

Atom, Atom-Type, and Total Linear Indices of the “Molecular Pseudograph’s Atom Adjacency Matrix”: Application to QSPR/QSAR Studies of Organic Compounds

Yovani Marrero Ponce ^{1,2,*}, Juan Alberto Castillo Garit ^{2,3}, Francisco Torrens ⁴, Vicente Romero Zaldivar ⁵ and Eduardo A. Castro ⁶

¹ Department of Pharmacy, Faculty of Chemical-Pharmacy, Central University of Las Villas, Santa Clara, 54830, Villa Clara, Cuba

² Department of Drug Design, Chemical Bioactive Center, Central University of Las Villas, Santa Clara, 54830, Villa Clara, Cuba

³ Applied Chemistry Research Center, Central University of Las Villas, Santa Clara, 54830, Villa Clara, Cuba

⁴ Institut Universitari de Ciència Molecular, Universitat de València, Dr. Moliner 50, E-46100 Burjassot (València), Spain

⁵ Faculty of Informatics, University of Cienfuegos, Cienfuegos, Cuba

⁶ INIFTA, División Química Teórica, Suc.4, C.C. 16, La Plata 1900, Buenos Aires, Argentina

* Author to whom correspondence should be addressed; Fax: (+53)-42-281130/281455; Telephone: (+53)-42-281192/281473; E-mail: yovanimp@qf.uclv.edu.cu or ymarrero77@yahoo.es

Received: 4 May 2004 / Accepted: 13 October 2004 / Published: 31 December 2004

Abstract: In this paper we describe the application in QSPR/QSAR studies of a new group of molecular descriptors: atom, atom-type and total linear indices of the molecular pseudograph’s atom adjacency matrix. These novel molecular descriptors were used for the prediction of boiling point and partition coefficient ($\log P$), specific rate constant ($\log k$), and antibacterial activity of 28 alkyl-alcohols and 34 derivatives of 2-furylethylenes, respectively. For this purpose two quantitative models were obtained to describe the alkyl-alcohols’ boiling points. The first one includes only two total linear indices and showed a good behavior from a statistical point of view ($R^2 = 0.984$, $s = 3.78$, $F = 748.57$, $q^2 = 0.981$, and $s_{cv} = 3.91$). The second one includes four variables [3 global and 1 local (heteroatom) linear indices] and it showed an improvement in the description of physical property ($R^2 = 0.9934$, $s = 2.48$, $F = 871.96$, $q^2 = 0.990$, and $s_{cv} = 2.79$). Later, linear multiple regression analysis was also used to describe $\log P$ and $\log k$ of the 2-furyl-

ethylenes derivatives. These models were statistically significant [$(R^2 = 0.984, s = 0.143,$ and $F = 113.38)$ and $(R^2 = 0.973, s = 0.26$ and $F = 161.22)$, respectively] and showed very good stability to data variation in leave-one-out (LOO) cross-validation experiment [$(q^2 = 0.938$ and $s_{cv} = 0.178)$ and $(q^2 = 0.948$ and $s_{cv} = 0.33)$, respectively]. Finally, a linear discriminant model for classifying antibacterial activity of these compounds was also achieved with the use of the atom and atom-type linear indices. The global percent of good classification in training and external test set obtained was of 94.12% and 100.0%, respectively. The comparison with other approaches (connectivity indices, total and local spectral moments, quantum chemical descriptors, topographic indices and E-state/biomolecular encounter parameters) reveals a good behavior of our method. The approach described in this paper appears to be a very promising structural invariant, useful for QSPR/QSAR studies and computer-aided “rational” drug design.

Keywords: *TOMOCOMD-CARDD*, Total and Local Linear Indices, QSPR, QSAR, alkyl-alcohol, 2-furylethylene.

Introduction

The graph-theory approach appears to be an important alternative to computer-aided molecular design methods. They provide for the discovery of new lead drugs at minimum cost [1]. The high cost of development of new bioactive molecular entities using traditional methods has led to the interest of the pharmaceutical industry in “rational” drug design assisted by computers. This is manifested by the gradually growing interest shown by these companies in quantitative studies of Structure-Activity/Property Relationships (QSAR/QSPR) directed to the rationalization of the search for new biologically active molecules. In this sense, rational combinatorial library design [2] and virtual screening [3] have emerged as important foci of attention in drug discovery research.

