ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/7300903

A Computer-Based Approach to the Rational Discovery of New Trichomonacidal Drugs by Atom-Type Linear Indices

Article *in* Current Drug Discovery Technologies · January 2006

DOI: 10.2174/157016305775202955 · Source: PubMed

CITATIONS
40

READS
30

13 authors, including:

Yovani Marrero-Ponce
Universidad San Francisco de Quito (USFQ)
214 PUBLICATIONS **3,515** CITATIONS
SEE PROFILE

Jose Antonio Escario
Complutense University of Madrid
112 PUBLICATIONS **1,301** CITATIONS
SEE PROFILE

Alicia Gómez Barrio
Complutense University of Madrid
111 PUBLICATIONS **1,834** CITATIONS
SEE PROFILE

J.J. Nogal-Ruiz
Complutense University of Madrid
65 PUBLICATIONS **1,225** CITATIONS
SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Biomolecular characterization of novel isolates of Trichomonas vaginalis View project

Novel Similarity Measures in Cheminformatics View project

All content following this page was uploaded by Alfredo Meneses-Marcel on 25 May 2017.

The user has requested enhancement of the downloaded file.

A Computer-Based Approach to the Rational Discovery of New Trichomonacidal Drugs by Atom-Type Linear Indices

Yovani Marrero-Ponce,[1,2*] Yanetsy Machado-Tugores,[3] David Montero Pereira,[4] José Antonio Escario,[4] Alicia Gómez Barrio,[4] Juan José Nogal-Ruiz,[4] Carmen Ochoa,[5] Vicente J. Arán,[5] Antonio R. Martínez-Fernández,[4] Rory N. García Sánchez,[4,6] Alina Montero-Torres,[1] Francisco Torrens,[2] and Alfredo Meneses-Marcel.[3,4]

[1]*Department of Pharmacy, Faculty of Chemistry-Pharmacy and Department of Drug Design, Chemical Bioactive Center. Central University of Las Villas, Santa Clara, 54830, Villa Clara, Cuba.*

[2]*Institut Universitari de Ciència Molecular, Universitat de València, Dr. Moliner 50, E-46100 Burjassot (València), Spain.*

[3]*Department of Parasitology, Chemical Bioactive Center. Central University of Las Villas, 54830, Villa Clara, Cuba.*

[4]*Departamento de Parasitología, Facultad de Farmacia, UCM, Pza. Ramón y Cajal s/n, 28040 Madrid.*

[5]*Instituto de Química Médica, CSIC, c/ Juan de la Cierva 3, 28006 Madrid, Spain.*

[6]*Laboratorio de Investigación de Productos Naturales Antiparasitarios de la Amazonía (LIPNAA), Universidad Nacional de la Amazonía Peruana (UNAP), Iquitos, 496 Perú.*

Abstract: Computational approaches are developed to design or rationally select, from structural databases, new lead trichomonacidal compounds. First, a data set of 111 compounds was split (design) into training and predicting series using hierarchical and partitional cluster analyses. Later, two discriminant functions were derived with the use of non-stochastic and stochastic atom-type linear indices. The obtained LDA (linear discrimination analysis)-based QSAR (quantitative structure-activity relationship) models, using non-stochastic and stochastic descriptors were able to classify correctly 95.56% (90.48%) and 91.11% (85.71%) of the compounds in training (test) sets, respectively. The result of predictions on the 10% *full-out* cross-validation test also evidenced the quality (robustness, stability and predictive power) of the obtained models. These models were orthogonalized using the Randi_ orthogonalization procedure. Afterwards, a simulation experiment of virtual screening was conducted to test the possibilities of the classification models developed here in detecting antitrichomonal chemicals of diverse chemical structures. In this sense, the 100.00% and 77.77% of the screened compounds were detected by the LDA-based QSAR models (Eq. 13 and Eq. 14, correspondingly) as trichomonacidal. Finally, new lead trichomonacidals were discovered by prediction of their antirichomonal activity with obtained models. The most of tested chemicals exhibit the predicted antitrichomonal effect in the performed ligand-based virtual screening, yielding an accuracy of the 90.48% (19/21). These results support a role for TOMOCOMD-CARDD descriptors in the *biosilico* discovery of new compounds.

Keywords: *TOMOCOMD-CARDD* Software, Atom-Based Linear Index, LDA-Based QSAR Model, Trichomonacidal Activity, Virtual Screening, Lead Antitrichomonal Compound, Cytocidal activity, Heterocycles

1. INTRODUCTION

Trichomonas vaginalis is a parasitic protozoan that is the cause of trichomoniasis, a sexually transmitted disease of worldwide importance [1-3]. Recent estimates have suggested that *T. vaginalis* infections account for nearly one-third of the 15.4 million cases of sexually transmitted diseases in the United States [4]. In 1995, the World Health Organization estimated the number of adults with trichomoniasis at 170 million worldwide, more than the combined numbers for gonorrhea, syphilis, and chlamydia [5]. Infection with this organism has been linked to various additional pathological manifestations, including cervical neoplasia [6-8], atypical pelvic inflammatory disease [9], and tubal infertility [10], and has been reported to be a risk factor in the development of posthysterectomy cuff cellulites [11]. Infection with *T. vaginalis* has been linked to premature rupture of placental membranes, premature birth, and low birth weight [12,13]. *T. vaginalis* infection has also been reported to increase intrauterine transmission of cytomegalovirus [14] and elevate the risk of acquiring human immunodeficiency virus [15].

Metronidazole has been the drug of choice for *T. vaginalis* infections since 1960 [16] and remains effective today, with a cure rate of approximately 95% [17]. Metronidazole-resistant trichomoniasis had been reported in 1962 [18], two years after metronidazole introduction to *T. vaginalis* therapy. Although metronidazole resistance has been considered rare, treatment of these rare patients who do

*Address correspondence to this author at the Department of Pharmacy, Faculty of Chemistry-Pharmacy and Department of Drug Design, Chemical Bioactive Center. Central University of Las Villas, Santa Clara, 54830, Villa Clara, Cuba; Tel: 53-42-281192, 281473; Fax: 53-42-281130, 281455; E-mail: ymponce@gmail.com

1570-1638/05 $50.00+.00 © 2005 Bentham Science Publishers Ltd.

not respond to treatment is extremely problematic for physicians and is associated with enormous patient suffering [19].

Clearly, new Trichomonacidal agents are needed to treat resistant organisms. Although there are many other nitroimidazoles, only metronidazole is available in North America. Furthermore, all the nitroimidazoles have similar modes of antimicrobial activity [20], and so resistance to metronidazole often includes resistance to the other nitroimidazoles [21].

However, the great cost associated to the development of new compounds and the small economic size of the market for antiprotozoal drugs makes this development slow. For this reason, it is necessary to develop computational methods permitting theoretical *–in silico-* evaluations of trichomonacidal activity for virtual libraries of chemicals before these compounds are synthesized in the laboratory. This 'in silico' world of data, analysis, hypothesis, and models that reside inside a computer is alternative to the 'real' world of synthesis and screening of compounds in the laboratory [22,23].

At present, many large pharmaceutical industries have reoriented their research strategies seeking to solve the problem of generation/selection of novel chemical entities (NCEs), one of the major bottlenecks in the drug discovery process. In fact, currently most integration projects include efforts to integrate the data associated with NCE generation [24]. Alternatively, several approaches to the computer-aided molecular design and high-throughput *in silico* screening (or virtual high-throughput screening) have been introduced in the literature [25]. Nevertheless, novel computational methods and strategies are required to deliver a system that significantly reduces the time-to-market and research and development (R&D) spendings, and increase the rate at which NCEs progress through the pipeline. Such studies if they are implemented successfully can deliver substantial benefits and act as the bedrock for NCE selection [24].

In this context, our research group has recently introduced a novel scheme to perform rational *–in silico-* molecular design (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) [26]. This method has been developed to generate 2D (topologic), 2.5 (3D-chiral) and 3D (topographic and geometric) molecular descriptors based on the application of the discrete mathematics and linear algebra theory to chemistry. In this sense, atomic, atom-type and total linear and quadratic molecular fingerprints have been defined in analogy to the linear and quadratic mathematical maps [27,28]. This *In silico,* method has been successfully applied to the prediction of several physical, physicochemical and chemical properties of organic compounds [27-30]. 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 [31].

The latter opportunity has allowed the description of the significance-interpretation and the comparison to other molecular descriptors [28,29]. The prediction of the pharmacokinetic properties of organic compounds is a problem that can also be addressed using this approach. This method has been used to estimate the intestinal–epithelial transport of drugs in human adenocarcinoma of colon cell line type 2 (Caco-2) cultures of a heterogeneous series of drug-like compounds [32-34]. The obtained results suggested that the *TOMOCOMD-CARDD* method was able to predict 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. It was successfully applied to the virtual (computational) screening of novel anthelmintic compounds, which were then synthesized and *in vivo* evaluated on *Fasciola hepatica* [35,36].

Studies for the fast-track discovery of novel paramphistomicides, antibacterial and antimalarial compounds were also conducted with this theoretical approach [37-40].

Later, promising results have been found in the modeling of the interaction between drugs and HIV _-RNA packaging-region in the field of bioinformatics using the *TOMOCOMD-**CANAR*** (***C***omputed-***A***ided ***N***ucleic ***A***cid ***R***esearch) approach [41,42]. Finally, an alternative formulation of our approach for structural characterization of proteins was carried out recently [43,44]. This extended method [*TOMOCOMD-**CAMPS*** (***C***omputed-***A***ided ***M***odeling 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 [43,44].

The main objective of this work was to use non-stochastic and stochastic atom-type linear indices to generate predictive LDA (linear discriminant analysis)-based QSAR models enabling the selection of new hits and lead drug-like compounds with antitrichomonal activity. The *in vitro* evaluation of a new lead series of heterocyclic compounds with antitrichomonal activity is also presented.

2. MATERIALS AND METHODS

2.1. TOMOCOMD-CARDD Approach and 2D Atom-Based Linear Indices

TOMOCOMD is an interactive program for molecular design and bioinformatic research [26]. It is composed of four subprograms; each one of them allows drawing the structures (drawing mode) and calculating molecular 2D/3D (calculation mode) descriptors. The modules are named 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 the present report, we outline salient features concerned with only one of these subprograms, CARDD and with the calculation of non-stochastic and stochastic 2D

atom-based linear indices. The mathematical basis concerning these novel molecular descriptors has been described in previous reports [28,30,35,38,40], thus we will outline only the fundamental remarks.

Briefly, this method codifies the molecular structure by means of mathematical linear maps. In order to calculate these applications for a molecule, the *n*x*n* k^{th} "non-stochastic and stochastic graph–theoretic electronic-density matrices", $\mathbf{M}^k$ and $\mathbf{S}^k$ are constructed, where *n* is the number of atoms in the molecule [28,30,35,38,40]. The coefficients ${}^k m_{ij}$ are the elements of the k^{th} power of the symmetric square matrix $\mathbf{M}$(G) of the molecular pseudograph G and are defined as follows [28,30,35,38,40]:

$$\begin{aligned} m_{ij} &= P_{ij} \text{ if } i \neq j \text{ and } \exists\, e_k \in \text{E(G)} \\ &= L_{ii} \text{ if } i = j \\ &= 0 \text{ otherwise} \end{aligned} \quad (\mathbf{1})$$

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.

The elements $m_{ij} = P_{ij}$ of such a matrix represent the number of bonds between an atom *i* and the other *j*. The matrix $\mathbf{M}^k$ provides the number of walks of length *k* that links the vertices v_i and v_j. For this reason, each edge in $\mathbf{M}^1$ represents 2 electrons belonging to the covalent bond between atoms (vertices) v_i and v_j; e.g. the inputs of $\mathbf{M}^1$ are equal to 1, 2 or 3 when single, double or triple bonds, appear respectively, between vertices v_i and v_j, respectively. On the other hand, molecules containing aromatic rings with more than one canonical structure are represented like a pseudograph. It happens for substituted aromatic compounds such as pyridine, naphthalene, quinoline, and so on, where the presence of pi (π) electrons is accounted by means of loops in each atom of the aromatic ring. Conversely, aromatic rings having only one canonical structure, such as furan, thiophene and pyrrole are represented like a multigraph.

As can be seen, $\mathbf{M}^k$, are graph–theoretic electronic–structure models, like an EHT MO model". The $\mathbf{M}^1$ matrix considers all valence-bond electrons (σ- and π-networks) in one step and their power (k = 0, 1, 2, 3…) can be considered 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 [45].

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 set free in space at an arbitrary initial time (t_0). At this time, the electrons are distributed around the atomic nuclei. 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* that can move (step-by-step) to other atoms at different discrete time periods t_k (k = 0, 1, 2, 3,…) throughout the chemical-bonding network.

