Available online at www.sciencedirect.com
SCIENCE DIRECT
European Journal of Medicinal Chemistry 41 (2006) 483–493
EUROPEAN JOURNAL OF MEDICINAL CHEMISTRY
http://france.elsevier.com/direct/ejmech

Original paper

Non-stochastic quadratic fingerprints and LDA-based QSAR models in hit and lead generation through virtual screening: theoretical and experimental assessment of a promising method for the discovery of new antimalarial compounds

Alina Montero-Torres [a,*], Rory N. García-Sánchez [b,e], Yovani Marrero-Ponce [a,c,d], Yanetsy Machado-Tugores [a], Juan J. Nogal-Ruiz [e], Antonio R. Martínez-Fernández [e], Vicente J. Arán [f], Carmen Ochoa [f], Alfredo Meneses-Marcel [a], Francisco Torrens [d]

[a] *Department of Drug Design, CBQ, Central University of Las Villas, 54830 Santa Clara, Villa Clara, Cuba*
[b] *Laboratorio de Investigación de Productos Naturales Antiparasitarios de la Amazonía, Universidad Nacional de la Amazonía Peruana, Pasaje Los Paujiles s/n, A A.H H Nuevo San Lorenzo, San Juan Bautista, Iquitos, Peru*
[c] *Department of Pharmacy, Faculty of Chemistry-Pharmacy, Central University of Las Villas, 54830 Santa Clara, Villa Clara, Cuba*
[d] *Institut Universitari de Ciència Molecular, Universitat de València, Edifici d'Instituts de Paterna, P.O. Box 22085, E-46071 Valencia, Spain*
[e] *Departamento de Parasitología, Facultad de Farmacia (UCM), Av. Complutense s/n E-28040 Madrid, Spain*
[f] *Instituto de Química Médica (CSIC), Juan de la Cierva 3, E-28006 Madrid, Spain*

Received in revised form 16 December 2005; accepted 20 December 2005
Available online 20 March 2006

Abstract

In order to explore the ability of non-stochastic quadratic indices to encode chemical information in antimalarials, four quantitative models for the discrimination of compounds having this property were generated and statistically compared. Accuracies of 90.2% and 83.3% for the training and test sets, respectively, were observed for the best of all the models, which included non-stochastic quadratic fingerprints weighted with Pauling electronegativities. With a comparative purpose and as a second validation experiment, an exercise of virtual screening of 65 already-reported antimalarials was carried out. Finally, 17 new compounds were classified as either active/inactive ones and experimentally evaluated for their potential antimalarial properties on the ferriprotoporphyrin (FP) IX biocrystallization inhibition test (FBIT). The theoretical predictions were in agreement with the experimental results. In the assayed test compound **C5** resulted more active than chloroquine. The current result illustrates the usefulness of the *TOMOCOMD-CARDD* strategy in rational antimalarial-drug design, at the time that it introduces a new family of organic compounds as starting point for the development of promising antimalarials.

Keywords: Antimalarial compounds; QSAR; LDA; Non-stochastic quadratic index; *TOMOCOMD-CARDD* software

1. Introduction

Being known since time immemorial, malaria remains a serious and complex health problem today. Approximately 300–400 million people worldwide suffer from this infectious disease, dying ca. 3 million every year, mostly children younger than 5 years [1–3]. This situation has become even more complex over the last decades with the increase in resistance to those drugs normally used to combat the malaria parasite [3–5]. Therefore, the international scientific community is called upon to use efficient strategies for the discovery of non-traditional antimalarial agents [6–9].

Computer-aided drug design has emerged in the pharmaceutical world as an important tool for the rational search of che-

* Corresponding author. Tel.: +53 42 28 1473; fax: +53 42 28 1130, +53 42 28 1455.
E-mail address: alinamontero@gmail.com (A. Montero-Torres).

0223-5234/$ - see front matter © 2006 Elsevier SAS. All rights reserved.
doi:10.1016/j.ejmech.2005.12.010

micals with desired properties. Different studies related to the in silico design have been reported in the literature during the last years [7–16]. In this context, our research group has recently introduced the *TOMOCOMD*-CARDD scheme for the codification of chemical information of small and medium-sized molecules considering different atomic features [17–23]. Using the *TOMOCOMD*-CARDD strategy, it is possible to encode "critical" structural patterns of data sets consisting of compounds with common molecular scaffolds as well as very diverse structures. Specifically, non-stochastic quadratic indices were defined upon the basis of linear algebra theory by one of the present authors, as the first descriptors family implemented in the *TOMOCOMD-CARDD* software [20]. These fingerprints have been successfully used in the generation of *quantitative structure–activity relationships* (QSAR) to predict physical and pharmacological properties of organic compounds [9,20,22,24–27]. In this context, biological properties like antibacterial [25], trichomonacidal [26] and trypanocidal activities [27] have been successfully studied. In an earlier publication, an overview of the significance of these *TOMOCOMD* descriptors and a comparison whit other molecular indices have been presented [22].