An important part of QSAR/QSPR research is the discovery of molecular descriptors applicable to physical, chemical and biological properties of interest. At present, there are a great number of molecular descriptors that can be used in QSAR/QSPR studies [4]. The so-called topological indices (TIs) are among the most useful molecular descriptors known nowadays [5-10]. TIs can be classified as “global” and “local”, according to the way in which they characterize the molecular structure [11]. However, most TIs known today can be considered as global molecular descriptors. One exception in this sense is the electrotopological state (E-state) index [12-13]. Other “global” molecular descriptors, such as the spectral moments of the edge adjacency matrix, can be obtained in local form [11]. The great success of the E-state and total and local spectral moments in QSPR/QSAR stimulated us to propose and validate here some novel local descriptors based on a topological characterization of the molecular structure.

In this sense, our research group has recently introduced the novel computer-aided molecular design scheme *TOMOCOMD-CARDD* (acronym of *TO*topological *MO*lecular *COM*puter *D*esign-*COM*puter *A*ided “*R*ational” *D*rug *D*esign) [14-16]. This method has been developed to generate molecular descriptors based on the linear algebra theory. This approach has been successfully employed in QSPR/QSAR studies [15-18], including studies related to nucleic acid-drug interactions

[19]. The approach describes changes in the electron distribution with time throughout the molecular backbone. The **TOMOCOMD-CARDD** strategy is very useful for the selection of novel subsystems of compounds having a desired property/activity, which can be further optimized by using some of the many molecular modeling methods at the disposition of the medicinal chemists. The method has also demonstrated flexibility in relation to many different problems. One of the applications involved the prediction of the anthelmintic activity of novel drugs [20]. More recently, the **TOMOCOMD-CARDD** approach has been applied to the fast-track experimental discovery of novel antimalarial compounds [21]. Codification of chirality and other 3D structural features constitutes another advantage of this method [22]. The latter opportunity has allowed the description of the significance-interpretation and the comparison to other molecular descriptors [16,23]. The features of the k^{th} total and local linear and quadratic indices was illustrated by examples of various types of molecular structures, including chain-lengthening, branching, heteroatoms-content, and multiple bonds. Besides, the linear independence of the quadratic and linear indices to others 0D, 1D, 2D, and 3D molecular descriptors is demonstrated by using principal component analysis for several heterogeneous molecules [16,23].

The main objective of the present paper was to test the QSPR/QSAR applicability of the **TOMOCOMD-CARDD** approach; and in a second place, to compare the results obtained with other cheminformatic methods in order to assess it. For this purpose, we will develop quantitative models to describe the boiling points of alkyl alcohols and the partition coefficient ($\log P$), specific rate constant ($\log k$) and antibacterial activity of 34 derivatives of 2-furylethylenes.

Theoretical Approach

The current approach is based on the calculation of the linear indices of the molecular pseudograph's atom (vertex) adjacency matrix. The general principle of this approach for small-to-medium size organic compounds has been explained in some detail elsewhere [16]. However, in this paper we offer a global consideration of this approach.

First, the molecular vector (X) is built to calculate the linear indices of a molecule, where the components of this vector are numeric values that represent a certain atomic property. These properties characterize each atom type in the molecule. Some of these properties can be the electronegativity, density, atomic radii, among others. For example the Mulliken electronegativity (X_A) [24] of an atom A takes the values $X_H = 2.2$ for Hydrogen, $X_C = 2.63$ for Carbon, $X_N = 2.33$ for Nitrogen, $X_O = 3.17$ for Oxygen, $X_{Cl} = 3.0$ for Chlorine, and so on. Therefore, a molecule having 5, 10, 15, ..., n atoms can be represented by means of vectors, with 5, 10, 15, ..., n components, belonging to the spaces \mathfrak{R}^5 , \mathfrak{R}^{10} , \mathfrak{R}^{15} , ..., \mathfrak{R}^n , respectively, where n is the dimension of the real sets (\mathfrak{R}^n).

