th stochastic total [$f_k(x)$ and $f_k^H(x)$] and local [$f_{kL}(x_E)$, $f_{kL}^H(x_E)$ and $f_{kL}^H(x_{E-H})$] linear indices were also computed.

2.2. Atom, atom-type, and total non-stochastic linear indices

Making use of the subprogram CARDD implemented in the TOMOCOMD software, the non-stochastic linear indices can be calculated for the ‘molecular pseudograph’s atom adjacent matrix’ for small-to-medium sized organic compounds. The theoretic aspects concerning to these indices have been explained in some detail elsewhere.¹⁷ However, an overview of this approach will be given in the current paper.

For a given molecule composed of n atoms, the ‘molecular vector’ (X) is constructed and the k th atom linear indices, $f_k(x_i)$, are calculated as linear maps on $\Re^n[f_k(x_i) : \Re^n \rightarrow \Re^n]$ as shown in Eq. 2,

$$f_k(x_i) = \sum_{j=1}^n {}^k a_{ij} X_j \quad (2)$$

where, ${}^k a_{ij} = {}^k a_{ji}$ (symmetric square matrix), n is the number of atoms of the molecule, and X_1, \dots, 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 numeric values, which can be considered as weights (atom labels) of the vertices of the molecular pseudograph. Certain atomic properties (electronegativity, density, atomic radii, etc.) can be used with this propose. In this work Paulin electronegativities are selected as atom weights.²⁴

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

$$\begin{aligned}
 a_{ij} &= P_{ij} & \text{if } i \neq j & \text{ and } \exists e_k \in E(G) \\
 &= L_{ii} & \text{if } i = j \\
 &= 0 & \text{otherwise}
 \end{aligned} \quad (3)$$

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 (see Table 1).

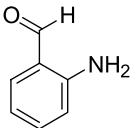
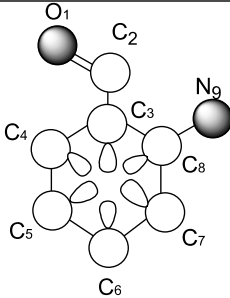
The atom linear indices are defined as a linear transformation $f_k(x_i)$ on a molecular vector space \mathfrak{R}^n . The defined equation 2 for $f_k(x_i)$ may be written as the single matrix equation:

$$f_k(x_i) = [X']^k = M^k[X] \quad (4)$$

where $[X]$ is a column vector (a $n \times 1$ matrix) of the coordinates of X in the canonical basis of \mathfrak{R}^n and M^k is the k th power of the matrix M of the molecular pseudograph (map's matrix).

This approach is rather similar to the **LCAO–MO** (Linear Combinations of Atomic Orbitals–Molecular Orbitals) method. Really, our approach (for $k=1$) is a quite similar approximation to the extended Hückel MO method, due to the formalism each MO ψ_i is composed of n valence AOs of atoms in a molecule.²⁵

Table 1. Definition and calculation of total (whole molecule) and local (atom) linear indices of the molecular pseudograph's atom adjacency matrix of the 2-aminobenzaldehyde molecule

		$X = [O_1, C_2, C_3, C_4, C_5, C_6, C_7, C_8, N_9]$ Molecular Vector: $X \in \mathfrak{R}^9$ In the definition of the X , as molecular vector, the chemical symbol of the element is used to indicate the corresponding electronegativity value. That is: if we write O it means $\chi(O)$, oxygen Paulin electronegativity or some atomic property, which characterizes each atom in the molecule. Therefore, if we use the canonical basis of \mathfrak{R}^9 , the coordinates of any vector X coincide with the components of that molecular vector [X] = [3.17, 2.63, 2.63, 2.63, 2.63, 2.63, 2.63, 2.63, 2.33] [X]: column vector of coordinates of X in the Canonical base of \mathfrak{R}^9 (a $n \times 1$ matrix)				
Molecular Structure	Molecular Pseudograph (G) (hydrogen suppressed-pseudograph)					
$f_1(x_i) = \sum_{j=1}^n a_{ij} X_j = \mathbf{M}^1[X] = \begin{bmatrix} 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 2 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} O_1 \\ C_2 \\ C_3 \\ C_4 \\ C_5 \\ C_6 \\ C_7 \\ C_8 \\ N_9 \end{bmatrix} = \begin{bmatrix} 2C_2 \\ 2O_1 + 1C_3 \\ 1C_2 + 1C_3 + 1C_4 + 1C_8 \\ 1C_3 + 1C_4 + 1C_5 \\ 1C_4 + 1C_5 + 1C_6 \\ 1C_5 + 1C_6 + 1C_7 \\ 1C_6 + 1C_7 + 1C_8 \\ 1C_3 + 1C_7 + 1C_8 + 1N_9 \\ 1C_8 \end{bmatrix}$						
Atom linear indices of first order is a <i>linear map</i> ; $f_1(x_i): \mathfrak{R}^n \rightarrow \mathfrak{R}^n$ such that, $f_1(O_1, C_2, C_3, C_4, C_5, C_6, C_7, C_8, N_9) = (2C_2, 2O_1 + 1C_3, 1C_2 + 1C_3 + 1C_4 + 1C_8, 1C_3 + 1C_4 + 1C_5, 1C_4 + 1C_5 + 1C_6, 1C_5 + 1C_6 + 1C_7, 1C_6 + 1C_7 + 1C_8, 1C_3 + 1C_7 + 1C_8 + 1N_9, 1C_8) = (5.26, 8.97, 10.52, 7.89, 7.89, 7.89, 7.89, 10.22, 2.63)$ and whole-molecule linear indices of first order is a <i>linear functional</i> ; $f_1(x) = \sum_{i=1}^n f_1(x_i) = f_1(O_1) + f_1(C_2) + f_1(C_3) + f_1(C_4) + f_1(C_5) + f_1(C_6) + f_1(C_7) + f_1(C_8) + f_1(N_9) = 69.16$						
<i>Local and total linear indices of order 0-5 (k = 0-5)</i>						
Atom (i)	$f_0(x_i)$	$f_1(x_i)$	$f_2(x_i)$	$f_3(x_i)$	$f_4(x_i)$	$f_5(x_i)$
O ₁	3.17	5.26	17.94	42.08	146.96	400.72
C ₂	2.63	8.97	21.04	73.48	200.36	676.25
C ₃	2.63	10.52	37.6	116.2	382.33	1193.57
C ₄	2.63	7.89	26.3	87.57	277.41	894.29
C ₅	2.63	7.89	23.67	73.64	234.55	739.87
C ₆	2.63	7.89	23.67	73.34	227.91	721.81
C ₇	2.63	7.89	26	80.93	259.35	820.73
C ₈	2.63	10.22	31.26	105.08	333.47	1080.23
N ₉	2.33	2.63	10.22	31.26	105.08	333.47
Total	23.91	69.16	217.7	683.58	2167.42	6860.94

Total (whole molecule) linear indices are linear functionals (some mathematicians use the term linear form, which means the same as linear functional) on \mathfrak{R}^n . The mathematical definition of these molecular descriptors is the following:

$$f_k(x) = \sum_{i=1}^n f_k(x_i) \quad (5)$$

where n is the number of atoms and $f_k(x_i)$ are the atom's linear indices (linear maps) obtained by Eq. 2. Then, a linear form $f_k(x)$ can be written in matrix form,

$$f_k(x) = [u]^t [X']^k \quad (6)$$

or

$$f_k(x) = [u]^t M^k [X] \quad (7)$$

for each molecular vector $X \in \mathfrak{R}^n$. $[u]^t$ is a n -dimensional unitary row vector. As can be seen, the k th total linear index is calculated by summing the local (atom) linear indices of all atoms in the molecule.