On the other hand, the $\mathbf{S}^k$(G) can be obtained directly from $\mathbf{M}^k$. Here, $\mathbf{S}^k = [{}^k\mathbf{s}_{ij}]$, is a squared table of order *n* (*n* = number of atoms) and the elements ${}^k\mathbf{s}_{ij}$ are defined as follows [38,40].

$${}^k s_{ij} = \frac{{}^k m_{ij}}{{}^k SUM_i} = \frac{{}^k m_{ij}}{{}^k \delta_i} \quad (\mathbf{2})$$

where, ${}^k m_{ij}$ are the elements of the k^{th} power of $\mathbf{M}$ and the SUM of the *i*th row of $\mathbf{M}^k$ is named the *k*-order vertex degree of atom *i*, ${}^k\delta_i$. Note that the matrix $\mathbf{S}^k$ in Eq. 2 has the property that *the sum of the elements in each row* is 1. An *n*x*n* matrix with non-negative entries having this property is called a "**stochastic matrix"** [46]. The k^{th} $\mathbf{s}_{ij}$ elements are the transition probabilities with the electrons moving from atom *i* to *j* in the discrete time periods t_k. Note that k^{th} element $\boldsymbol{s}_{ij}$ takes into consideration the molecular topology in the k^{th} step throughout the chemical-bonding (σ- and π-) network. The ${}^2\mathbf{s}_{ij}$ values can distinguish between hybrid states of atoms in bonds. For instance, the self-return probability of second order (${}^2\mathbf{s}_{ii}$) [they are the probabilities with which electron return to the original atoms at t_2], varies regularly according to the different hybrid states of atom *i* in the molecule, e.g. an electron will have a higher probability of returning to the sp C atom than to the sp_2 (or sp_3) C atom in t_2 [$p(C_{sp})$>$p(C_{sp2})$>$p(C_{sp2arom})$>$p(C_{sp3})$]. This is a logical result if the electronegativity of these hybrid states is taken into account.

The k^{th} non-stochastic [28,30,35,38,40] and stochastic [38,40] linear indices for atom *i* in a molecule, $\boldsymbol{f_k}(x_i)$ and ${}^s\boldsymbol{f_k}(x_i)$, are computed from these k^{th} non-stochastic and stochastic graph–theoretic electronic-density matrices, $\mathbf{M}^k$ and $\mathbf{S}^k$ as shown in Eqs. **3** and **4**, respectively:

$$f_k(x_i) = \sum_{j=1}^{n} {}^k m_{ij} X_j \quad (\mathbf{3})$$

$${}^s f_k(x_i) = \sum_{j=1}^{n} {}^k s_{ij} X_j \quad (\mathbf{4})$$

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 canonical basis set in $\Re^n$. The components of the molecular vector are numeric values, which can be considered as weights (atom-labels) that characterize each atom in the molecule. Different weighing schemes can be used with this purpose, such as: 1) the atomic masses, 2) the van der Waals volumes, 3) the atomic electronegativities in the Pauling scale, (4) the atomic polarizabilities, and so on [47]. In this work, the Pauling electronegativities were selected as atom weights because they take into account the electronic features of each atom in the molecule, and permit adequately differentiating among atoms [48].

The atomic linear indices are defined as a linear transformation $\boldsymbol{f_k}(x_i)$ on a molecular vector space $\mathfrak{R}^n$. The defined equations (**3**) and (**4**) for $\boldsymbol{f_k}(x_i)$ and ${}^s\boldsymbol{f_k}(x_i)$ may be written as the single matrix equation:

$$\boldsymbol{f_k}(x_i) = [\text{X'}]^k = \mathbf{M}^k[\text{X}] \quad (\mathbf{5})$$

$${}^s\boldsymbol{f_k}(x_i) = [{}^s\text{X'}]^k = \mathbf{S}^k[\text{X}] \quad (\mathbf{6})$$

where [X] is a column vector (an *n*x1 matrix) of the coordinates of X in the canonical basis of $\mathfrak{R}^n$.

This approach is rather similar to the **LCAO-MO** (**L**inear **C**ombination of **A**tomic **O**rbitals-**M**olecular **O**rbitals) method. Really, our approach (for $k = 1$) is a quite similar approximation to the EHT, due to the formalism each MO ψ_i is composed of n valence AOs of atoms in a molecule [49].

Total (whole-molecule) linear indices are linear functionals (or linear forms) [46] on $\Re^n$. The mathematical definition of these molecular descriptors (non-stochastic and stochastic) is the following:

$$f_k(x) = \sum_{i=1}^{n} f_k(x_i) \quad (7)$$

$$^{s}f_k(x) = \sum_{i=1}^{n} {}^{s}f_k(x_i) \quad (8)$$

where n is the number of atoms and $\boldsymbol{f_k}(x_i)$ and $^{s}\boldsymbol{f_k}(x_i)$ are the non-stochastic and stochastic atomic linear indices obtained by Eqs. **3** and **4**, respectively. Then these linear forms, $\boldsymbol{f_k}(x)$ and $^{s}\boldsymbol{f_k}(x)$, can be written in matrix form;

$$\boldsymbol{f_k}(x) = [u]^t [X']^k \quad (9)$$

$$^{s}\boldsymbol{f_k}(x) = [u]^t [{}^{s}X']^k \quad (10)$$

for each molecular vector $X \in \Re^n$. $[u]^t$ is an n-dimensional unitary row vector (a 1xn matrix). As can be seen, the k^{th} total linear index is calculated by summing the local (atomic) linear indices of all atoms in the molecule.

In addition to atomic 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 is calculated by summing the k^{th} atom linear indices of all atoms of the same atom type in the molecule [28,30,35,38,40]. Consequently, if a molecule is partitioned into Z molecular fragments, the total linear indices can be partitioned into Z local linear indices $\boldsymbol{f_{kL}}(x)$ [or $^{s}\boldsymbol{f_{kL}}(x)$], L = 1, …, 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:

$$f_k(x) = \sum_{L=1}^{Z} f_{kL}(x) \quad (11)$$

$$^{s}f_k(x) = \sum_{L=1}^{Z} {}^{s}f_{kL}(x) \quad (12)$$

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) non-stochastic and stochastic linear indices provide much useful information. The atom-type descriptors combine three important aspects of structure information: 1) electron accessibility for the atoms of the same type, 2) presence/absence of the atom type, and 3) count of the atoms in the atom type.

2.2. Computational Strategies

The main steps for the application of present method in QSAR/QSPR and drug design can be briefly summarized in the following set of steps: 1) Draw the molecular pseudographs for each molecule of the data set, using the software 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 appropriate weights in order to differentiate the molecular atoms. In this study, we used the Pauling electronegativity [58] as atomic property for each kind of atom, 3) Compute the total and local (atomic and atom-type) non-stochastic and stochastic linear indices. It can be carried out in the software calculation mode, where you can select the atomic properties and the descriptor family 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 linear indices or other molecular descriptors family implemented in this program, 4) Find 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. That is to say, we can find a quantitative relation between an activity **A** and the linear indices having, for instance, the following appearance, $\mathbf{A} = a_0\boldsymbol{f}_0(x) + a_1\boldsymbol{f}_1(x) + a_2\boldsymbol{f}_2(x) + \ldots + a_k\boldsymbol{f}_k(x) + c$, where **A** is the measured activity, $\boldsymbol{f}_k(x)$ are the k^{th} total linear indices, and the $\boldsymbol{a_k}$*'s* are the coefficients obtained by the linear regression analysis, 5) Test the robustness and predictive power of the QSPR/QSAR equation by using internal [cross-validation] and external (using a test set and an external predicting set) validation techniques, and 6) Apply the obtained LDA-based QSAR models as cheminformatic tool for identifying leads through ligand-based virtual screening-drug discovery process.

The following atom-type descriptors were calculated in this work for describing the antitrichomonal activity of some compounds via LDA models:

i) $\boldsymbol{f}_{kL}(x_E)$ and $\boldsymbol{f}_{kL}{}^{H}(x_E)$ are the k^{th} local (atom-type = heteroatoms: S, N, O) linear indices not considering and considering H-atoms in the molecule, correspondingly. These local descriptors are putative molecular charge, dipole moment, and H-bonding acceptors.

ii) $\boldsymbol{f}_{kL}{}^{H}(x_{E\text{-}H})$ are the k^{th} local (atom-type = H-atoms bonding to heteroatoms: S, N, O) linear indices considering H-atoms in the molecule. These local descriptors are putative H-bonding donors (hydrogen bonding capacity), lipophilicity, and so on.

iii) The k^{th} stochastic atom-type [$^{s}\boldsymbol{f}_k(x_E)$, $^{s}\boldsymbol{f}_k{}^{H}(x_E)$ and $^{s}\boldsymbol{f}_k{}^{H}(x_{E\text{-}H})$] linear indices were also computed.

2.3. Database Selection

A data set of 111 organic-chemicals having a great structural variability was collected from the literature for the present study [50,51]. The data set of active compounds (49 chemicals used as trichomonacidal in clinic) was chosen considering a representation of most of the different structural patterns and action modes for the case of

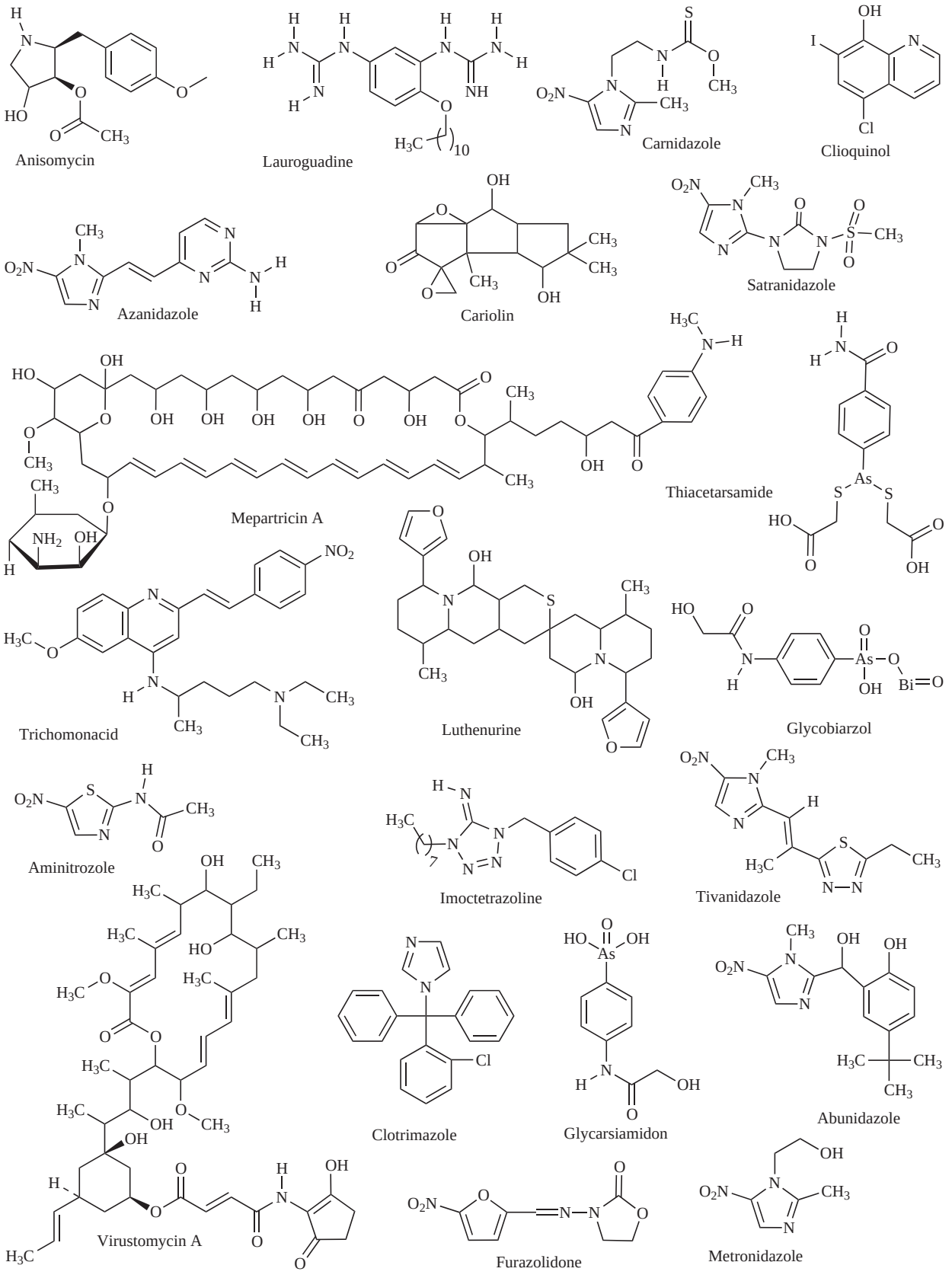


Fig. (1). Random sample of the molecular families of trichomonacidal agents studied here.

compounds with antitrichomonal activity. Figure **1** shows a representative sample of such active compounds.