With the aim of extending the application field of the *TOMOCOMD*-CARDD method, the present study is intended to develop quantitative models to discriminate antimalarial compounds from inactive ones using, in a comparative way, four families of atomic labels. It is also an objective of the present report to conduct exercises of virtual screening for antimalarials of known activity as well as for compounds for which the activity of interest has been left undetected so far. Both experiments can be considered as part of the external validation process. Results of virtual evaluation and in vitro antimalarial activity of 17 already synthesized homocyclic and heterocyclic compounds will be presented.

2. Materials and methods

2.1. Computational approach

Calculations were carried out on a PC Pentium-4 2.0 GHz. The *CARDD* module implemented in the *TOMOCOMD* software [17] was used to calculate total and local non-stochastic quadratic indices for the data set of organic molecules reported by R. Gozalbes et al. [7]. Weighting schemes depicted in Table 1 were used as atomic labels (molecular vector's components) [18,19].

We calculated k^{th} non-stochastic total quadratic indices $\boldsymbol{q_k}(x)$ and $\boldsymbol{q_k}^{\text{H}}(x)$ (without and with consideration of H atoms, respectively) as quadratic forms as follows:

$$q_k(x) = \sum_{i=1}^{n}\sum_{j=1}^{n} {}^{k}a_{ij}x_ix_j$$

where n is the number of atoms in the molecule; $x_1,\ldots,x_n$ are the coordinates or components of the "molecular vector" (X) (which can be seen as weights for each kind of atom in the molecule), and ${}^{k}a_{ij}$ are the elements of the k^{th} power of the symmetric square matrix **M**(G) of the molecular pseudograph (G). This matrix is constructed following the spirit of an "extended Hückel" molecular orbital model and encloses information of all valence-bond electrons (σ- and π-networks). For powers greater than 1, the electronic interaction through the chemical network is also taken into account [20].

Table 1
Values of the atomic weights used for quadratic index calculation

ID	Atomic mass	VdW[a] volume	Polarizability	Pauling electronegativity
H	1.01	6.709	0.667	2.20
B	10.81	17.875	3.030	2.04
C	12.01	22.449	1.760	2.55
N	14.01	15.599	1.100	3.04
O	16.00	11.494	0.802	3.44
F	19.00	9.203	0.557	3.98
Al	26.98	36.511	6.800	1.61
Si	28.09	31.976	5.380	1.9
P	30.97	26.522	3.630	2.19
S	32.07	24.429	2.900	2.58
Cl	35.45	23.228	2.180	3.16
Fe	55.85	41.052	8.400	1.83
Co	58.93	35.041	7.500	1.88
Ni	58.69	17.157	6.800	1.91
Cu	63.55	11.494	6.100	1.90
Zn	65.39	38.351	7.100	1.65
Br	79.90	31.059	3.050	2.96
Sn	118.71	45.830	7.700	1.96
I	126.90	38.792	5.350	2.66

[a] VdW: van der Waals.

k^{th} local quadratic indices of heteroatoms (S,N,O) [$\boldsymbol{q_{kL}}(x_{\text{E}})$ and $\boldsymbol{q_{kL}}^{\text{H}}(x_{\text{E}})$], hydrogen bonded to heteroatoms (S,N,O) [$\boldsymbol{q_{kL}}(x_{\text{E-H}})$] and halogens [$\boldsymbol{q_{kL}}(x_{\text{H}})$], as fragments [20], were calculated as:

$$q_{kL}(x) = \sum_{i=1}^{n}\sum_{j=1}^{n} {}^{k}a_{ijL}x_ix_j$$

where ${}^{k}\boldsymbol{a}_{ijL}$ is the element of the "i" row and "j" column of $\mathbf{M}_{\mathbf{L}}^{\boldsymbol{k}}$. Matrix $\mathbf{M}^{\boldsymbol{k}}$ encodes information of the selected fragment as well as of the molecular environment.

2.2. Statistical analysis

Linear discriminant analysis (LDA) was performed with the STATISTICA 5.5 package for Windows [28]. Forward stepwise procedure was used for variable selection [28]. Quantitative models with the following form were obtained:

$$\begin{aligned}\mathrm{P} = {} & \mathrm{a_0q_0(x)} + \mathrm{a_1q_1(x)} + \ldots + \mathrm{a_nq_n(x)} + \mathrm{a_{n+1}q_0L(x)} \\ & + \mathrm{a_{n+2}q_1L(x)} + \ldots + \mathrm{a_mq_mL(x)}\end{aligned}$$

where $\boldsymbol{P}$ is the biological property, $\boldsymbol{q_n}(x)$, the n^{th} total quadratic index, $\boldsymbol{q_{mL}}(x)$, the m^{th} local quadratic index and a_n's and a_m's, the coefficients obtained by LDA. The antimalarial activity was coded by a dummy variable "***Class***", which indicates the presence of either an active compound (***Class*** = 1) or an inactive one (***Class*** = –1). The classification of all cases was per-