In addition to atom linear indices computed for each atom in the molecule, a local-fragment (atom-type) formalism can be developed. The k th atom-type linear index of the molecular pseudograph's atom adjacency matrix is calculated by summing the k th atom linear indices of all atoms of the same atom type in the molecule. Consequently, if a molecule is partitioned in Z molecular fragments, the total linear indices can be partitioned in Z local linear indices $f_{kL}(x)$, $L = 1, \dots, Z$. The total linear indices of order k can be expressed as the sum of the local linear indices of the Z fragments of the same order:¹⁷

$$f_k(x) = \sum_{L=1}^Z f_{kL}(x) \quad (8)$$

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

2.3. Atom, atom-type, and total stochastic linear indices

The linear indices' matrices, M^k , are graph-theoretic electronic-structure models, like an 'extended Hückel MO model'. The M^1 matrix considers all valence-bond electrons (σ - and π -networks) in one step and their power ($k = 0, 1, 2, 3, \dots$) can be considering as an interacting-electron chemical-network model in k step. This model can be seen as an intermediate between the quantitative quantum-mechanical Schrödinger equation and classical chemical bonding ideas.²⁵

The present approach is based on a simple model for the intramolecular movement of all outer-shell 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 atom nucleus. Alternatively, these electrons can be distributed around cores in discrete intervals of time t_k . In this sense, the electron in an arbitrary atom i can move to other atoms at different discrete time periods t_k ($k = 0, 1, 2, 3, \dots$) throughout the chemical-bonding network.

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

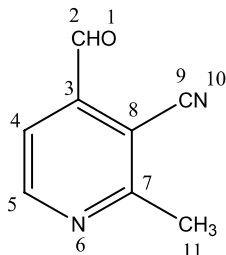
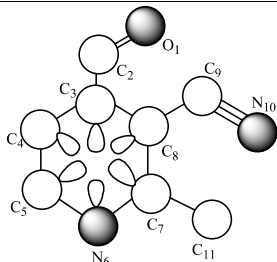
$${}^k s_{ij} = \frac{{}^k a_{ij}}{{}_k \text{SUM}_i} = \frac{{}^k a_{ij}}{{}_k \delta_i} \quad (9)$$

where, ${}^k a_{ij}$ are the elements of the k th power of M and the SUM of the i th row of M^k are named the k -order vertex degree of atom i , ${}_k \delta_i$. The ${}^k s_{ij}$ elements are the transition probabilities with the electrons move from atom i to j in the discrete time periods t_k . Note, that k th element s_{ij} take into consideration the molecular topology in k step throughout of the chemical-bonding (σ - and π -) network. For instance, the ${}^2 s_{ij}$ values can distinguish between hybrid states of atoms in bonds. In this sense, It can clearly be seen from Table 2 that electrons will have a higher probability of returning to the sp nitrogen ($N_{10} = 0.75$) than to the sp₂ nitrogen ($N_6 = 0.33$) in t_2 . A similar behavior can be observed among the different hybrid states of carbon atom in the molecule of 2-formyl-6-methyl-benzonitrile (see Table 2): Csp₃ ($C_{11} = 0.25$); Csp₂ ($C_2 = 0.625$); Csp_{2arom} ($C_3 = 0.285$, $C_4 = 0.3$, $C_5 = 0.33$, $C_7 = 0.33$, $C_8 = 0.25$), and Csp ($C_9 = 0.769$). This is a logical result if the electronegativity scale of these hybrid states is taken into account. The k th total [and local (atom and atom type) stochastic linear indices], ${}^s f_k(x)$ [${}^s f_k(x_i)$] are calculated in the same way that the linear indices (non-stochastic), but using k th stochastic molecular pseudograph's atom adjacency matrix, $S^k(G)$, like mathematical linear maps' matrices (Table 2).

2.4. Developing classification functions

The use of linear discriminant analysis (LDA) in rational drug design has become in an important tool for the prediction of chemicals properties.^{26–30} In the current work we make use of the advantages of TOMOCOMD strategy to developing classification models including linear stochastic and non-stochastic descriptors. As first step of this methodology, the selection of a training data set composed of antimalarials and inactive compounds was performed. The quality of the classification models is highly dependent on the quality of this data set. For this reason, an extensive search of compounds with great structural variability was carried out. A general data set made up of 1562 compounds, 597 with antimalarial properties and different action modes, and 965 having other clinical uses (antivirals, sedative/hypnotics, diuretics, anticonvulsivants, hemostatics, oral hypoglucemics, antihypertensives, antihelmintics, anticancer compounds, and so on)^{2,7–9,31–72} was conformed. From these 1562 compounds, 1120 were

Table 2. Calculation of $M^k(G)$ and $S^k(G)$ for 2-formyl-6-methyl-benzonitrile when k varies from 0 to 2 and i is a specific atom in the molecule