In addition, the set of inactive compounds was obtained by selecting at random 62 drugs with different

pharmacological uses. These drugs include, for instance, antibiotic, antivirals, sedative/hypnotics, diuretics, anticonvulsants, hemostatics, oral hypoglycemics, antihypertensives, anthelminthics, anticancer, antifungal, etc; guaranteeing also a great structural variability. However, the declaration of these compounds as "inactive" antitrichomonal *per se* does not guarantee that there do not exist trichomonacidal side-effects for some of those 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.

Later, two kinds of cluster analyses (CA) were performed for active and inactive series of compounds, in order to split (design) the dataset (111 organic-chemicals) into training and predicting series in a "rational" way [52,53].

2.4. Data Analysis and Processing

2.4.1. Clustering

Cluster analysis encompasses a number of different classification algorithms and it permits to organize the observed data into meaningful structures. Conceptually, the approach used by CA in order to address this problem can be described well by the saying "birds of a feather flock together" [54]. Many CA algorithms have been invented and they belong to two categories: hierarchical clustering and partitional (non-hierarchical) clustering. Hierarchical clustering rearranges objects in a tree-structure (joining clustering) and these methods are implemented in agglomerative (bottom-up) or divisive (top-down) procedure. On the other hand, the partitional clustering assumes that the objects have non-hierarchical characters [52-54].

Most popular partitional cluster algorithms are *k*-mean cluster algorithms (*k*-MCA) and Jarvis-Patrick (also known as k-nearest neighbor cluster algorithm; *k*-NNCA) algorithms. K-mean clustering algorithms use an interchange (or switching) method to divide *n* data points into *k* groups (clusters) so that the sum of distances/dissimilarities among the objects within the same cluster is minimized. The k-mean approach requires that *k* (the number of clusters) is known before clustering. The Jarvis-Patrick method requires the user specifies the number of nearest neighbors, and the number of neighbors in common to merge to objects. Jarvis-Patrick is a deterministic algorithm; it does not require iterations for computations [52-54].

In order to design training and test series and to demonstrate the structural diversity of the present database, we carried out the two kinds of cluster analyses (*k*-MCA and *k*-NNCA) for both active and inactive series of compounds [52-54]. The statistical software package STATISTICA [55] was used to develop these CAs.

In this study, we used "average linkage" metric as method to merge objects into clusters. The average linkage distance between two clusters is defined as the average distance (squared Euclidean) between pairs of objects, one in each cluster. Average linkage tends to join clusters with small variances and, is biased toward producing clusters with roughly the same variance.

Takeing into consideration that the number of combinations of partitioning N objects into K groups is an astronomical high figure, we forced the STATISTICA program to abort after of 10 iterations in order to produce result in a feasible period of time. The number of members in each cluster and the standard deviation of the variables in the cluster (kept as low as possible) were taken into account, to have an acceptable statistical quality of data partition in clusters. The values of the standard deviation (SS) between and within clusters, of the respective Fisher ratio and their *p*-level of significance were also examined [52-54]. Finally, before carrying out the cluster processes, 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.

2.4.2. Linear Discriminant Analysis

The discriminant functions were obtained by using the Linear Discriminant Analysis (LDA) [56] as implemented in the STATISTICA [55]. The default parameters of this program were used in the development of the model. Forward stepwise was fixed as the strategy for variable selection. The principle of parsimony (Occam's razor) was taken into account as they strategy for model selection. In its original form, the Occam's razor states that »*Numquam ponenda est pluritas sin necesitate*«, which can be translated as »Entities should not be multiplied beyond necessity« [57]. In this case, simplicity is loosely equated with the number of parameters in the model. If we understand the predictive error to be the error rate for unseen examples, the Occam's razor can be stated for the selection of QSAR/QSPR models as ("*QSAR/QSPR Occam's Razor*"): Given two QSAR/QSPR models with the same predictive error, the most simple one should be preferred because simplicity is desirable in itself [57]. In this connection, we select the model with higher statistical signification but having as few parameters (a_k) as possible.

The quality of the models were determined by examining Wilks' λ parameter (*U*-statistic), squared 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 [56]. Models with a proportion between the number of cases and variables in the equation lower than 5 were rejected.

The Wilks' λ for the overall discrimination can take values in the range of 0 (perfect discrimination) to 1 (no discrimination). The D^2 statistics indicates the separation of the respective groups, showing whether the model possesses an appropriate discriminatory power for differentiating between the two respective groups.

By using the models, one compound can then be classified as either active, if $\Delta P\% > 0$, being $\Delta P\% = [P(Active) - P(Inactive)]x100$ or inactive otherwise. P(Active) and P(Inactive) are the probabilities with which the equations classify a compound as active and inactive, respectively.

The statistical robustness and predictive power of the obtained model were assessed using a prediction (test) set [58]. Also a leave-group-out (LGO) cross-validation strategy was carried out. In this case, 10% of the data set was used as

group size, i.e. groups including 10% of the training data set were 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, stability and predictive powers of the obtained models [58].

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 (***C***) in the training and test sets permitted the assessment of the model [59].

The interrelation among the molecular descriptors makes difficult the interpretation of the QSAR model. To overcome this difficulty, an approach based on the orthogonalization of the descriptors has been introduced in the literature [60-66]. The main philosophy of this approach is to avoid the exclusion of descriptors on the basis of its collinearity with other variables included in the model. In addition, it is well known that the interrelatedness among the different descriptors can result in highly unstable regression coefficients, which makes it impossible to know the relative importance of an index and underestimates the utility of the regression coefficient in a model [60-66]. 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 is stable to the inclusion of novel descriptors, which permit to interpret the regression coefficients and to evaluate the role of individual fingerprints in the QSAR model.

In this study, the Randi_ method of orthogonalization was used [60-66]. This method has been described in details in several publications. Thus, we will give only a general overview here. As a first step, an appropriate order of orthogonalization was considered following the order with which the variables were selected in the forward stepwise search procedure of the statistical analysis. The first variable (V_1) is taken as the first orthogonal descriptor ${}^1O(V_1)$, and the second one (V_2) is orthogonalized with respect to it [${}^2O(V_2)$]. The residual of its correlation with ${}^1O(V_1)$, 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.

2.5. Determination of *In Vitro* Trichomonacidal Activity

The biological activity was assayed on *Trichomonas vaginalis* JH31A #4 No. ref. 30326 (ATCC, Maryland, USA) in modified diamond medium supplemented with equine serum and grown at 37 ˚C (5% CO_2). The compounds were added to the cultures at several concentrations (100, 10, and 1 _g/ml) after 6 h of the seeding (0 h). Viable protozoa were assessed at 24 and 48 h after incubation at 37 ˚C by using the Neubauer chamber. Metronidazole (Sigma-Aldrich SA, Spain) was used as reference drug at concentrations of 2, 1, 0,5 _g/ml. Cytocidal and cytostatic activities were determined by calculation of percentages of cytocidal (%C) and cytostatic activities (%CA) in relation with controls as previously reported [67-71].

3. RESULTS AND DISCUSSION

3.1. Construction of Training and Test Sets Using Hierarchical and Non- Hierarchical Cluster Analyses

It is well known that the quality of a classification model is highly dependent on the quality of the selected data set. The most critical aspect for constructing the training set is to warrant enough molecular diversity on it. Taking this into account, we selected a data set of 111 organic chemicals having a great structural variability. In order to demonstrate the structural diversity of this data set, we performed a hierarchical cluster analysis of the active and inactive series [52-53]. The hierarchical clustering approach finds a hierarchy of objects represented by a number of descriptors. The two dendrograms given in Figure **2** and **3**, using the Euclidean distance (X-axis) and the complete linkage (Y-axis), illustrate the results of the k-NNCA developed in active and inactive sets, respectively. As can be seen in both dendrograms, there is a great number of different subsets, which prove the molecular variability of the selected chemicals in these databases.

Also this procedure permits to select compounds for the training and test sets in a representative way, in all levels of the linking distance. In addition, and in order to split also the whole group into two data sets (training and predicting ones), two *k*-MCA [52-53] were performed for active and inactive compounds, respectively.

The main idea of this procedure consists in making a partition of either active or inactive series of chemicals in several statistically representative classes of compounds. This procedure ensures that any chemical class (as determined by the clusters) will be represented in both compounds' series. This "rational" design of training and predicting series allowed us to design both sets that are representative of the whole "experimental universe".

First, we carried out a *k*-MCA algorithm with active compounds and afterwards with inactive ones. The first *k*-MCA (I) divided trichomonacidals into 10 clusters. On the other hand, the inactive compounds were partitioned into 12 clusters (*k*-MCA II). Tables **1** and **2** depict the members of each cluster as active (*k*-MCA I) or inactive (*k*-MCA II) groups, respectively.

Afterwards the selection of the training and prediction sets was performed by taking, in a random way, compounds belonging to each cluster. From these 111 chemicals, 90 were chosen at random to form the training set, being 39 of them actives and 51 inactive ones. The great structural variability of the selected training data set makes possible the discovery of lead compounds, not only with determined mechanisms of antitrichomonal activity, but also with novel

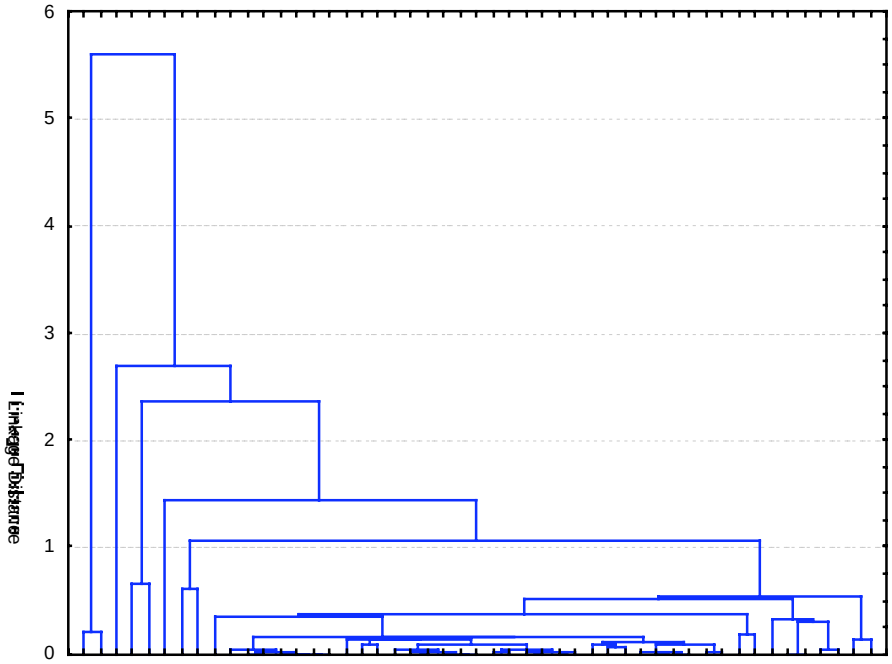


Fig. (2). A dendrogram illustrating the results of the hierarchical *k*-NNCA of the set of active compounds used in the training and prediction set of the present work.

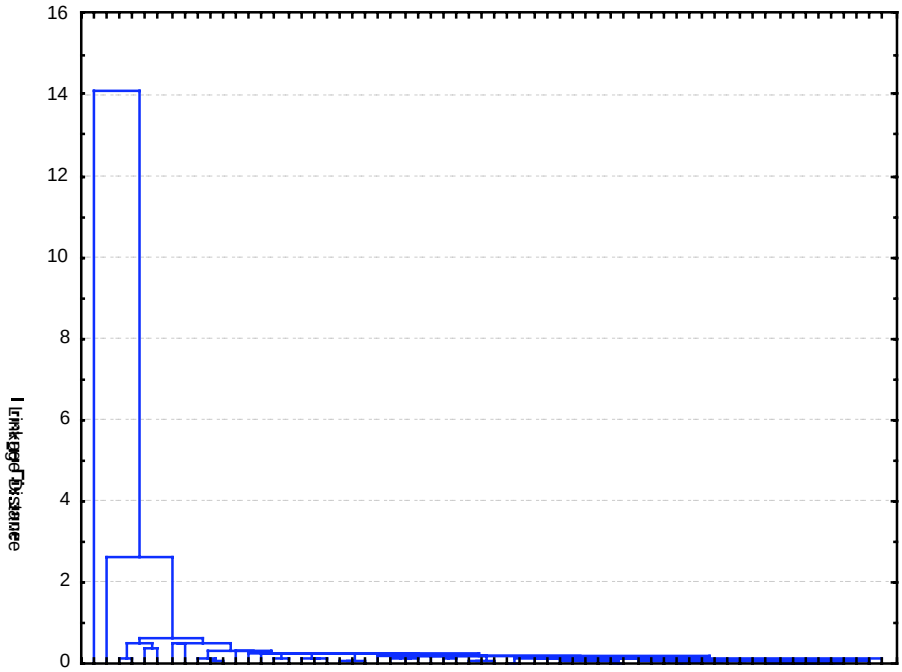


Fig. (3). A dendrogram illustrating the results of the hierarchical *k*-NNCA of the set of inactive compounds used in the training and prediction set of the present work.

modes of action. This will be well-illustrated in this paper in a virtual experiment for lead compounds generation.