This focus allows us to code molecules like acetic acid (suppressed H-atoms) through the molecular vector $X = [X_C, X_C, X_O, X_O] = [2.63, 2.63, 3.17, 3.17]$, in the X_A -electronegativity scale [24]. This vector belongs to the product space of \mathfrak{R}^4 . The use of other atomic properties defines other vectors. In this context, total (and local) linear indices include "bulk" properties and physicochemical properties (such hydrophobicity [25], molecular polar surface area [26], molar refractivity [27], molecular polarizability [28] and atomic charge summatory [29]), if some atomic physicochemical parameters (such as atomic Log P [25], surface contributions of polar atoms [26], atomic molar refractivity [27], atomic hybrid polarizabilities [28], and Gasteiger-Marsilli atomic charge [29],

respectively) are considered as atom-property (atom-label) for building the n -dimensional molecular vector, X .

Local (Atom) Linear Indices of the “Molecular Pseudograph’s Atom (Vertex) Adjacency Matrix”

If we have a molecule composed by n atoms (vector of \mathfrak{R}^n), then the k^{th} atom linear indices, $f_k(x_i)$, will be calculated as linear maps in \mathfrak{R}^n [$f_k(x_i): \mathfrak{R}^n \rightarrow \mathfrak{R}^n$; thus $f_k(x_i): \text{End on } \mathfrak{R}^n$] in the canonical bases of this space as is shown in Eq. 1,

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

where, ${}^k a_{ij} = {}^k a_{ji}$ (symmetric square matrix), n is the number of atoms of the molecule, and X_j are the coordinates of the molecular vector (X) in a set of basis vectors of \mathfrak{R}^n . One can choose the basis vectors; the coordinates of the same vector will be different [30-33]. The values of the coordinates depend thus in an essential way on the choice of the basis. With the so-called canonical (‘natural’) basis, e_j denotes the n -tuple having 1 in the j^{th} position and 0’s elsewhere. In the canonical basis, the coordinates of any vector X coincide with the components of this vector [30-33]. For this reason, those coordinates can be considered as weights (atom labels) of the vertices of the molecular pseudograph [15-23].

The coefficients ${}^k a_{ij}$ are the elements of the k^{th} power of the matrix $\mathbf{M}(G)$ of the molecular pseudograph (G). The term pseudograph in chemical graph-theory was introduced by Frank Harary [34]. According to him, a pseudograph is a graph with multiple edges or loops between the same vertices or the same vertex. Loop-multigraph [35] or general graphs [36] are other terms also used in this research area [37].

Here, $\mathbf{M}(G) = \mathbf{M} = [a_{ij}]$, denotes the matrix of $f_k(x_i)$ with respect to the natural basis. In this matrix n is the number of vertices (atoms) of G and the elements a_{ij} are defined as follows [15-23]:

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

where $E(G)$ represents the set of edges of G . In this adjacency matrix $\mathbf{M}(G)$ the row i and column i correspond to vertex v_i from G . P_{ij} is the number of edges between vertices v_i and v_j , and L_{ii} is the number of loops in v_i .

Given that $a_{ij} = P_{ij}$, the elements a_{ij} of this matrix represent the number of bonds between an atom i and other j . The matrix \mathbf{M}^k provides the number of walks of length k that link 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 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 between vertices v_i and v_j , respectively. On the other hand, molecules containing aromatic rings with more than one canonical structure are represented by a pseudograph. This happens for substituted aromatic compounds such as pyridine, naphthalene, quinoline, and so on, where the presence of $\text{PI}(\pi)$ electrons are accounted for 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 by a multigraph.