																									
Molecular Structure												Molecular Pseudograph (G)													
a_{ij}	O ₁	C ₂	C ₃	C ₄	C ₅	N ₆	C ₇	C ₈	C ₉	N ₁₀	C ₁₁	k	δ_i	O ₁	C ₂	C ₃	C ₄	C ₅	N ₆	C ₇	C ₈	C ₉	N ₁₀	C ₁₁	
$M^0(G)$													$S^0(G)$												
O ₁	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	
C ₂	0	1	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	
C ₃	0	0	1	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	
C ₄	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	
C ₅	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	
N ₆	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	
C ₇	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	
C ₈	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	
C ₉	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	
N ₁₀	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	1	0	0	
C ₁₁	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	1	
$M^1(G)$													$S^1(G)$												
O ₁	0	2	0	0	0	0	0	0	0	0	0	2	0	1	0	0	0	0	0	0	0	0	0	0	
C ₂	2	0	1	0	0	0	0	0	0	0	0	3	0.66	0	0.33	0	0	0	0	0	0	0	0	0	
C ₃	0	1	1	1	0	0	0	1	0	0	0	4	0	0.25	0.25	0.25	0	0	0	0	0.25	0	0	0	
C ₄	0	0	1	1	1	0	0	0	0	0	0	3	0	0	0.33	0.33	0.33	0	0	0	0	0	0	0	
C ₅	0	0	0	1	1	1	0	0	0	0	0	3	0	0	0	0.33	0.33	0.33	0	0	0	0	0	0	
N ₆	0	0	0	0	1	1	1	0	0	0	0	3	0	0	0	0	0.33	0.33	0.33	0	0	0	0	0	
C ₇	0	0	0	0	0	1	1	1	0	0	1	4	0	0	0	0	0	0.25	0.25	0.25	0.25	0	0	0.25	
C ₈	0	0	1	0	0	0	1	1	1	0	0	4	0	0	0.25	0	0	0	0.25	0.25	0.25	0.25	0	0	
C ₉	0	0	0	0	0	0	0	1	0	3	0	4	0	0	0	0	0	0	0	0.25	0	0.75	0	0	
N ₁₀	0	0	0	0	0	0	0	0	3	0	0	3	0	0	0	0	0	0	0	0	1	0	0	0	
C ₁₁	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	
$M^2(G)$													$S^2(G)$												
O ₁	4	0	2	0	0	0	0	0	0	0	0	6	0.66	0	0.33	0	0	0	0	0	0	0	0	0	
C ₂	0	5	1	1	0	0	0	1	0	0	0	8	0	0.625	0.125	0.125	0	0	0	0.125	0	0	0	0	
C ₃	2	1	4	2	1	0	1	2	1	0	0	14	0.143	0.071	0.287	0.143	0.071	0	0.071	0.143	0.071	0	0	0	
C ₄	0	1	2	3	2	1	0	1	0	0	0	10	0	0.1	0.2	0.3	0.2	0.1	0	0.1	0	0	0	0	
C ₅	0	0	1	2	3	2	1	0	0	0	0	9	0	0	0.111	0.222	0.333	0.222	0.111	0	0	0	0	0	
N ₆	0	0	0	1	2	3	2	1	0	0	1	10	0	0	0	0.1	0.2	0.3	0.2	0.1	0	0	0	0.1	
C ₇	0	0	1	0	1	2	4	2	1	0	1	12	0	0	0.083	0	0.083	0.166	0.333	0.166	0.083	0	0.083	0	
C ₈	0	1	2	1	0	1	2	4	1	3	1	16	0	0.063	0.125	0.063	0	0.063	0.125	0.25	0.063	0.188	0.063	0	
C ₉	0	0	1	0	0	0	1	1	10	0	0	13	0	0	0.077	0	0	0	0.077	0.077	0.769	0	0	0	
N ₁₀	0	0	0	0	0	0	0	3	0	9	0	12	0	0	0	0	0	0	0	0.25	0	0.75	0	0	
C ₁₁	0	0	0	0	0	1	1	1	0	0	1	4	0	0	0	0	0	0.25	0.25	0.25	0	0	0.25	0	

chosen at random to be included in the training set, being 437 of them actives and 683 inactive ones. The resting group composed of 160 antimalarials and 282 compounds with different biological properties were never used in the development of the classification models. In this sense, this conjunct was prepared as test data set for the external cross-validation of the models.

Making use of the LDA technique implemented in the STATISTICA software,⁷³ the following classification models were obtained:

$$\begin{aligned}
 \text{Class} = & -8.91067 + 0.49305f_0^H(x) \\
 & - 0.1002f_2^H(x) + 0.35985f_1(x) \\
 & - 0.00946f_3(x) - 0.40409f_{1L}(x_{E-H}) \\
 & + 3.1135 \times 10^{-5}f_{10L}(x_{E-H}) \\
 & - 0.12309f_{3L}^H(x_E) + 0.0033f_{5L}^H(x_E) \\
 N = & 1120 \quad \lambda = 0.35 \quad D_2 = 7.92 \\
 F(8, 1111) = & 261.61 \quad p < 0.0000
 \end{aligned} \tag{10}$$

$$\begin{aligned}
\text{Class} = & -8.3859 + 0.96266^s f_0^H(x) \\
& + 2.0570^s f_1^H(x) - 3.0412^s f_2^H(x) \\
& - 0.5636^s f_4(x) + 0.6599^s f_{15}(x) \\
& - 2.6885^s f_{1L}(x_{E-H}) + 2.8129^s f_{14L}(x_{E-H}) \\
& + 1.5025^s f_{0L}^H(x_E) - 1.4299^s f_{15L}^H(x_E) \\
N = & 1120 \quad \lambda = 0.38 \quad D_2 = 6.9 \\
F(9, 1110) = & 202.73 \quad p < 0.0000 \quad (11)
\end{aligned}$$

where, N is the number of compounds, λ is Wilk's coefficient, F is the Fisher ratio, D^2 is the squared Mahalanobis distance and p -value is the significance level. The antimalarial activity was codified by a dummy variable 'Class', which indicates either the presence of an active compound (Class = 1) or an inactive one (Class = -1).

This first model (Eq. 10) classified correctly 92.45% (positive predictive value) of compounds with antimalarial activity and the 95.02% (negative predictive value) of compounds without the desired property (inactive ones). The global good classification value for the data set was 94.02% (accuracy). This model showed a high Matthews' correlation coefficient (MCC) of 0.87. MCC quantified the strength of the linear relation between the molecular descriptors and the classifications.⁷⁴ These result and the two most commonly used *operating characteristics* of 'diagnostic' tests (sensitivity and specificity) are depicted in Table 3. Also in this Table appear the results for model 11. In this case 91.52% of accuracy was obtained. This function, which includes stochastic linear indices, showed 91.52% of global good classification and a Matthews Correlation Coefficient of 0.82.

A discriminant model could be accepted or rejected depending on its predictive power. This important characteristic can be tested carrying out internal or external validation processes. In our case, an external prediction data set was used for this purpose. In this sense, the calculation of percentages of global good classification (accuracy), sensibility, specificity, positive, and negative predictive values and Matthews correlation coefficient in all validation experiments permitted us to carry out the assessment of the models. In Table 3 are also depicted the results of the validation process using the external prediction data set.

The classification of cases was performed by means of the posterior classification probabilities. This is the probability that the respective case belongs to a particu-

lar group (active or inactive). By using the models, each compound can be then classified as active, if $\Delta P\% > 0$, being $\Delta P\% = [P(\text{active}) - P(\text{inactive})] \times 100$ or as inactive otherwise. The classification results (for some compounds in both training and test sets) using model 10 and 11 are shown in Tables 4 and 5. The complete set of compounds in training and prediction series is given as Supplementary data (see Tables 4.1, 4.2, and 5.1).