The remaining subseries composed of 10 trichomonacidals and 11 compounds with different biological properties were prepared as test sets for the external cross-validation of the models (21 chemicals). These compounds were never used in the development of the classification models. Figure **4** graphically illustrates the above-described procedure where two independent cluster analyses (one for active and the other for inactive compounds) were performed to select a representative sample for the training and test sets.

Table 1. Result of the *k*-MCA I.

Active compounds	Cluster	Dist.	Active compounds	Cluster	Dist.	Active compounds	Cluster	Dist.
Lauroguadine	1	0.20	Nimorazole	5	0.07	Mepartricin A	7	0.11
Azomycin	1	0.20	Ornidazole	5	0.09	Mepartircin B	7	0.11
Acertarsone	2	0.30	Benzoylmetroni-dazole	5	0.05	Metronidazole	8	0.02
Glycobiarzole	2	0.32	Misonidazole	5	0.09	Nifuroxime	8	0.11
Glycarsiamidon	2	0.18	Fexinidazole	5	0.03	Secnidazole	8	0.03
Thiacetarsamide	2	0.24	Pirinidazole	5	0.08	Chlomizol	8	0.05
Virustomycin A	3	0.41	Nimorazole	5	0.07	Isometronidazole	8	0.02
Pentamycin	3	0.41	Carnidazole	6	0.12	Ternidazole	8	0.02
Aminitrozole	4	0.08	Propenidazole	6	0.11	Gynotabs	8	0.05
2 -Amino -5 -nitrotiazole	4	0.21	Furazolidone	6	0.11	Moxnidazole hidrocloride	9	0.20
Trichomonacid	4	0.15	Nifuratel	6	0.14	Satranidazole	9	0.20
Luthenurine	4	0.19	Mertronidazole phosphate	6	0.25	Anisomycin	10	0.23
Abunidazole	4	0.07	Bamnidazole	6	0.15	Cariolin	10	0.18
Imoctetrazoline	4	0.13	Piperanitrozole	6	0.11	Clioquinol	10	0.12
Forminitrazole	4	0.10	Metronidazole hydrogen succinate	6	0.13	Clotrimazol	10	0.27
Acinitrazole	4	0.08	Tivanidazole	6	0.14	Diyodohidroxi-quinoline	10	0.13
Tolamizol	4	0.09	Azanidazole	6	0.35			

Table 2. Result of the *k*-MCA II.

Inactive compounds	Cluster	Dist.	Inactive compounds	Cluster	Dist.	Inactive compounds	Cluster	Dist.
Norantoin	1	0.16	Petidina	5	0.05	Basedol	9	0.13
Rolipram	1	0.09	Tenalidine	5	0.09	Didym levulinate	9	0.17
N-hidroxymethyl-N-methylurea	1	0.03	Bamipine	5	0.19	Cyclopramine	9	0.11
Mecysteine	1	0.10	Nonaferone	5	0.14	Colestipol	9	0.14
Cirazoline	1	0.13	Acetylcholine	5	0.05	4-Chlorobenzoic acid	9	0.11
Zoxazolamine	1	0.09	Amitraz	5	0.11	Acetanilide	9	0.15
Thiacetazone	2	0.23	Diponium Bromide	5	0.05	Proclonol	9	0.15
Orotonsan Fe	2	0.20	Methenamine	6	0.11	Dopamine	10	0.21
Naftazone	2	0.16	Carbimazole	6	0.07	Bufeniode	10	0.15
Ag 307	2	0.12	Ethydine	6	0.12	Carazolol	10	0.21
Eticoumarolum	2	0.16	Chloral betaine	7	0.13	Amantadine	11	0.25
Guanazole	2	0.17	Frigen 113	7	0.11	Propamin “soviet”	11	0.09
Lysergide	2	0.21	Perchloroethane	7	0.03	Vinyl ether	11	0.12
Alibendol	2	0.19	Bisoxatin acetate	8	0.22	Trimethylsulfonium hydroxide	11	0.13
Phenoltetrachloro-phthalein	3	0.15	Besunide	8	0.35	Tetramin	11	0.13
Methocarbamol	3	0.15	Celiprolol	8	0.20	KC-8973	11	0.15
Barbismetylii iodidum	4	0.24	Erysimin	8	0.26	Picosulfate	12	0.80
Pancuronium bromide	4	0.20	Peruvoside	8	0.32	Acetazolamide	12	0.58
Magnesii metioglicas	4	0.27	Asame	8	0.42	Glicondamide	12	0.48
Pyrantel tartrate	5	0.03	Carbavin	9	0.11	Streptomycin	12	1.45
Fentanilo	5	0.04	RMI 11894	9	0.10			

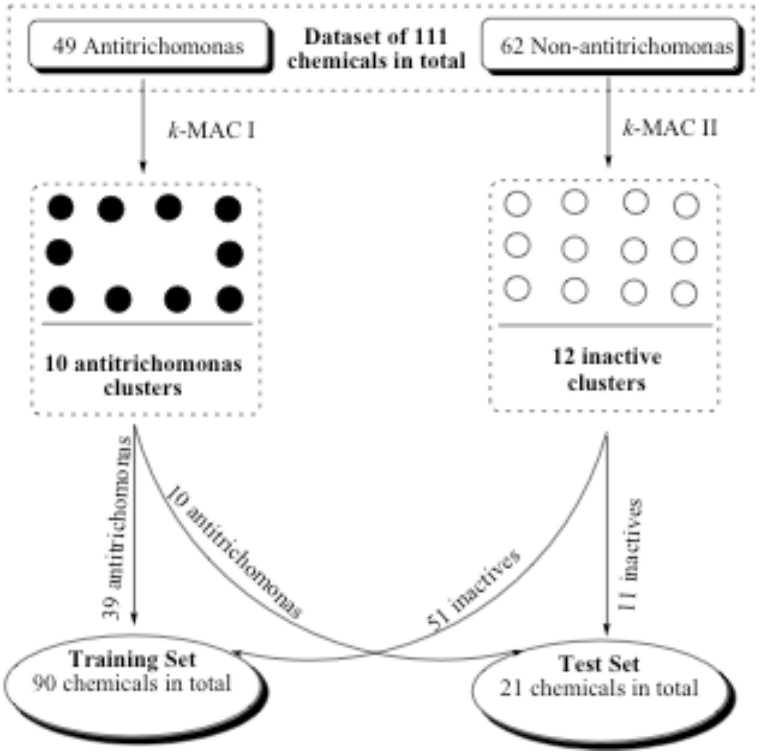


Fig. (4). General algorithm used to design training and test sets throughout *k*-MCA.

Table 3. Main results of the *k*-MCAs, for antitrichomonal and non-antitrichomonal drug-like compounds.

Analysis of Variance				
Variables	**Between SS[a]**	**Within SS[b]**	**Fisher ratio (F)**	***p*-level[c]**
Antitrichomonal agents clusters (k-MCA I)				
$f_{0L}(x_E)$	39.74	1.18	145.46	0.00
$f_{1L}(x_E)$	18.67	1.64	49.27	0.00
$f_{2L}{}^{H}(x_E)$	21.72	1.44	65.20	0.00
$f_{8L}(x_{E\text{-}H})$	37.88	1.83	89.42	0.00
Non-Antitrichomonal agents clusters (k-MCA II)				
$f_{0L}(x_E)$	30.51	6.36	21.79	0.00
$f_{1L}(x_E)$	30.22	1.90	71.96	0.00
$f_{2L}{}^{H}(x_E)$	37.77	2.26	75.79	0.00
$f_{8L}(x_{E\text{-}H})$	23.30	9.38	11.28	0.00

[a]Variability between groups. [b]Variability within groups. [c]Level of significance.

The k^{th} non-stochastic atom-type linear indices were used, with all variables showing *p*-levels <0.05 for the Fisher test. The results are depicted in Table **3**.

From the CAs, it can be concluded that the structural diversity of several up-to-date known antitrichomonal compounds (as codified by *TOMOCOMD-CARDD* descriptors) may be described at least by 10 statistically homogeneous clusters of chemicals.

3.2. Development and Validation of the Discriminant Functions

Although the number of existing statistical methods to get classification functions is relatively extensive, we select linear discriminant analysis (LDA) given the simplicity of the method [56]. The use of LDA in rational drug design has been extensively reported by different authors [23,29-31,33-41]. The best discrimination functions obtained using non-

stochastic and stochastic linear indices for the training set are given below:

$$\textit{Class} = 1.7422 + 28.0505 f_{1L}(x_E) - 31.1775 f_{2L}^{H}(x_E) + 6.5594 f_{0L}(x_E) + 2.1860 f_{8L}^{H}(x_{E\text{-}H}) \quad (\mathbf{13})$$

$N = 90 \quad \lambda = 0.308 \quad D^2 = 8.91 \quad F(4, 85) = 47.558 \quad p<0.0001$

$$\textit{Class} = -7.783 + 5.108\,{}^{s}f_{1L}^{H}(x_E) - 4.406\,{}^{s}f_{14L}(x_E) + 2.9609\,{}^{s}f_{1L}^{H}(x_{E\text{-}H}) - 3.60839\,{}^{s}f_{4L}^{H}(x_{E\text{-}H}) \quad (\mathbf{14})$$

$N = 90 \quad \lambda = 0.32 \quad D^2 = 8.43 \quad F(4, 85) = 44{,}999 \quad p<0.0001$

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 significance level.

Model (**13**) classifies correctly 92.31% of active and 98.04% of inactive compounds in the training set for a global good classification (accuracy) of 95.56%. Model (**14**) correctly classifies 91.11% of the compounds in training set. Specifically, the model correctly classifies 33 out of 39 (84.62%) trichomonacidal compounds and 49 out of 51 (96.08%) inactive chemicals in the training series. On the other hand, Eqs. **13** and **14** show a 90.48% (19/21) and 85.71% (18/3) global predictability in the prediction series, respectively. In Table 4 we give the names of all compounds in the training and test active set together with their posterior probabilities calculated from the Mahalanobis distance using both equations. The same information of all compounds in the training and test inactive set appears in Table **5**. Table **6** summarizes the results of the classifications for both models in the training and test sets. These results validate the models for use in the ligand-based virtual screening taking into consideration that 85.0% is considered as an acceptable threshold limit for this kind of analysis [72].