It should be noted that atom's linear indices are defined as a linear transformation $f_k(x_i)$ on an molecular vector space \mathfrak{R}^n . This map is a correspondence that assigns to every vector X in \mathfrak{R}^n a vector $f(x)$ in such a way that:

$$f(\lambda_1 X_1 + \lambda_2 X_2) = \lambda_1 f(X_1) + \lambda_2 f(X_2) \quad (3)$$

for any scalar λ_1, λ_2 and any vector X_1, X_2 in \mathfrak{R}^n . The defining equation (1) for $f_k(x_i)$ may be written as the single matrix equation:

$$f_k(x_i) = \begin{bmatrix} X'_1 \\ \vdots \\ X'_n \end{bmatrix}^k = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix}^k \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} \quad (4)$$

or in the more compact form,

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

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

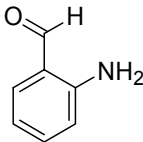
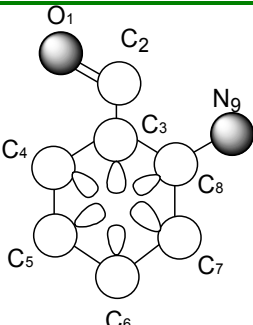
It should also be noted that this approach is rather similar to the **LCAO-MO** (Linear Combinations of Atomic Orbitals-Molecular Orbitals) method. Reality, the approach (for $k = 1$) is a quite similar approximation to the extended Hückel MO method, due to the formalism each MO ψ_i is composed of n valence AOs of atoms in a molecule.

The main idea of the **LCAO-MO** method is that the electrons in a molecule are accommodated in definite MOs just as in an atom where they are accommodated in definite AOs. Normally MOs made up as LCAO of atoms composing the system, i.e. are written in the form,

$$\psi_i = \sum_{j=1}^n c_{ij} \varphi_j \quad (6)$$

where i is the number of the MO ψ [in our case, $f_i(x_i)$]; j are the numbers of atomic φ -orbitals (in our case, X_j); and c_{ij} (in our case, $^1a_{ij}$) are the numerical coefficients defining the contributions of individuals AOs into the given MO. Such a way of constructing a MO is based on the assumption that an atom represented by a definite set of orbitals remains distinctive in the molecule.

Table 1. Definition and Calculation of Total (whole-molecule) and Local (Atom) Linear Indices of the Molecular Pseudograph's Atom Adjacency Matrix of the 2-Aminobenzaldehyde Molecule.

 <p>2-Amino-benzaldehyde</p>		$\mathbf{X} = [O_1, C_2, C_3, C_4, C_5, C_6, C_7, C_8, N_9]$ Molecular Vector: $\mathbf{X} \in \mathfrak{R}^9$ In the definition of the \mathbf{X} , as molecular vector, the chemical symbol of the element is used to indicate the corresponding electronegativity value. That is: if we write O it means $\chi(O)$, oxygen Mulliken electronegativity or some atomic property, which characterizes each atom in the molecule. Therefore, if we use the canonical basis of \mathfrak{R}^9 , the coordinates of any vector X coincide with the components of that molecular vector $[X] = [3.17, 2.63, 2.63, 2.63, 2.63, 2.63, 2.63, 2.63, 2.33]$ $[X]$: column vector of coordinates of \mathbf{X} in the Canonical base of \mathfrak{R}^9 (a $n \times 1$ matrix)
<p>Molecular Structure</p>	<p>Molecular Pseudograph (G) (hydrogen suppressed-pseudograph)</p>	

$$f_1(x_i) = \sum_{j=1}^n a_{ij} X_j = \mathbf{M}^1[\mathbf{X}] = \begin{bmatrix} 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 2 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} O_1 \\ C_2 \\ C_3 \\ C_4 \\ C_5 \\ C_6 \\ C_7 \\ C_8 \\ N_9 \end{bmatrix} = \begin{bmatrix} 2C_2 \\ 2O_1 + 1C_3 \\ 1C_2 + 1C_3 + 1C_4 + 1C_8 \\ 1C_3 + 1C_4 + 1C_5 \\ 1C_4 + 1C_5 + 1C_6 \\ 1C_5 + 1C_6 + 1C_7 \\ 1C_6 + 1C_7 + 1C_8 \\ 1C_3 + 1C_7 + 1C_8 + 1N_9 \\ 1C_8 \end{bmatrix}$$