2.5. TOMOCOMD-CARDD method versus other cheminformatic approaches

In the last decade, two 'in silico' methods have been used to develop structure-based classification models on antimalarial activity, which give rise to a good discrimination of this activity in large and heterogeneous series of organic compounds.⁶ In the current work we developed also two models for the discrimination of compounds having this kind of activity. As before expressed we pretend to compare both approaches. 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 methodology reported previously⁶ and the current approach. However, a relative comparison could be based on the kind of method used for deriving the QSAR and their statistical parameter, the explored molecular descriptors, the number and diversity of chemical structural patterns contained in the data, the overall accuracy (%), and the method which was used for the validation of the models. Table 6 shows all these parameters for both approaches.

The global good classification in the training set of TOMOCOMD-CARDD models (Eq. 10 = 94.02% and Eq. 11 = 91.52%) was higher than the two reported LDA equations (see Table 6). 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

On the other hand, Golbraikh and Tropsha emphasize recently that the predictive ability of a QSAR model can only be estimated using an external test set (external validation) of compounds that was not used for building the model and formulated a set of criteria for evaluation of predictive ability of QSAR model.⁷⁵ All LDA-QSAR models reported were successfully validated by means of external prediction series. In this sense, the overall accuracy in test sets of TOMOCOMD-CARDD models

Table 3. Overall measures of accuracy obtained in the training and prediction sets for the models 10 and 11

	Matthews corr. coefficient	Accuracy (%)	Sensitivity (%)	Specificity (%)	Predictive value (+) (%)	Predictive value (-) (%)
<i>Training set</i>						
Ec.1	0.87	94.02	92.23	95.16	92.45	95.02
Ec.2	0.82	91.52	89.40	92.86	88.79	93.27
<i>Test set</i>						
Ec.1	0.86	93.42	91.15	94.66	90.63	95.03
Ec.2	0.79	90.50	88.82	91.40	84.38	93.97

Table 4. Classification of active compounds in the training and test sets by the use of LDA–QSAR models developed using non-stochastic (Eq. 10) and stochastic (Eq. 11) total and local linear indices

Name	$\Delta P\%_a$	$\Delta P\%_b$	Name	$\Delta P\%_a$	$\Delta P\%_b$
<i>Training active group</i>					
Cinchonine	−64.88	−90.20	PS-15	92.86	55.35
Hydroxychloroquine	83.96	−2.84	Artemisinin	84.14	96.59
CDRI 87209	31.66	−29.61	Bispyroquine	99.69	58.94
WR 33,063	100.00	99.80	Fluornemethanol	72.51	97.19
WR 122,455	43.04	37.79	Artemisitene	67.70	93.35
Pamaquine	88.66	50.61	Lapinone	99.98	99.97
Primaquine	−52.34	−56.04	Benzonaphthyridine 7351	99.99	97.66
WR 225498	97.38	98.09	12278 R	99.69	89.75
WR 242511	97.59	95.10	Secoartemisinin	91.60	98.07
M8506, Trifluoracetylprimaquine	88.53	−28.35	(+)-4,5-Secoartemisinin	89.91	98.61
CDRI 8053	97.63	88.40	9-Desmethylartemisinin	57.20	72.50
Chloroguanide	−42.15	−90.51	6,9-Didesmethylartemisinin	57.20	72.50
Chloroproguanil	21.64	−69.14	B-artether	98.54	97.27
Pyrimetamine	32.38	16.56	Gossypol	99.99	99.99
Trimethoprim	74.22	48.24	CN 10443	99.98	99.97
Sulfalene	88.67	−76.03	Sodium artesunate	98.49	99.36
Dapsone	−47.39	−80.77	6-Demethyl-6-difluoromethyl B-artether	99.88	99.59
Sulfisoxazole	73.58	21.99	Tripiperakin	99.50	84.61
WR 99210	90.11	94.87	Nitroguanil	−99.09	−84.96
Menctone	98.23	97.14	Fluoroquine	89.22	79.32
Pyronaridine	99.95	91.02	Pentaquine	74.36	0.07
Quinacrine	98.00	92.95	Dabekhin	−71.43	−72.58
Amodiaquine	96.15	34.22	Methylchloroquine	92.33	78.46
Tetracycline	97.40	98.87	RC-12	97.71	90.98
Doxycycline	95.74	98.53	Dimeplasmin	91.29	78.14
Clindamycin	99.77	99.15	Azamepacrine	99.40	96.77
Ciprofloxacin	−78.24	39.62	Mepacrine	98.83	93.19
Yingzhaosu A	96.01	91.42	Aristochin	99.70	99.39
Yingzhaosu C	5.35	64.68	Axonitrile-3	−52.64	41.52
Arteflene	98.04	91.27	Berbamine	99.07	99.89
Fenozan-50 F	92.68	98.93	Malagashanine	15.53	−28.40
Chalcone	−92.17	−50.42	Berberine	−94.49	−49.82
Exifone	95.58	60.64	10,12 Peroxycalamenene	−28.75	67.49
Methylene blue	−98.11	−80.10	Simalikalactone D	99.76	99.87
WR 197236	77.43	93.35	Gutolactone	99.48	99.71
Piperaquine	93.21	72.00	Lissoclinotoxin A	35.28	−66.20
Desferrioxamine	99.81	99.99	Licochalcone A	−41.54	61.31
2-(4-Methoxybenzoyl)-1-naphthoyl acid	−98.95	−72.24	Oxalic bis(2-hydroxy-1-naphthylmethylene)hydrazide	−64.06	−53.98
Buquinolate	99.53	99.44	Decoquinat	99.98	99.84
Methyl benzoate	77.31	80.48	Cycloleucine	−98.99	−94.85
Cloguanil	−99.44	−95.12	Supazine	−61.64	−61.77
Metachloridine	86.39	−74.99	Cilional	30.40	−61.95
WR 10 488	29.71	61.65	Acedapsone	2.22	−38.26
Gentiopicrocin	−2.47	−12.10	Antimalarine	49.99	−17.80
Oxychlorochin	49.97	−37.02	Brindoxime	60.08	86.76
Aminopterin	99.87	98.30	Amquinate	40.72	85.27
12-(3'-Hydroxy- <i>n</i> -propyl)-deoxoartemisinin	98.85	97.28			
<i>Test active group</i>					
Quinine	−9.22	−65.72	B-Artemether	96.39	94.12
Chloroquine	86.68	44.23	Artelinic acid	98.93	98.81
Mefloquine	94.83	57.15	Cycloquin	100.00	99.70
Halofrantine	94.93	97.73	7,7 Difluoro-B-artether	99.83	99.74
Quinocide	−46.79	−69.71	Naphthol blue-black	99.98	99.81
WR 238,605	99.31	99.53	9-Epiartemisinin	73.30	91.49
WR182393	82.92	70.03	Lapacol	−85.29	−1.18
Cycloguanil	−94.38	−64.08	Tebuquine	97.24	82.62
Sulfadoxine	97.03	−40.89	Octanoylprimaquine	99.66	98.32
Sulfamethoxazole	58.11	−48.07	Brusatol	99.87	99.95
Clociguanil	−43.44	50.86	Floxacrine	51.48	84.21
Nitroquine	55.25	66.58	Aecachinium	19.99	21.43
Dihydroartemisinin	89.13	87.35	Methotrexate	99.33	98.80
Iso-artemisitene	82.79	88.62	Apicidin	100.00	100.00
Hydrolapachol	−61.40	29.98	Dioncophyline B	61.96	69.61