Table 4. Names and classification of active compounds in training and test series according to the two TOMOCOMD-CARDD models developed in this work.

name	ΔP%[a]	Score[b]	ΔP%[c]	Score[d]	name	ΔP%[a]	Score[b]	ΔP%[c]	Score[d]
Active training set									
Anisomycin	-87.34	1.07	-96.01	1.01	Abunidazole	99.91	-2.41	99.59	-2.46
Virustomycin A	24.23	0.00	95.00	-1.59	Imoctetrazoline	99.67	-1.98	87.51	-1.27
Azanidazole	99.97	-2.81	99.88	-2.89	Forminitrazole	99.00	-1.60	92.60	-1.46
Carnidazole	99.55	-1.88	95.88	-1.66	Chlomizole	99.67	-1.97	90.28	-1.36
Propenidazole	98.84	-1.55	99.01	-2.16	Acinitrazole	98.79	-1.54	92.97	-1.47
Lauroguadine	-93.35	1.30	-81.49	0.45	Moxnidazole	99.99	-3.15	99.83	-2.78
Mepartricin A	99.93	-2.50	96.22	-1.69	Isometronidazole	99.39	-1.77	97.97	-1.91
Metronidazole	99.39	-1.77	97.51	-1.84	Mertronidazole phosphate	100.00	-4.88	99.75	-2.64
Nifuratel	99.97	-2.84	99.79	-2.69	Benzoylmetronidazole	98.64	-1.50	99.27	-2.26
Nifuroxime	100.00	-3.49	99.80	-2.71	Bamnidazole	93.58	-0.97	24.62	-0.51
Nimorazole	99.90	-2.39	97.40	-1.82	Glycarsiamidon	68.52	-0.39	55.98	-0.77
Secnidazole	99.38	-1.76	98.11	-1.93	Fexinidazole	99.87	-2.29	99.61	-2.48
Cariolin	-66.78	0.71	-82.65	0.48	Piperanitrozole	99.35	-1.75	98.25	-1.96
2-Amino-5-nitro-tiazole	99.34	-1.74	92.70	-1.46	Gynotabs	99.52	-1.85	99.39	-2.33
Glycobiarzole	99.99	-3.36	84.19	-1.18	Pirinidazole	99.97	-2.76	99.83	-2.76
Clioquinol	19.75	0.03	-61.41	0.16	Metronidazole hydrogen succinate	97.23	-1.26	91.69	-1.41
Diyodohidroxiquinoline	15.10	0.07	-52.51	0.07	Tolamizole	99.16	-1.66	98.69	-2.06
Ornidazole	99.99	-3.22	99.80	-2.71	Thiacetarsamide	15.35	0.07	-40.28	-0.04
Trichomonacid	100.00	-3.68	99.99	-3.78	Tivanidazole	99.94	-2.57	99.98	-3.52
Luthenurine	8.57	0.11	55.77	-0.77					
Active test set									
Acertarsone	80.75	-0.58	56.94	-0.78	Pentamycin	98.52	-1.47	95.73	-1.65
Furazolidone	99.91	-2.43	99.51	-2.40	Azomycin	99.93	-2.48	96.83	-1.76
Mepartircin B	99.93	-2.51	95.43	-1.63	Ternidazole	99.37	-1.76	98.01	-1.92
Aminitrozole	98.79	-1.54	92.97	-1.47	Misonidazole	99.69	-2.00	96.76	-1.75
Clotrimazole	-94.25	1.35	-92.61	0.79	Satranidazole	93.01	-0.94	97.32	-1.81

[a,c]Antitrichomonal activity predicted by Eq **13** and **14**, respectively: ΔP% = [P(Active) - P(Inactive)]x100. [b,d]Canonical scores obtained from canonical analysis, Eq. **15** and **16**, correspondingly.

Table 5. Names and classification of inactive compounds in training and test series according to the two TOMOCOMD-CARDD models developed in this work.

name	ΔP%[a]	Score[b]	ΔP%[c]	Score[d]	name	ΔP%[a]	Score[b]	ΔP%[c]	Score[d]
Inactive training set									
Amantadine	-97.27	1.60	-99.65	1.85	Nonaferone	-84.66	1.00	-95.58	0.97
Thiacetazone	-43.43	0.48	-37.98	-0.06	Rolipram	-91.31	1.20	-94.58	0.90
Chloral betaine	-83.12	0.97	-99.90	2.28	N-hidroxymethyl -N-methylurea	-96.86	1.56	-99.73	1.94
Carbavin	-99.77	2.43	-99.74	1.95	4-Chlorobenzoic acid	-96.66	1.53	-95.49	0.97
Norantoin	-99.54	2.20	-99.54	1.76	Acetanilide	-97.76	1.67	-99.44	1.69
Orotonsan Fe	-99.66	2.30	-99.83	2.11	Guanazole	-87.17	1.07	-99.94	2.47
Picosulfate	-81.32	0.93	-1.45	-0.32	Tetramin	-92.44	1.25	-98.53	1.36
Naftazone	-91.67	1.22	-91.84	0.75	Mecysteine	-96.60	1.53	-100.00	3.91
Besunide	-75.73	0.83	36.06	-0.59	Cirazoline	-94.13	1.34	-98.83	1.43
Acetazolamide	-59.08	0.62	-89.84	0.68	Methocarbamol	-97.25	1.60	-99.91	2.33
Propamin “soviet”	-97.43	1.62	-99.85	2.15	Lysergide	-74.51	0.81	-94.29	0.88
RMI 11894	-97.74	1.67	-99.76	1.98	Dopamine	-87.09	1.06	-96.24	1.03
Ag 307	-98.18	1.74	-89.99	0.68	Bufeniode	-4.42	0.20	-15.90	-0.22
Barbismetylii iodidum	-99.20	2.02	-99.76	1.99	Celiprolol	-88.84	1.12	-45.77	0.01
Pancuronium bromide	-96.72	1.54	-96.60	1.06	Erysimin	-50.18	0.54	31.11	-0.55
Vinyl ether	-96.07	1.48	-99.86	2.18	Peruvoside	-29.66	0.37	-9.42	-0.27
Basedol	-98.06	1.72	-99.17	1.55	Amitraz	-98.57	1.82	-94.92	0.92
Carbimazole	-99.28	2.05	-99.51	1.74	Proclonol	-59.31	0.63	-86.92	0.58
Didym levulinate	-99.15	2.00	-99.91	2.32	Asame	-93.04	1.28	-99.77	2.00
Perclhoroethane	-14.85	0.27	-99.58	1.79	KC-8973	-93.73	1.32	-98.80	1.43
Pyrantel tartrate	-94.51	1.36	-96.81	1.09	Ethydine	-98.09	1.72	-99.52	1.74
Fentanil	-97.43	1.62	-96.47	1.05	Magnesii metioglicas	-95.55	1.44	-98.65	1.39
Petidine	-96.37	1.51	-96.73	1.08	Alibendol	-77.09	0.85	-79.27	0.41
Tenalidine tartrate	-78.66	0.88	-92.60	0.79	Diponium Bromide	-97.36	1.61	-99.63	1.83
Bamipine	-91.51	1.21	-98.23	1.29	Streptomycin	75.15	-0.49	-59.47	0.14
Colestipol	-97.51	1.63	-99.54	1.76					
Inactive test set									
Methenamine	-90.02	1.16	-99.98	2.85	Cyclopramine	-85.17	1.01	-98.68	1.39
Phenoltetrachloro-phthalein	56.34	-0.26	91.20	-1.39	Trimetilsulfonium hidroxide	-92.60	1.26	-99.56	1.78
Bisoxatin acetate	-97.59	1.65	-24.29	-0.16	Zoxazolamine	-97.54	1.64	-96.62	1.07
Glicondamide	-84.81	1.01	1.38	-0.34	Acetylcholine	-96.29	1.50	-99.95	2.53
Frigen 113	-62.80	0.66	-100.00	3.90	Carazolol	-19.93	0.30	-86.94	0.58
Eticoumarolum	-93.24	1.29	-63.93	0.19					

[a,c]Antitrichomonal activity predicted by Eq **13** and **14**, respectively: ΔP% = [P(Active) - P(Inactive)]x100. [b,d]Canonical scores obtained from canonical analysis, Eq. **15** and **16**, correspondingly.

Table 6. Prediction performances for two LDA-based QSAR models (using non-stochastic and stochastic atom-type linear indices) in the training and test sets.

	Matthews Corr. Coefficient (*C*)	Accuracy ‘Q_{Total}’ (%)	Sensitivity ‘hit rate’ (%)	Specificity (%)	False positive rate ‘false alarm rate’ (%)
Non-Stochastic Atom-type Linear Indices [Eq. (13)]					
Training set	0.91	95.56	92.31	97.3	2.0
Test set	0.81	90.48	90.00	81.82	10.0
Non-Stochastic Atom-type Linear Indices [Eq. (14)]					
Training set	0.82	91.11	84.62	94.29	3.92
Test set	0.72	85.71	90.00	81.82	18.18

Table 7. Results of the 10-*fold* full cross-validation procedure.

Groups	% Class[a]	λ	D^2	F	% Class[b]	% Class[a]	λ	D^2	F	% Class[b]
Eq. **13** (Non-Stochastic Atom-based Linear indices)						Eq. **14** (Stochastic Atom-based Linear indices)				
1	93.67	0.28	10.05	46.80	81.81	91.36	0.30	8.93	42.72	88.88
2	94.94	0.28	9.83	45.76	90.90	92.59	0.28	9.84	47.04	77.77
3	96.20	0.31	8.52	39.66	90.90	90.12	0.34	7.42	35.49	100.00
4	94.94	0.28	10.20	47.01	81.81	92.59	0.30	9.12	43.60	77.77
5	92.40	0.34	7.69	35.80	100.00	91.35	0.31	8.56	40.92	88.88
6	93.67	0.31	8.46	39.38	100.00	90.12	0.33	7.73	36.96	100.00
7	93.67	0.30	8.85	41.20	81.81	90.12	0.33	7.89	37.75	100.00
8	93.63	0.31	8.41	39.42	90.90	90.12	0.31	8.78	42.00	88.88
9	92.40	0.34	7.69	35.80	100.00	88.88	0.31	8.68	41.52	88.88
10	94.94	0.28	9.83	45.76	90.90	91.46	0.32	8.27	39.18	88.88
Mean	94.05	0.30	8.95	41.66	90.90	90.87	0.31	8.52	40.72	89.99
SD	1.20	0.02	0.96	4.37	7.43	1.20	0.02	0.72	3.45	8.20

[a,b] Global good classification from both models in training (90% of the data) and test (10% of the data) sets, respectively.

A more serious analysis was carried out by calculating most of the parameters commonly used in medical statistics (accuracy, sensitivity, specificity and false positive rate) and the Matthews correlation coefficient (***C***). Table 6 also lists these parameters for both obtained models [59]. 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, ***C*** 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 [59]. The obtained models, Eqs. **13** and **14**, showed a high ***C*** of 0.91 (0.81) and 0.82 (0.72) in training (test) sets, correspondingly.

Although, the most important criterion for the quality of the discriminant model is based on the statistics for the external prediction set, for a more exhaustive testing of the predictive power of the models, we carried out a leave-10-*fold* full-out (LGO) cross-validation procedure. For each group of observations left out (10% of the whole data set, 9 compounds), a model was developed from the remaining 90% of the data (81 compounds). This process was carried out ten times on ten unique subsets. The statistical results are depicted in Table **7**. The overall mean of the correct classification in training (test) set for this process for Eq. **13** and Eq. **14** was 94.05% (90.90%) and 90.87% (89.99%), correspondingly. The result of predictions on the 10% *full* cross-validation test evidenced the quality (robustness, stability and predictive power) of the obtained models.

Later, we also developed the linear discriminant canonical analysis by checking the following statistics: canonical regression coefficient (R_{can}), Chi-squared and its *p*-level [$p(\chi^2)$] [73]. This statistical analysis also permitted us to order these compounds accordingly with their activity profile. Atom-type non-stochastic and stochastic linear indices & LDA antitrichomonal activity canonical analysis principal root are given below:

$$\textit{Classroot} = 2.347 - 1.759\,{}^{s}f_{1L}{}^{H}(x_E) + 1.52\,{}^{s}f_{14L}(x_E) - 1.019\,{}^{s}f_{1L}{}^{H}(x_{E\text{-}H}) + 1.243\,{}^{s}f_{4L}{}^{H}(x_{E\text{-}H}) \quad (\mathbf{15})$$

$$N = 90 \quad \lambda = 0.32 \quad R_{can} = 0.82 \quad \chi^2 = 97.78 \quad \text{Mean}(+) = -1.65 \quad \text{Mean}(-) = 1.26 \quad p<0.0001$$

$$\textit{Classroot} = -0.4146 - 9.396 f_{1L}(x_E) + 10.4439 f_{2L}{}^{H}(x_E) - 2.1972 f_{0L}(x_E) - 0.7323 f_{8L}{}^{H}(x_{E\text{-}H}) \quad (\mathbf{16})$$

$$N = 90 \quad \lambda = 0.308 \quad R_{can} = 0.83 \quad \chi^2 = 101.15 \quad \text{Mean}(+) = -1.69 \quad \text{Mean}(-) = 1.29 \quad p<0.0001$$

The canonical transformation of the LDA results with non-stochastic and stochastic atom-type linear fingerprints gives rise to canonical roots with good canonical regression coefficients of 0.82 and 0.83, respectively. Chi-squared test permits us to assess the statistical signification of this analysis as having a *p*-level <0.0001. The canonical scores of all active and inactive compounds appear in Table **4** and **5**, correspondingly.

On the other hand, a close inspection of the molecular descriptors included in both LDA-based QSAR models showed that several of these molecular fingerprints are strongly interrelated to each other. In Table **8** we give the correlation coefficients of the molecular descriptors in Eqs **13** and **14**.

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 [60-66]. This process is an approach in which molecular descriptors are transformed in such a way that they do not mutually correlate. Both the non-orthogonal

Table 8. Intercorrelation of the molecular descriptors included in the LDA-based QSAR models and results of Randić's orthogonalization analysis.