(continued on next page)

Table 4 (continued)

Name	$\Delta P\%^a$	$\Delta P\%^b$	Name	$\Delta P\%^a$	$\Delta P\%^b$
Atovaquone	−55.47	−36.17	Strychnobrasiline	−56.34	−42.27
Amopyroquine	94.45	−26.57	Hexalorxylol	−74.87	−97.05
Azithromycin	100.00	100.00	WR 135 403	56.88	−19.44
Norfloxacin	−86.47	61.65	WR 226 253	93.35	64.91
Enpiroline	95.56	62.22	CI-608	95.33	97.68
Refigallo	98.56	93.71	Endochin	75.26	81.45
WR 194905	99.85	99.95	Isopentachin	78.60	20.34

^a Results of the classification of compounds obtained from Eq. 10 (using non-stochastic linear indices).

^b Results of the classification of compounds obtained from Eq. 11 (using stochastic linear indices): $\Delta P\% = [P(\text{active}) - P(\text{inactive})] \times 100$.

Table 5. Classification of inactive compounds in the training and test set by the use of LDA–QSAR models developed using non-stochastic (Eq. 10) and stochastic (Eq. 11) total and local linear indices

Name	$\Delta P\%^a$	$\Delta P\%^b$	Name	$\Delta P\%^a$	$\Delta P\%^b$
<i>Training inactive group</i>					
Amantadine	−91.85	−98.78	Tiazesin hydrochloride	−80.81	−90.66
Glycerol	−99.81	−99.79	5-fluorocitosine	−98.35	−89.41
Triacetonamine	−72.62	−60.11	Benzotef	−99.58	−98.58
Hemedin	−98.49	−98.74	Glucin	−84.81	−97.25
Plegarol	−43.80	−72.34	Hexestrol	−44.98	−52.07
Methioflurane	−89.18	−98.10	Mtrafazoline	−95.25	−83.31
Dimepranol	−99.24	−99.33	Perithiaden	−98.79	−93.07
Alarmino	−96.93	−98.77	Caracemide	−99.84	−68.30
Pildralazine	−90.27	−72.43	Chloromethylsilatrane	−99.56	−99.64
Supazine	−61.64	−61.77	Ericolol	35.34	50.55
Ketoxal	−95.54	−96.67	Dimetridazole	−98.66	−95.45
Citenazone	−99.63	−98.10	Hachimycin	7.27	10.21
Disotat-Amp	−97.46	−97.03	Guanazole	−100.00	−99.88
Guancidine	−94.26	−85.64	Carbetimer	−99.87	−99.33
Olmidine	−96.11	−97.04	Leucenol	−97.40	−82.20
Moxonidine	11.00	−42.55	Basidalin	−97.30	−96.30
Methylthiouracil	−95.03	−96.07	Clodanolene	−77.46	−8.83
Tolazamide	−26.03	−47.60	O-acetilsalicylamide	−98.86	−92.21
Metadiphenii bromidum	−92.13	−72.46	Chinoi-401	21.20	7.28
Benfosformin	−99.99	−99.98	Mebenformin	−99.88	−99.46
Urethane	−99.97	−99.55	Fentanilo	−73.53	−80.18
Fencadium	−27.12	−95.98	Clorotepine	−69.92	−90.10
Styptol	−88.50	−87.15	Cianothepin	−94.05	−91.09
Hydroxindasate	−91.37	47.23	Azabuperone	3.05	−23.31
Vernelan	−99.90	−98.03	Morfina	−42.12	−84.79
Sedanfactor Solucion	−99.92	−98.93	Petidina	−91.40	−85.49
Bromothiazide	32.83	−96.78	Amifostine	−100.00	−99.95
Geroquinol	19.25	−11.73	Fosforotioic acid	−99.70	−98.18
Methonal	−99.45	−94.49	Bextrometorfano	−58.97	−68.91
Clomethiazole	−97.59	−97.74	Alylprodine	−59.72	−34.24
Fenaclon	−96.00	−96.25	Tironamine	−97.39	−88.76
Dicumarol	−68.47	60.72	Metaraminol	−97.54	−97.10
Methylpentynol	−99.15	−98.08	Dezocine	−56.14	−60.95
Carbromide	−97.25	−87.51	Diampromide	−64.78	−36.77
Cyclopropane	−99.74	−99.77	Racefemine	−70.03	−89.94
Norantoin	−99.60	−96.83	Bamipine	−90.32	−96.28
Paraldehyde	−95.66	−98.05	Dipipanone	−25.58	10.90
Hedonal	−99.62	−94.97	Hydroxypethidine	−65.95	−67.84
Aminoglutethimide	−78.79	−62.89	Difenilhidramine hydrochloride	−96.00	−98.04
Carbavin	−99.94	−98.43	Bromazine	33.40	−67.61
Chlorphenacemide	−99.81	−98.11	Medrylamine	−85.46	−92.25
Ethchlorvynol	−96.72	−96.71	Meptazinol	−63.94	−78.08
Fertaron	−99.69	−97.40	Mapyroxal	−99.92	−99.66
Ferocal	−99.24	−96.59	Penferon	−94.78	−92.36
Phenolphthalol	−93.63	−94.45	Nicoclonate	−69.52	−25.91
Imekhin	−89.50	−70.97	Tuaminoheptane	−97.53	−96.29
Dimecamine	−83.55	−59.36	Tyrosam	−99.62	−99.06
Cryofluorane	−85.64	−99.98	Xylazine	−89.33	−84.88

Table 5 (continued)

Name	$\Delta P\%$ ^a	$\Delta P\%$ ^b	Name	$\Delta P\%$ ^a	$\Delta P\%$ ^b
Tribromoethanol	−98.93	−99.77	Phedrazine	−63.75	−70.46
Ethyl bromide	−99.49	−99.35	Naphazoline	−99.73	−99.02
Isopryl	−98.74	−98.99	Meprobamate	−99.80	−34.39
Anaesthaminol	−69.86	−84.33	Guaifenesin	−93.43	−96.99
Butamben	−78.41	−81.61	Strychnocarpine	−80.03	−93.81
Thiophenobarbital	−97.83	−78.87	Benzyllephedrine	−94.47	−96.97
Alloxanthine	−32.43	−91.73	Isoprenaline	−36.93	−81.81
Auxinutril	−99.50	−99.38	Mephentermine	−98.34	−97.23
Pivalylindandione	−93.91	30.05	Octodrine	−95.27	−93.40
Cetovex	−96.75	−77.28	Tiopronin	−99.34	−95.28
Nitrodimethylin	−92.91	−99.26	Ordenina	−97.60	−98.24
Vincosfos	−55.19	−67.29	KC−8973	−86.03	19.71
Nitrodimethylin	−92.91	−99.26	Ordenina	−97.60	−98.24
<i>Test inactive group</i>					
DCF, BW 683c	−67.78	−80.83	Hydroxyamfetamine	−98.85	−98.19
Phenbutamide	−86.02	−80.16	Amidefrini mesilas	−96.48	−98.22
Dimepheptanol	−77.46	−62.48	Canfochinid	−9.22	−65.72
Dimetrizadole	−98.54	−86.47	Phenylephrine	−97.61	−98.60
Guanoxabenz	−92.68	−85.91	Tymazoline hydrochloride	−87.29	−64.56
Ciproximide	−97.27	−89.60	Fenoverine	−53.14	−48.76
Clofibrate	−52.15	−9.38	Apoatropine	−85.72	−74.30
Urefibrate	−86.49	−24.16	Caroverine	−7.69	51.22
Colestipol	−99.98	−99.69	Cititolone	−98.27	−97.17
Dextromoramide	−13.54	−53.41	Flusoxolol	94.76	69.79
Cyclocholine tosylate	−99.85	−99.86	Nafomine	−99.92	−98.50
Calcii diethylacetate	−98.51	−94.80	Thiosalicylic acid	−98.74	−99.38
Cobalti besilas	−99.81	−99.77	Hadacidin	−99.87	−99.88
Iprazochrome	−87.20	−29.27	Imifos	−97.17	−95.31
Lemidosul	−83.11	17.05	Spiroplatin	−99.35	−97.04
Sulclamide	−75.03	−91.18	Xylose	−98.64	−99.46
Atrolactamide	−99.29	−96.78	3-Hydroxyacetanilide	−93.55	−97.99
Aloe–Emodin	−51.45	9.85	Propanoic acid	−99.85	−99.39
Cintramide	−91.24	−75.47	Hydramitrazine	14.26	86.77
Sodium dipantoylferrate	−98.49	−93.72	Fluoxypilin	−96.30	−96.06
Metiapine	−77.76	−89.13	Flavoxate	−25.16	82.65
Gaplegin	−82.16	−33.56	Promoxolane	−49.41	−24.81
Clioquinol	−73.37	−78.02	Betaxolol	55.17	−56.27
Hidrocodona	−46.04	33.91	Bornaprolol hydrochloride	19.57	−47.40
Meflophenhidramine	−74.89	−33.34	Arphamenine A	−97.24	−75.58
Tioxidazole	−90.74	−66.73	Oxyephedrine	−94.37	−96.44

^a Results of the classification of compounds obtained from Eq. 10 (using non-stochastic linear indices).

^b Results of the classification of compounds obtained from Eq. 11 (using stochastic linear indices): $\Delta P\% = [P(\text{active}) - P(\text{inactive})] \times 100$.

(Eq. 10 = 93.42% and Eq. 11 = 90.50%) was higher than the rest of two reported LDA equations (see Table 6).

Another remarkable aspect is referred 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 (see Tables 4.1 and 5.1 in Supplementary data to obtain the complete list of 597 antimalarial compounds used in training and test sets).

3. Concluding remarks

The search of effective and rational methodologies for the discovery of new drugs has become a first-line objective for the pharmaceutical research. The introduction of

graph theoretical descriptors for rational drug design (or selection) has turned into an attractive cheminformatic tool in this area. With the present work, we have shown that TOMOCOMD-CARDD approach can be applied to generate useful quantitative models for the classification of antimalarials, taking into account the information obtained from a training data set of compounds with a considerable structural variability. The refereed methodology permits a quick in silico discovery of new candidates to lead compounds using a minimum of resources and reducing the degree of uncertainty of the process. The collected data of active compounds used in this study, constitutes a useful tool for the work in this area.

The interactive and flexible character of the TOMOCOMD-CARDD approach permits the future inclusion of new antimalarial drugs in the training data set and the generation of models considering more structural information

Table 6. Comparison between TOMOCOMD-CARDD method and others cheminformatic approaches for antimalarial activity

Models' features to be compared ^a	Structure-based classification models of antimalarial activity			
	Eq. 10	Eq. 11	Eq. 12	Eq. 13
<i>N</i> Total	1562	1562	59	60
<i>N</i> Antimalarials	597	597	25	25
Technique ^b	LDA	LDA	LDA	LDA
Wilks' λ (<i>U</i> -statistics)	0.35	0.38	0.55	0.35
<i>F</i>	261.61	202.73	9.83	8.88
<i>D</i> ²	7.92	6.90	—	—
<i>p</i> -Level	<0.000	<0.000	—	—
Variables in the model	8	9	3	8
<i>Training set</i>				
<i>N</i> Total	1120	1120	41	45
<i>N</i> Antimalarials	437	437	17	19
Accuracy (%)	94.02	91.52	82.92	91.11
Families of drugs ^c	Broader range	Broader range	Low range	Low range
<i>Test set</i>				
<i>N</i> Total	442	442	20	21
<i>N</i> Antimalarials	160	160	10	8
Predictability (%)	93.42	90.50	88.88	60.00
Families of drugs	Broader range	Broader range	Low range	Low range

^a Eqs. 10 and 11 are reported in this work, models 12 and 13 were reported by Gozalbez et al.⁶ for two different studies: Eq. 12 was performed for the classification of antimalarial drugs and non-antiprotazoan drugs and, Eq. 13 for the discrimination between antimalarials and antiprotazoan drugs without antimalarial activity.

^b LDA refers to Linear discriminant analysis.

^c Only largely represented families were considered.

such as 3D features. However, this point, which will be the objective of a forthcoming paper, is out of the general scope of the present work.

4. Experimental

4.1. Computational approach

Calculations were carried out on a PC PENTIUM-4 2.0 GHz. The CARDD-module implemented in the TOMOCOMD software¹⁵ was used to the calculation of total and local non-stochastic and stochastic linear indices for the data set of organic molecules. In all cases, Pauling electronegativities²⁴ were employed as atomic weights (molecular vector's components).

4.2. Chemometric method

The linear discriminant analysis was performed with the STATISTICA 5.5 for WINDOWS package.⁷³ The tolerance parameter (proportion of variance that is unique to the respective variable) used 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.⁷⁶ In this connection; we select the model with higher statistical significance but having as few parameters (a_k) as possible. The quality of the models were determined by examining Wilk's 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 training and test sets. Models with a proportion between the number of cases and variables in the equation lower than 4 were rejected. The statistical robustness

and predictive power of the obtained model was assessed using an external prediction (test) set.