Non-orthogonal atom-type non-stochastic linear indices				**Non-orthogonal atom-type stochastic linear indices**			
$f_{1L}(x_E)$	$f_{2L}^{H}(x_E)$	$f_{0L}(x_E)$	$f_{8L}^{H}(x_{E\text{-}H})$	${}^{s}f_{1L}^{H}(x_E)$	${}^{s}f_{14L}(x_E)$	${}^{s}f_{1L}^{H}(x_{E\text{-}H})$	${}^{s}f_{4L}^{H}(x_{E\text{-}H})$
1.00	0.97	0.85	0.23	1.00	0.99	0.61	0.59
	1.00	0.92	0.41		1.00	0.54	0.52
		1.00	0.50			1.00	0.98
			1.00				1.00
Orthogonal atom-type non-stochastic linear indices				**Orthogonal atom-type stochastic linear indices**			
${}^{1}O(f_{1L}(x_E))$	${}^{2}O(f_{2L}^{H}(x_E))$	${}^{3}O(f_{0L}(x_E))$	${}^{4}O(f_{8L}^{H}(x_{E\text{-}H}))$	${}^{1}O({}^{s}f_{1L}^{H}(x_E))$	${}^{2}O({}^{s}f_{14L}(x_E))$	${}^{3}O({}^{s}f_{1L}^{H}(x_{E\text{-}H}))$	${}^{4}O({}^{s}f_{4L}^{H}(x_{E\text{-}H}))$
1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
	1.00	0.00	0.00		1.00	0.00	0.00
		1.00	0.00			1.00	0.00
			1.00				1.00
LDA-based model derived with orthogonal atom-type non-stochastic linear indices				LDA-based model derived with orthogonal atom-type stochastic linear indices			
Class = 0.50 +3.57${}^{1}O(f_{1L}(x_E))$ -12.35${}^{2}O(f_{2L}^{H}(x_E))$ +6.04${}^{3}O(f_{0L}(x_E))$ +1.61${}^{4}O(f_{8L}^{H}(x_{E\text{-}H}))$ (**13a**)				***Class*** = -0.96 +3.03${}^{1}O({}^{s}f_{1L}^{H}(x_E))$ -33.43${}^{2}O({}^{s}f_{14L}(x_E))$ +1.02${}^{3}O({}^{s}f_{1L}^{H}(x_{E\text{-}H}))$ -22.69${}^{4}O({}^{s}f_{4L}^{H}(x_{E\text{-}H}))$ (**14a**)			
N = 90 λ = 0.308 D^2 = 8.91 F(4, 85) = 47.558 ***C*** = 0.91 Q_{total} = 95.56 p<0.0001				N = 90 λ = 0.32 D^2 = 8.43 F(4, 85) = 44,999 ***C*** = 0.82 Q_{total} = 91.11 p<0.0001			

descriptors and derived orthogonal descriptors contain the same information. Therefore, the same statistical parameters of the QSAR models are obtained [60-66].

In Table **8** we also resume the results of the orthogonalization of molecular descriptors included in both equations. In this case, the models **13a** and **14a** correspond to the final models with the orthogonalized linear indices. Here, we used the symbols ${}^{m}O(f_k(x))$, where the superscript *m* expresses the order of importance of the variable ($f_k(x)$) after a preliminary forward stepwise analysis and *O* means orthogonal.

It has to be highlighted here that the orthogonal descriptor-based models coincide with the collinear (i.e., ordinary) linear descriptors-based models in all statistical parameters. The statistical coefficients of LDA-QSARs λ, D^2, F, ***C***, accuracy (Q_{total}) are the same whether we use a set of non-orthogonal descriptors or the corresponding set of orthogonal indices (see Table **8**). [60-66].

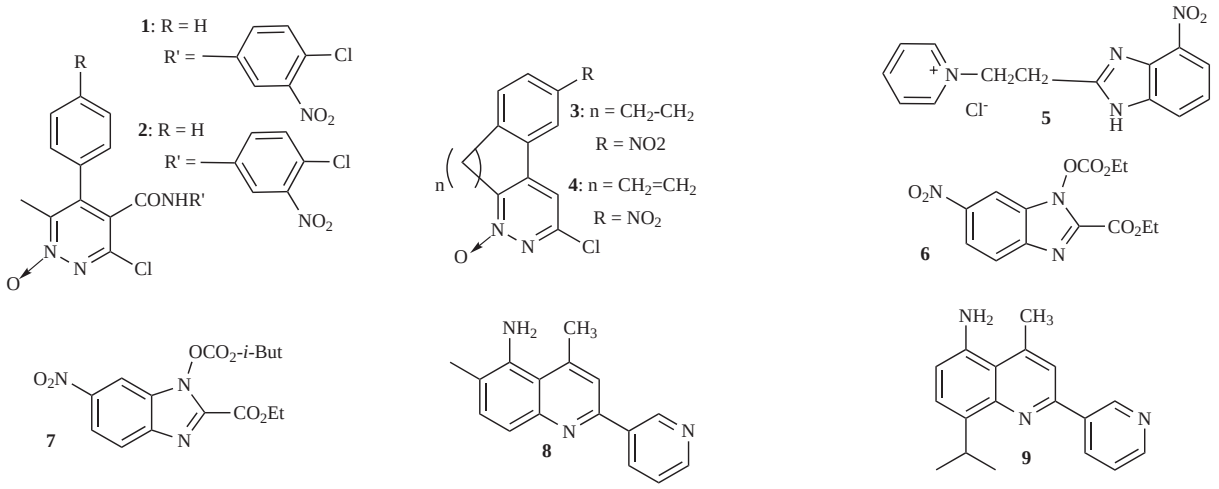
This fact also makes possible the interpretation of the coefficients in the LDA-QSAR equations. Therefore, ${}^{m}O(f_k(x))$ may be classified according to the distance *k* into short- (0-5), mid- (6-10), and long-range non-stochastic and stochastic linear indices. The information in Table 8 clearly shows that the major contribution to antitrichomonal activity is provided by short-range atom-type (heteroatoms and H-atoms bonding to heteroatoms) linear indices. These short-range local descriptors are putative molecular charge, dipole moment, and H-bonding acceptors, and H-bonding donors.

3.3. Simulation of an Experiment of Lead Generation by Computational Screening

In addition to high-throughput screening technology, virtual (*in silico*) screening has become one of the main tools for identifying leads [24,25,54]. Virtual screening is actually one of the computational tools used to filter out unwanted chemicals from physical and/or *in silico* libraries [24,25,54]. Virtual screening techniques may be classified according to their particular modeling of molecular recognition and the type of algorithm used in database searching [24,25,54]. If the target (or at least its active site) 3D structure is known, one of the structure-based virtual screening methods can be applied. By contrast, ligand-based methods are founded on the principle of similarity, that is, similar compounds are assumed to produce similar effects. The absence of a receptor 3D structure is the main reason for the application of ligand-based methods. However, most (Q)SAR methods are focused on one single family of compounds or a specific action mode. Nevertheless, our group has shown that new lead drugs can be designed and/or selected even if their mechanism of action is completely unknown, by using algorithms based on the structural characterization of a structurally diverse database with molecular descriptors and some pattern recognition technologies such as LDA [22,23,35,36,66].

In order to prove the possibilities of the present approach for the virtual screening of trichomonacidal compounds, we have selected a series of 9 compounds whose structures are given in Table **9**. They have been selected from the current

Table 9. Lead identified as Trichomonacidal from literature by using LDA-based QSAR models in simulate virtual screening.



Comp.[a]	Ref.[b]	ΔP%[c]	Score[d]	ΔP%[e]	Score[f]	Antitrichomonal Activity
1	74	100.00	-4.10	100.00	-5.75	MIC = 31.5μg/ml[g]
2	74	100.00	-7.62	100.00	-9.76	MIC = 12.5μg/ml [g]
3	75	100.00	-3.45	100.00	-5.20	MIC= 31.3μg/ml [g]
4	75	100.00	-3.45	100.00	-5.35	MIC= 3.9μg/ml [g]
5	76	99.91	-2.42	99.84	-2.79	MLC =50μg/ml[h] LD_{50} = 50μg/ml[h]
6	67	99.91	-2.43	99.99	-3.9573	100μg/ml = 71.3[I] 10μg/ml = 14.4 [I] 1μg/ml = 0.0 [i]
7	67	99.90	-2.39	99.99	-3.9591	100μg/ml = [87.5] [I] 10μg/ml = 17.3 [I] 1μg/ml = 9.6 [i]
8	69	26.35	-0.01	-93.21	0.82	100 = 58.3[82.3][j] 10 = 29.1[21.6] [j] 1= 18.1[6.8] [j]
9	69	21.22	0.02	-92.26	0.77	100 = 65.4 [73.9] [j] 10 = 56.7 [16.7] [j] 1= 40.1 [0.0] [j]

[a]The molecular structure of the compounds represented with numbers are shown at the top of this Table. [b]bibliographical references where were taken the molecules together with *in vitro* activity. [c]Antitrichomonal activity predicted by Eq **13**; ΔP% = [P(Active) - P(Inactive)]x100. [d]Canonical scores obtained from canonical analysis (Eq. **15**). [e]Antitrichomonal activity predicted by Eq. **14**; ΔP% = [P(Active) - P(Inactive)]x100. [f]Canonical scores obtained from canonical analysis (Eq. **15**). [g]MIC: minimum inhibitory concentration. Most of these compounds showed better activity against *T. vaginalis* than metronidazole (MIC = 25μg/ml). [h]MLC: minimum lethal concentration used that killed all the parasites by 24h. LD_{50}: minimum concentration used that reduced the number of parasites at least 50%. This compound showed a weak inhibitory activity compared with metronidazole (MLC =1μg/ml and LD_{50} =1μg/ml). [i]Percentage of inhibition of *T. Vaginalis* growth at the indicated doses at 24h. In brackets; percentage of reduction. [j]Specific activity against *T. Vaginalis* expressed as

medicinal chemistry literature reporting them as promising trichomonacidal compounds and showing that they are active in one or several experimental assays [67,69,74,75].

In this experiment of 'simulation' of virtual screening, 9 previously reported trichomonacidal compounds with potent-moderate antitrichomonal activity were evaluated with models **13** and **14** as active/inactive ones. The results of the classification are shown in Table **9**.

As can be seen, both models classify correctly most of the 9 selected compounds. In the second case (Eq. **14**) only two lead-compound were classified as false inactives (77.77% of correct classification), while with model **13**, the prediction has an overall accuracy of 100%. This result is the most important validation for the model developed here because it has been able to detect a series of compounds as active from a database composed of compounds selected from literature and these chemicals have shown the predicted activity.

Finally, these compounds taken from the latest literature can be included in the training series in order to derive more

Table 10. Results of the computational evaluation using LDA-based QSAR models and percentages of citostatic and/or citocidal activity [brackets] for the three concentrations assayed in *vitro* against *Trichomonas vaginalis.*

Compound[*]	Theoretical results						*in vitro* activity (μg/ml)[h]					
	Class [a]	ΔP% [b]	Score [c]	Class [dΩ]	ΔP% [e]	Score [f]	Class [g]	$\%CA_{24h}$ [$\%C_{24h}$]			$\%CA_{48h}$ [$\%C_{48h}$]		
								100	10	1	100	10	1
VA1-31	+	99.95	-2.59	-	-99.92	2.35	-	0.00	1.73	8.65	10.39	2.61	3.46
VA1-33	-	-80.98	0.92	-	-99.92	2.35	-	23.38	0.00	0.00	14.72	9.09	0.00
VA2-10	+	99.99	-3.34	+	100.00	-4.02	+	**[98.34]**	**[91.10]**	0.00	**[99.75]**	88.25	12.47
VA2-17	+	99.99	-3.34	+	100.00	-4.02	+	**[98.73]**	76.62	0.00	**[99.92]**	62.88	6.98
VA2-25	+	99.97	-2.74	-	-99.92	2.35	+	**[99.61]**	**[96.48]**	0.00	**[100]**	**[97.38]**	4.86
VA2-26	-	-71.76	0.77	-	-99.92	2.35	-	21.76	0.00	0.00	6.98	0.00	0.00
VA2-38	+	99.98	-2.88	+	100.00	-4.02	+	**[100.00]**	**[99.07]**	69.31	**[100.00]**	**[100.00]**	38.01
VA3-3c	-	-97.13	1.59	-	-99.92	2.35	-	0.00	0.00	0.00	7.01	0.00	0.00
VA3-3f	+	99.62	-1.93	-	-99.92	2.35	-	0.00	0.00	0.00	3.18	0.00	0.00
VA3-8a	-	-97.19	1.59	-	-99.92	2.35	-	78.61	5.23	0.00	84.71	2.34	0.00
VA4-10	+	100.00	-7.38	+	100.00	-6.66	+	**[100.00]**	83.21	0.00	**[100.00]**	28.54	0.00
VA4-18	-	-58.07	0.61	-	-99.92	2.35	-	21.27	0.00	0.00	0.00	11.55	0.00
VA5-5b	+	99.99	-3.31	+	100.00	-4.02	+	**[99.86]**	**[98.37]**	20.81	**[100.00]**	**[97.52]**	2.83
VA5-6	+	100.00	-3.48	+	100.00	-4.02	+	**[100.00]**	**[93.30]**	2.01	**[100.00]**	86.32	1.53
VA5-9a	+	99.99	-3.15	+	99.85	-2.82	+	**[99.51]**	43.25	0.00	**[99.32]**	10.61	0.00
VA5-10	+	100.00	-3.75	+	99.85	-2.82	+	**[97.85]**	24.54	0.00	**[98.83]**	16.03	0.00
VA5-15c	+	100.00	-3.68	+	99.82	-2.75	+	**[99.62]**	0.00	0.00	**[100.00]**	0.00	0.00
VA6-9a	+	99.39	-1.77	+	100.00	-4.19	+	**[95.58]**	12.04	9.85	**[95.94]**	21.82	13.47
VA6-10a	+	99.97	-2.81	+	100.00	-13.07	+	**[97.01]**	18.98	5.84	**[95.26]**	10.82	10.44
VA6-17b	-	-97.48	1.63	+	100.00	-8.34	-	40.51	18.61	0.00	43.83	10.44	0.00
VA6-22	+	99.98	-2.89	+	99.99	-3.86	+	91.51	17.15	16.42	**[96.41]**	14.99	10.82
MTZ	+	97.23	-1.26	+	91.69	-1.41	+	**[99.63]**	**[99.18]**	**[98.19]**	**[100]**	**[99.72]**	**[98.79]**