Supplementary data

The complete list of compounds used in training and prediction sets, as well as their structures and posterior classification according to model 10 and 11 is available via the Internet at <http://www.sciencedirect.com>. Supplementary data associated with this article can be found in the online version at [doi:10.1016/j.bmc.2004.11.008](https://doi.org/10.1016/j.bmc.2004.11.008).

References and notes

- Walsh, J. A. *Ann. N. Y. Acad. Sci.* **1989**, 569, 1.
- Torok, D. S.; Ziffer, H. *J. Med. Chem.* **1995**, 38, 5045.
- Posner, G. H.; O'Dowd, H.; Ploypradith, P.; Cumming, J. N.; Xie, S.; Shapiro, T. A. *J. Med. Chem.* **1998**, 41, 2164.
- Posner, G. H.; Cumming, J. N.; Woo, S. H.; Ploypradith, P.; Xie, S.; Shapiro, T. A. *J. Med. Chem.* **1998**, 41, 940.
- Lin, A. J.; Zikry, A. B.; Kyle, D. E. *J. Med. Chem.* **1997**, 40, 1396.
- Gozalbes, R.; Gálvez, J.; Moreno, A.; García-Domenech, R. *J. Pharm. Pharmacol.* **1999**, 52, 111.
- Go, M. L.; Ngiam, T. L.; Tan, A. L. C.; Kuaha, K.; Wilairat, P. *Eur. J. Pharm. Sci.* **1998**, 6, 19.
- 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. *J. Med. Chem.* **1998**, 41, 1367.
- De Dibyendu; Krogstad, F. M.; Byers, L. D.; Krogstad, D. J. *J. Med. Chem.* **1998**, 41, 4918.
- Estrada, E.; Peña, A.; García-Domenech, R. *J. Comput.-Aided Mol. Des.* **1998**, 12, 583.

11. Estrada, E.; Peña, A. *Bioorg. Med. Chem.* **2000**, *8*, 2755.
12. Estrada, E.; Uriarte, E.; Montero, A.; Teijeira, M.; Santana, L.; De Clercq, E. *J. Med. Chem.* **2000**, *43*, 1975.
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. *Chem. Res. Toxicol.* **2003**, *16*, 1318.
14. Julián-Ortiz, J. V. de; Alapont, C. de. G.; Ríos-Santamarina, I.; García-Doménech, R.; Gálvez, J. *J. Mol. Graphics Mod.* **1998**, *16*, 14.
15. Marrero-Ponce, Y.; Romero, V. TOMOCOMD software; Central University of Las Villas; **2002**. TOMOCOMD (TOPological MOlecular COMputer Design) 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. *Molecules* **2003**, *8*, 687.
17. Marrero-Ponce, Y. *J. Chem. Inf. Comput. Sci.*, doi:10.1021/ci049950k.
18. Marrero-Ponce, Y.; Nodarse, D.; González-Díaz, H.; Ramos de Armas, R.; Romero-Zaldivar, V.; Torrens, F.; Castro, E. A. *CPS : physchem/0401004*.
19. Marrero-Ponce, Y.; Cabrera, M. A.; Romero, V.; Ofori, E.; Montero, L. A. *Int. J. Mol. Sci.* **2003**, *4*, 512.
20. Marrero-Ponce, Y.; Cabrera, M. A.; Romero, V.; González, D. H.; Torrens, F. *J. Pharm. Pharm. Sci.* **2004**, *7*, 186.
21. Marrero-Ponce, Y.; González-Díaz, H.; Romero-Zaldivar, V.; Torrens, F.; Castro, E. A. *Bioorg. Med. Chem.* **2004**, *12*, 5331.
22. Marrero-Ponce, Y. *Bioorg. Med. Chem.* **2004**, *12*, 6351–6369.
23. Marrero-Ponce, Y.; Cabrera, M. A.; Romero-Zaldivar, V.; Bermejo, M.; Siverio, D.; Torrens, F. *Internet Electronic J. Mol. Des.*, in press.
24. Pauling, L. *The Nature of Chemical Bond*; Cornell University Press: New York, 1939, pp 2–60.
25. Klein, D. J. *Internet Electron. J. Mol. Des.* **2003**, *2*, 814.
26. Julian-Ortiz, J. V.; Gálvez, J.; Muñoz-Collado, C.; García-Doménech, R.; Gimeno-Cardona, C. *J. Med. Chem.* **1999**, *42*, 3308.
27. Cabrera, M. A.; Bermejo, M. *Bioorg. Med. Chem.* **2004**, *12*, 5833–5843.
28. Ríos-Santamarina, I.; García-Doménech, R.; Cortijo, J.; Santamaría, P.; Morcillo, E. J.; Gálvez, J. *Internet Electron. J. Mol. Des.* **2002**, *1*, 70.
29. Mishra, R. K.; García-Doménech, R.; Galvez, J. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 387.
30. Cronin, M. T. D.; Aptula, A. O.; Dearden, J. C.; Duffy, J. C.; Netzeva, T. I.; Patel, H.; Rowe, P. H.; Schultz, T. W.; Worth, A. P.; Voutzoulidis, K.; Schüürmann, G. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 869.
31. Ann, D. C. In *Burger's Medicinal Chemistry and Drug Discovery*; Wolff, M. E., Ed.; Wiley-Interscience: New York, 1997; Vol. 5.
32. Negwer, M. *Organic-Chemical Drugs and their Synonyms*; Akademie: Berlin, 1987.
33. Domínguez, J. N.; López, S.; Charris, J.; Iarrosio, L.; Lobo, G.; Semenov, A.; Olson, J. E.; Rosenthal, P. J. *J. Med. Chem.* **1997**, *40*, 2726.
34. Hawley, S. R.; Bray, P. G.; Mungthin, M.; Atkinson, J. D.; O'Neill, P. M.; Ward, S. A. *Antimicrob. Agents Chemother.* **1998**, *42*, 682.
35. Rucker, G.; Schenkel, E. P.; Manns, D.; Mayer, R.; Heiden, K.; Heinzmann, B. M. *Planta Med.* **1996**, *62*, 565.
36. Ring, C. S.; Sun, E.; Mekerrow, J. H.; Lee, G. K.; Rosenthal, P. J.; Kurtz, I. D.; Cohen, F. E. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 3583.
37. Ryley, J. F.; Peters, W. *Ann. Trop. Med. Parasitol.* **1970**, *64*, 209.
38. Basco, L. K.; Dechy-Cabaret, O.; Ndounga, M.; Meche, F. S.; Robert, A.; Meunier, B. *Antimicrob. Agents Chemother.* **2001**, *45*, 1886.
39. 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. *J. Med. Chem.* **1999**, *42*, 1428.
40. Raynes, K. *Int. J. Parasitol.* **1999**, *29*, 367.
41. Tsai, C. S.; Shen, A. Y. *Arch. Pharm.* **1994**, *327*, 677.
42. Posner, G. H.; Tao, X.; Cumming, J. N.; Klinedinst, D.; Shapiro, T. A. *Tetrahedron Lett.* **1996**, *37*, 7225.
43. Philipp, A.; Kepler, J. A.; Johnson, B. H.; Carroll, F. I. *J. Med. Chem.* **1998**, *31*, 870.
44. Figgitt, D.; Denny, W.; Chvalitschewinkoon, P.; Wilairat, P.; Ralph, R. *Antimicrob. Agents Chemother.* **1992**, *36*, 1644.
45. Nga, T. T. T.; Menage, C.; Begue, J. P.; Delpon, D. B.; Gantier, J. C. *J. Med. Chem.* **1998**, *41*, 4101.
46. Avery, M. A.; Bonk, J. D.; Chong, W. K. M.; Mehrotra, S.; Miller, R.; Milhous, W.; Goins, D. K.; Venkatesan, S.; Wyandt, C.; Khan, I.; Avery, B. A. *J. Med. Chem.* **1995**, *38*, 5038.
47. Posner, G. H.; Wang, D.; González, L.; Tao, X.; Cumming, J. N.; Klinedints, D.; Shapiro, T. A. *Tetrahedron Lett.* **1996**, *37*, 815.
48. 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.
49. Posner, G. H.; O'Dowd, H.; Caferro, T.; Cumming, J. N.; Ploypradith, P.; Xie, S.; Shapiro, T. A. *Tetrahedron Lett.* **1998**, *39*, 2273.
50. Posner, G. H.; McGarvey, D. J.; Oh, C. H.; Kumar, N.; Meshnick, S. R.; Asawamahasadka, W. *J. Med. Chem.* **1995**, *38*, 607.
51. Posner, G. H.; González, L.; Cumming, J. N.; Klinedints, D.; Shapiro, T. A. *Tetrahedron* **1997**, *53*, 37.
52. Venugopalan, B.; Bapat, C. P.; Karnik, P. J. *Bioorg. Med. Chem. Lett.* **1994**, *4*, 751.
53. Avery, M. A.; Gao, F.; Chong, W. K. M.; Hendrickson, T. F.; Inman, W. D.; Crews, P. *Tetrahedron* **1994**, *50*, 957.
54. Avery, M. A.; Mehrotra, S.; Johnson, T. L.; Bonk, J. D.; Vroman, J. A.; Miller, R. *J. Med. Chem.* **1996**, *39*, 4149.
55. Venugopalan, B.; Bapat, C. P.; Karnik, P. J.; Chatterjee, D. K.; Iyer, N.; Lepcha, D. *J. Med. Chem.* **1995**, *38*, 1992.
56. Zouhiri, F.; Desmaele, D.; d' Angelo, J.; Riche, C.; Gay, F.; Cicéron, L. *Tetrahedron Lett.* **1998**, *39*, 2969.
57. Posner, G. H.; Parker, M. H.; Northrop, J.; Elias, J. S.; Ploypradith, P.; Xie, S.; Shapiro, T. A. *J. Med. Chem.* **1999**, *42*, 300.
58. Cumming, J. N.; Wang, D.; Park, S. B.; Shapiro, T. A.; Posner, G. H. *J. Med. Chem.* **1998**, *41*, 952.
59. Posner, G. H.; Oh, C. H.; Gerena, L.; Milhous, W. K. *J. Med. Chem.* **1992**, *35*, 2459.
60. Galas, M.; Cordina, G.; Bompert, J.; Bari, M. B.; Jei, T.; Ancelin, M. L.; Vial, H. *J. Med. Chem.* **1997**, *40*, 3557.
61. Ismail, F. M. D.; Dascombe, M. J.; Carr, P.; North, S. E. *J. Pharm. Pharmacol.* **1996**, *48*, 841.
62. Ram, V. J.; Saxena, A. S.; Srivastavab, S.; Chandrab, S. *Bioorg. Med. Chem. Lett.* **2000**, *10*, 2159.
63. Posner, G. H.; Northrop, J.; Paik, I. H.; Borstnik, K.; Dolan, P.; Kensler, T. W.; Xie, S.; Shapiro, T. A. *Bioorg. Med. Chem.* **2000**, *10*, 227.
64. Gironés, X.; Gallegos, A.; Carbó-Dorca, R. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 1053.

65. Cheng, F.; Shen, J.; Luo, X.; Zhu, W.; Gu, J.; Ji, R.; Jiang, H.; Chen, K. *Bioorg. Med. Chem.* **2002**, *10*, 2883.
66. Santos-Filho, O. A.; Hopfinger, A. J. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 1.
67. Jain, R.; Vangapandu, S.; Jain, M.; Kaur, N.; Singhb, S.; Singhb, P. P. *Bioorg. Med. Chem. Lett.* **2002**, *12*, 1701.
68. Itoh, T.; Shirakami, S.; Ishida, N.; Yamashita, Y.; Yoshida, T.; Kimb, H. S.; Watayab, Y. *Bioorg. Med. Chem. Lett.* **2000**, *10*, 1657.
69. Reichenberg, A.; Wiesner, J.; Weidemeyer, C.; Dreiseidler, E.; Sanderbrand, S.; Altincicek, B.; Beck, E.; Schlitzerc, M.; Jomaa, H. *Bioorg. Med. Chem. Lett.* **2001**, *11*, 833.
70. Ryckebusch, A.; Deprez-Poulain, R.; Maes, L.; Debreu-Fontaine, M. A.; Mouray, E.; Grellier, P.; Sergheraert, C. *J. Med. Chem.* **2003**, *46*, 542.
71. Nöteberg, D.; Hamelink, E.; Hulten, J.; Wahlgren, M.; Vrang, L.; Samuelsson, B.; Hallberg, A. *J. Med. Chem.* **2003**, *46*, 734.
72. 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. Bioorg. Med. Chem. Lett.* **2001**, *11*, 773.
73. STADISTICA, version 5.5; Statsoft Inc.; 1999.
74. Baldi, P.; Brunak, S.; Chauvin, Y.; Andersen, C. A.; Nielsen, H. *Bioinformatics* **2000**, *16*, 412.
75. Golbraikh, A.; Tropsha, A. *J. Mol. Graphic Modell.* **2002**, *20*, 269.
76. Estrada, E. In *Topological Indices and Related Descriptors in QSAR and QSPR*; Devillers, J., Balaban, A. T., Eds.; Gordon and Breach: Amsterdam, 1999, pp 403–453.