[*]The molecular structures of the compounds represented with codes are shown in Figure **5**. [a,d]*In silico* classification obtained from models **13** and **14** using non-stochastic and stochastic atom-type linear indices, respectively. [b,e]Results of the classification of compounds obtained from model **13** and **14**, correspondingly: ΔP% = [P(Active) - P(Inactive)]x100. [c,f]Canonical scores obtained from models **15** and **16**, correspondingly. [g]Observed (experimental activity) classification against *T. vaginalis* . [h]Pharmacology activity of each tested compounds, which were added to the cultures at doses of 100, 10 and 1μg/ml: $\%CA_{\#}$ = Citostatic activity$_{(24\ ó\ 48\ hours)}$ and **[$\%C_{\#}$]** = Citocidal activity$_{(24\ ó\ 48\ hours)}$. MTZ = Metronidazole (concentration for metronidazole were 2, 1 and 0.5 mg/ml, respectively).

robust classification models. By this means, the derivation of the classifier model is considered as an iterative process in which novel compounds with novel structural features are incorporated into the training set for improving the quality of the models so developed [22].

3.4. Ligand-Based Virtual *in Silico* Screening and Lead Trichomonacidals Discovery

In order to test the potential of TOMOCOMD-CARDD method and LDA for detecting novel antiprotozoal leads, we predicted the biological activity of all the chemicals contained in our 'in-house' collections of indazole, indole, cinnoline and quinoxaline derivatives, which have been recently obtained by our chemical synthesis team [67,77-84]. On the basis of computer-aided predictions we selected potential trichomonacidal compounds (virtual hits). The following criteria were used for the hits' selection: 1) compounds were selected as hits if the value of posterior probability of possessing antitrichomonal activity exceeded 97% (ΔP≥95%) by both LDA-based QSAR models, and 2) If, among the compounds designed by our chemical team, too many similar compounds satisfied criterion 1, then only several representative structures were selected. Some compounds classified as inactive by both LDA-based QSAR models were also *in vitro* tested.

The structures of potential trichomonicidal (and inactives ones) from different chemical series (VA1-VA6), selected on the basis of these criteria, are presented in Figure **5**. Later, all selected hits were re-synthesized and experimentally tested for their antitrichomonal effect. The results of the activity against *T. vaginalis* of the compound study objects are shown in the Table **10**. Our trained LDA-based QSAR models (Eq. **13** and **14**) successfully classified 19 out of 21 compounds yielding an accuracy of the 90.48% (19/21).

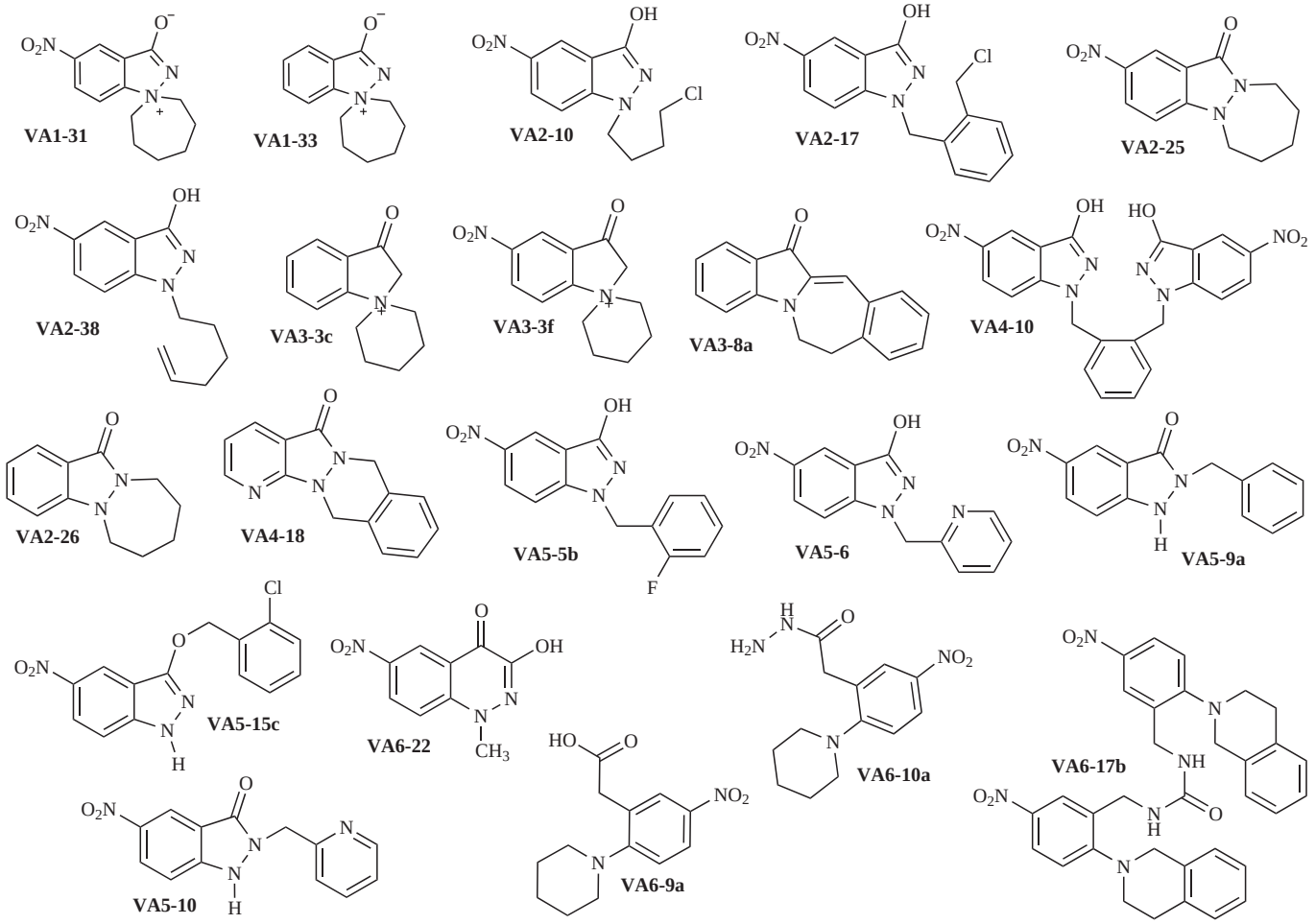


Fig. (5). Chemical structures of the selected compounds.

These outcomes exemplify how the present approach could be used for the selection/identification of new trichomonacidal drug candidates and also to remove undesired chemicals as early as possible.

In general, the compounds VA2-25, VA2-38 and VA5-5b that belong to the fused indazolinone, 1-hexenylindazolols, and 1-benzylindazolols series maintain their efficacy at 10 μg/ml with an important trichomonacidal (cytocidal) activity at 48h (24h): 97.38% (96.48%), 100.00% (99.07%) and 97.52% (98.37%), respectively. It is remarkable that these compounds did not show toxic activity in macrophages cultivations at these concentrations.

Other two compounds, VA2-10 and VA5-6 (1-substituted indazolols) showed appreciable activity at the concentration of 10 μg/ml at 24h and low non-specific cytotoxicity [85]. On the other hand, eight compounds (VA2-17, VA4-10, VA5-9a, VA5-10, VA5-15c, VA6-9a, VA6-10a, and VA6-22) showed activity against *T. vaginalis* between 96% and 100% at the concentration of 100 μg/ml and low cytotoxicity at this concentration [85]. All these compounds can be considered as new antitrichomonal agents. Even so, none of the studied compounds was more active than metronidazole. Our current results are significant because they demonstrate the straightforward way in which TOMOCOMD-CARDD method can identify new trichomonacidal leads.

4. CONCLUDING REMARKS

The main conclusion of this work is that it has been able to develop *in silico* models for one of the main steps of drug discovery: lead selection (or generation). These models permitted the identification of new hits and lead drug-like chemicals with antitrichomonas activity. The straightforward way in which TOMOCOMD-CARDD method can identify new trichomonacidal leads demonstrates a significant aspect of our work.

ACKNOWLEDGEMENTS

One of the authors (M-P. Y) thanks the program 'Estades Temporals per a Investigadors Convidats' for a fellowship to work at Valencia University. F.T. thanks support from Spanish MEC DGI (Project No. CTQ2004-07768-C02-01/BQU) and Generalitat Valenciana (DGEUI INF01-051 and INFRA03-047, and OCYT GRUPOS03-173).

REFERENCES

[1] Brown M.T.: Trichomoniasis. Practitioner 209, 639, (1972).

[2] Catterall R.D.: Trichomonal infections of the genital tract. Med. Clin. North Am. 56, 1203, (1972).

[3] Wisdom A.R., Dunlop E.M.C.: Trichomoniasis: study of the disease and its treatment in women and men. Br. J. Vener. Dis. 4, 90, (1965).

[4] Cates W.Jr.: Estimates of the incidence and prevalence of sexually transmitted diseases in the United States. Sex. Transm. Dis. *26*(Suppl.), S2, (1999).

[5] World Health Organization. An overview of selected curable sexually transmitted diseases, 1995, p. 2–27. *In* Global program on AIDS. World Health Organization, Geneva, Switzerland.

[6] Gram I., Macaluso M., Churchill J., Stalsberg H.: *Trichomonas vaginalis* (TV) and human papillomavirus (HPV) infection and the incidence of cervical intraepithelial neoplasia (CIN) grade III. Cancer Causes Control 3, 231, (1992).

[7] Kharsany A.B.M., Hoosen A.A., Moodley J., Bagaratee J., Gouws E.: The association between sexually transmitted pathogens and cervical intra-epithelial neoplasia in a developing community. Genitourin. Med. 69, 357, (1993).

[8] Zhang Z.-F., Begg C.B.: Is *Trichomonas vaginalis* a cause of cervical neoplasia? Results from a combined analysis of 24 studies. Int. J. Epidemiol. 23, 682, (1994).

[9] Cates W., Joesoef M.R., Goldman M.B. Atypical pelvic inflammatory disease: can we identify clinical predictors? Am. J. Obstet. Gynecol. 169, 341, (1993).

[10] Grodstein F., Goldman M.B., Cramer D.W.: Relation of tubal infertility to a history of sexually transmitted diseases. Am. J. Epidemiol. 137, 577, (1993).

[11] Soper D.E., Bump R.C., Hurt W.G.: Bacterial vaginosis and trichomoniasis vaginitis are risk factors for cuff cellulitis after abdominal hysterectomy. Am. J. Obstet. Gynecol. 163, 1016, (1990).

[12] Cotch M.F.: Vaginal infections and prematurity study group. Carriage of *Trichomonas vaginalis* (Tv) is associated with adverse pregnancy outcome, abstr. 681. *In* Program and abstracts of the 30th Interscience Conference on Antimicrobial Agents and Chemotherapy. American Society for Microbiology, Washington, D.C. 1990.

[13] Minkoff H., Grunebaum A.N., Schwarz R.H., Feldman J., Cummings M., Crombleholme W., Clark L., Pringle G., McCormack W.M.: Risk factors for prematurity and premature rupture of membranes: a prospective study of the vaginal flora in pregnancy. Am. J. Obstet. Gynecol. 150, 965, (1984).

[14] Fowler K.B., Pass R.F.: Sexually transmitted diseases in mothers of neonates with congenital cytomegalovirus infection. J. Infect. Dis. 164, 259, (1991).

[15] Laga M., Manoka A., Kivuvu M., Malele B., Tuliza M., Nzola N., Goeman J., Behets F., Batter V., Alary M., Heyward W.L., Ryder R.W., Piot P.: Non-ulcerative sexually transmitted diseases as risk factors for HIV-1 transmission in women: results from a cohort study. AIDS (London). 7, 95, (1993).

[16] Durel P., Roiron V., Siboulet A., Borel L.J.: ystemic treatment of human trichomoniasis with a derivative of nitroimidazole 8823 R.P. Br. J. Vener. Dis. 36, 21, (1960).

[17] Centers for Disease Control and Prevention**.:** Sexually transmitted diseases treatment guidelines. Morb. Mortal. Wkly. Rep. 42(RR-14), 70, (1993).

[18] Robinson S.C.: Trichomonal vaginitis resistant to metronidazole. Can. Med. Assoc. J. 86, 665, (1962).

[19] Sobel J.D., Nyirjesy P., Brown W.: Tinidazole therapy for metronidazole-resistant vaginal trichomoniasis. Clin. Inf. Dis. 33, 1341, (2001).

[20] Lumsden W.H.R., Robertson D.H.H., Heyworth R., Harrison C.: Treatment failure in *Trichomonas vaginalis* vaginitis. Genitourin. Med. 64, 217, (1988).

[21] Narcisi E.M., Secor W.E. : In vitro effect of tinidazole and furazolidone on metronidazole-resistant *Trichomonas vaginalis*. Antimicrob. Agents Chemother. 40, 1121, (1996).

[22] Estrada E., Peña A.I.: In Silico Studies for the Rational Discovery of Anticonvulsant Compounds. Bioorg. Med. Chem. 8, 2755, (2000).

[23] Estrada E., Uriarte E., Montero A., Teijeira M., Santana L., De Clercq E.A.: A Novel Approach for the Virtual Screening and Rational Design of Anticancer Compounds**.** J. Med. Chem. 43, 1975, (2000).

[24] Scott R.K.: Informatics integration: the bedrock of NCE selection. Biosilico. 1, 14, (2003).

[25] Seifert M.H.J., Wolf K., Vitt D.: Virtual high-throughput *in silico* screening Biosilico. 1, 143, (2003).

[26] 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

[27] 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. 8, 687, (2003).

[28] 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. 44, 2010, (2004).

[29] 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. 12, 6351, (2004).

[30] 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. 9, 1100, (2004).

[31] 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. 12, 5331, (2004).

[32] 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. 4, 512, (2003).

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

[34] 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 Electron. J. Mol. Des. 4, 124, (2005).

[35] 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. 13, 1005, (2005).

[36] 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. Design. 18, 615, (2004).

[37] 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. Mol. Struct. (Theochem) 717, 67, (2005).

[38] 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. 13, 1293, (2005).

[39] Marrero-Ponce Y., Medina-Marrero R., Torrens F., Martinez Y., Romero-Zaldivar V., Castro E.A.: Atom, Atom-type, and Total Non-Stochastic and Stochastic Quadratic Fingerprints: A Promising Approach for Modeling of Antibacterial Activity. Bioorg. Med. Chem. 13, 2881, (2005).

[40] Marrero-Ponce Y., Medina-Marrero R., Martinez Y., Torrens F., Romero-Zaldivar V., Castro E.A. J. Mol. Mod. In Press.

[41] 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. 5, 276, (2004).

[42] 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. 13, 3397, (2005).

[43] 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. 9, 1124, (2004).

[44] 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. 13, 3003, (2005).

[45] Klein D.J.: Graph Theoretically Formulated Electronic–Structure Theory. Internet Electron. J. Mol. Des. 2, 814, (2003).

[46] Edwards C.H., Penney D.E.: Elementary Linear Algebra. Prentice-Hall, Englewood Cliffs: New Jersey, USA, (1988).

[47] Kier L.B, Hall L.H.: Molecular Connectivity in Structure–Activity Analysis, Research Studies Press: Letchworth, U.K, (1986).

[48] Pauling L.: The Nature of Chemical Bond, Cornell University Press: New York, (1939).

[49] Dmitriev I.S. Molecules Without Chemical Bonds, Mir publishers: Moscow, (1981).

[50] Negwer M. Organic-Chemical Drugs and their Synonyms, Akademie-Verlag: Berlin, (1987).

[51] Chapman & Hall. The Merck Index. Twelfth Edition, (1996).

[52] Mc Farland J.W. Gans D.J.: In Chemometric Methods in Molecular Design; van de Waterbeemd, H., Ed.; VCH Publishers: Weinheim, pp 295–307, 1995.

[53] Johnson R.A., Wichern, D.W.: Applied Multivariate Statistical Analysis. Prentice-Hall: Englewood Cliffs (NJ), (1988).

[54] Xu J., Hagler A.: Chemoinformatics and Drug Discovery. Molecules. 7, 566, (2002).

[55] STATISTICA vs. 5.5, StatSoft, Inc. (1999).

[56] van de Waterbeemd H.: In Chemometric Methods in Molecular Design, van Waterbeemd, H., Ed.; VCH Publishers: Weinheim, pp. 265-288, (1995).

[57] Estrada E., Patlewicz G.: On the Usefulness of Graph-theoretic Descriptors in Predicting Theoretical Parameters. Phototoxicity of Polycyclic Aromatic

Hydrocarbons (PAHs). Croat. Chem. Acta. 77, 203, (2004).
[58] Wold S, Erikson L. In Chemometric Methods in Molecular Design, van Waterbeemd, H., Ed.; VCH Publishers: Weinheim, pp. 309-318, (1995).
[59] Baldi P., Brunak S., Chauvin Y., Andersen C.A., Nielsen H. Assessing the Accuracy of Prediction Algorithms for Classification: an Overview. Bioinformatics. 16, 412, (2000).
[60] Randić M.J.: Resolution of Ambiguities in Structure-Property Studies by Us of Orthogonal Descriptors. Chem. Inf. Comput. Sci. 31, 311, (1991).
[61] Randić M.: Orthogonal Molecular Descriptors. New J. Chem. 15, 517, (1991).
[62] Randić M.J.: Correlation of Enthalpy of Octanes with Orthogonal Connectivity Indices. Mol. Struct. (Theochem) 233, 45, (1991).
[63] Lučić B., Nikolić S., Trinajstić N., Jurić D.: The Structure-Property Models can be Improved Using the Orthogonalized Descriptors. J. Chem. Inf. Comput. Sci. 35, 532, (1995).
[64] Klein D.J., Randić M., Babić D., Lučić B., Nikolič S., Trinajstič N.: Hierarchical Orthogonalization of Descriptors. Int. J. Quantum Chem. 63, 215, (1997).
[65] 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. 42, 1194, (2002).
[66] Estrada E., Uriarte E.: Recent Advances on the Role of Topological Indices in Drug Discovery Research. Curr. Med. Chem. 8, 1573, (2001).
[67] Aguirre G., Boiani M., Cerecetto H., Gerpe A., González M., Fernandez-Sainz Y., Denicola A., Ochoa de Ocáriz C., Nogal J.J., Montero D., Escario J.A.: Novel antiprotozoal products: imidazole and benzimidazole N-oxide derivatives and related compunds. Arch. Pharm. (Weinheim, Ger.) 337, 259, (2004).
[68] Kouznetsov V.V, Rivero C. J., Ochoa P.C., Stashenko E., Martínez J.R., Montero P. D., Nogal J.J., Fernández P.C., Muelas S.S., Gómez B.A., Bahsas A., Amaro L.J.: Synthesis and antiparasitic properties of new 4-N-benzylamino-4-hetarylbut-1-enes. Arch. Pharm. (Weinheim, Ger.) 338, 1, (2005).
[69] Kouznetsov V.V., Vargas-Mendez L.Y., Tibaduiza B., Ochoa C., Montero P.D., Nogal J.J., Fernández C., Muelas S., Gómez A., Bahsas A., Amaro-Luis J.: Aryl(benzyl)amino-4-heteroarylbut-1-enes as building blocks in heterocyclic synthesis. Synthesis of 4,6 Dimethyl-5-nitro(amino)-2-pyridylquinolines and their antiparasitic activities. Arch. Pharm. (Weinheim, Ger.) 337, 127, (2004).
[70] Ochoa A., Pérez E., Pérez R., Suárez M., Ochoa E., Rodríguez H., Gómez A., Muelas S., Nogal J.J., Martínez R.A.: Synthesis and antiprotozoan properties of new 3,5 disubstitued tetrahydro-2H-1,3,5-thiadiazine-2-thione derivatives. Arzneim.-Forsch./Drug Res. 49, 764, (1999).
[71] Meneses A., Montero D., Escario J.A., Nogal-Ruiz J.J., Ochoa C., Arán V.J., Martínez-Fernández A.R.: Trichomonacidal activity of some heterocyclic head-series compounds: indole, indazole, and quinoxaline derivatives. IX European multicolloquium of Parasitology.Valencia, Spain. 18-23, July (2004).
[72] Gálvez J., García,R., Salabert M.T., Soler R.: Charge Indexes. New Topological Descriptors. J. Chem. Inf. Comput. Sci. 34, 520, (1994).
[73] Ford M.-G., Salt D.-W.: In chemometric methods in molecular design, van de Waterbeemd, H., Ed.; VCH Publishers: New York, pp. 283-292, (1995).
[74] Gavini E., Juliano C., Mulé A., Pirisino G.: Pyridazine N-oxides. II. Síntesis and in invitro antimicrobial evaluation of 3-chloro-4-carbamoyl-5-aryl-6-methyl-pyridazine N-oxides. IL Fármaco. 52, 67, (1997).
[75] Gavini E., Juliano C., Mulé A., Pirisino G., Murineddu G., Pinna A.: Pyridazine N-oxides. III. Synthesis and in vitro antimicrobial properties of N-oxide derivatives based on tricyclic indeno[2,1-*c*]pyridazine and benzo[*f*]cinnoline systems. Arch. Pharm. (Weinheim, Ger.) 333, 341, (2000).
[76] Alcalde E., Pérez L., Dinarés I., Frigola J.: Heterocyclic Betaines. XXII. Azinium(Azolium) 4-nitrobenzimidazolate inner salts and their derivatives with several interannular spacers. Synthesis, characterization and antitrichomonal activity. Chem.Pharm.Bull. 43, 493, (1995).
[77] Arán V.J., Asensio J.L., Ruiz J.R., Stud M.: The heterocyclization of N',N'-disubstituted 2-halogenobenzohydrazides to 1,1-disubstituted indazol-3-ylio oxides. J. Chem. Res. (M) 1319, (1993).
[78] Arán V.J., Asensio J.L., Ruiz J.R., Stud M.: Reactivity of 1,1-disubstituted indazol-3-ylio oxides: Synthesis of some substituted indazolols and indazolinones. J. Chem. Soc., Perkin Trans. 1. 1119, (1993).
[79] Ruiz J.R., Arán V.J., Asensio J.L., Flores M., Stud M.: Synthesis of quaternary indoxyl derivatives by intramolecular cyclization of some substitud acetophenones. Liebigs Ann. Chem. 679, (1994).
[80] Arán V.J., Flores M., Muñoz P., Ruiz J.R., Sánchez-Verdú P., Stud M.: Cytostatic activity against HeLa cells of a series of indazole and indole derivatives; sinthesis and evaluation of some analogues. Liebigs Ann. Chem. 817, (1995).
[81] Arán V.J., Flores M., Muñoz P., Páez J.A., Sánchez-Verdú P., Stud M.: Analogues of cytostatic, fused indazolinones: Synthesis, conformational analysis and cytostatic activity against HeLa cells of some 1-substituted indazolols, 2-substituted indazolinones, and related compounds. Liebigs Ann. Chem. 683, (1996).
[82] Arán V.J., Asensio J.L., Molina J., Muñoz, P., Ruiz J.R., Stud M.: Approaches to 1,1-disubstituted cinnolin-3-ylio oxides: Synthesis and reactivity of a new class of heterocyclic betaines. J. Chem. Soc., Perkin Trans.1. 2229, (1997).
[83] Castro S., Chicharro R., Arán V.J.: Synthesis of quinoxaline derivatives from substituted acetanilides through intramolecular quaternization reactions. J. Chem. Soc., Perkin Trans. 1. 790, (2002).

[84] Chicharro R., Castro S., Reino J.L., Arán V.J.: Synthesis of tri- and tetracyclic condensed quinoxalin-2-ones fused across the C-3-N-4 Bond. Eur. J. Org. Chem. 2314, (2003).

[85] Rolón M., Vega C., Gómez-Barrio A., Ochoa C., Arán V.J., Martinez- Fernandez A.R. In vitro antitrypanosomal activity of heterocyclic compounds: indole,indazole and quinoxaline derivatives. IX European multicolloquium of Parasitology.Valencia, Spain. 18-23, July (2004).

View publication stats