Atom linear indices of first order is a *linear map*, $f_1(x_i): \mathfrak{R}^n \rightarrow \mathfrak{R}^n$ such that,
 $f_1(O_1, C_2, C_3, C_4, C_5, C_6, C_7, C_8, N_9) = (2C_2, 2O_1 + 1C_3, 1C_2 + 1C_3 + 1C_4 + 1C_8, 1C_3 + 1C_4 + 1C_5, 1C_4 + 1C_5 + 1C_6, 1C_5 + 1C_6 + 1C_7, 1C_6 + 1C_7 + 1C_8, 1C_3 + 1C_7 + 1C_8 + 1N_9, 1C_8) = (5.26, 8.97, 10.52, 7.89, 7.89, 7.89, 7.89, 10.22, 2.63)$ and whole-molecule linear indices of first order is a *linear functional*;

$$f_1(x) = \sum_{i=1}^n f_1(x_i) = f_1(O_1) + f_1(C_2) + f_1(C_3) + f_1(C_4) + f_1(C_5) + f_1(C_6) + f_1(C_7) + f_1(C_8) + f_1(N_9) = 69.16$$

Local and total linear indices of order 0-5 ($k = 0-5$)

Atom (i)	$f_0(x_i)$	$f_1(x_i)$	$f_2(x_i)$	$f_3(x_i)$	$f_4(x_i)$	$f_5(x_i)$
O₁	3.17	5.26	17.94	42.08	146.96	400.72
C₂	2.63	8.97	21.04	73.48	200.36	676.25
C₃	2.63	10.52	37.6	116.2	382.33	1193.57
C₄	2.63	7.89	26.3	87.57	277.41	894.29
C₅	2.63	7.89	23.67	73.64	234.55	739.87
C₆	2.63	7.89	23.67	73.34	227.91	721.81
C₇	2.63	7.89	26	80.93	259.35	820.73
C₈	2.63	10.22	31.26	105.08	333.47	1080.23
N₉	2.33	2.63	10.22	31.26	105.08	333.47
Total	23.91	69.16	217.7	683.58	2167.42	6860.94

It is useful to perform a calculation on a molecule to illustrate the steps in the procedure. For this, we use the 2-aminobenzaldehyde molecule. Table 1 depicts the calculation of the linear indices of the molecular pseudograph's atom adjacency matrix for 2-aminobenzaldehyde. From Table 1, we extract the X -values (Mulliken electronegativity) [24] for each atom and the molecular vector X , for encoding whole-organic molecule, is obtained. Additionally, all valence-bond electrons (σ - and π -networks) in one step are revealed in \mathbf{M}^1 matrix. Then, the local (and total) linear indices of first-order values, $f_1(x_i)$, for each atom are calculated. Nevertheless, the k^{th} ($k = 0-5$) local and total values are shown at the bottom of Table 1.

Total (Whole-Molecule) Linear Indices of the "Molecular Pseudograph's Atom (Vertex) Adjacency Matrix"

Total linear indices are a *linear functions* [30-33] (some mathematicians use the term *linear form*, which means the same as linear functional) on \mathfrak{R}^n . That is, the k^{th} total linear indices is a linear map from \mathfrak{R}^n to the scalar \mathfrak{R} [$f_k(x): \mathfrak{R}^n \rightarrow \mathfrak{R}$]. The mathematical definition of these molecular descriptors is the following:

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

where n is the number of atoms and $f_k(x_i)$ are the atom's linear indices (linear maps) obtained by Eq. 1. Then, a linear form $f_k(x)$ can be written in the matrix form,

$$f_k(x) = [u]^t [X']^k \quad (8)$$

or

$$f_k(x) = [u]^t \mathbf{M}^k [X] \quad (9)$$

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

Local (Atom-type) Linear Indices of the "Molecular Pseudograph's Atom (Vertex) Adjacency Matrix"

In addition to this, atom linear indices computed for each atom in the molecule, a local-fragment (atom-type) formalism can be developed. The k^{th} atom-type linear index of the molecular pseudograph's atom adjacency matrix is calculated by summing the k^{th} atom linear indices of all atoms of the same type in the molecule.

Consequently, if a molecule is partitioned in Z molecular fragments, the total linear indices can be partitioned in Z local linear indices $f_{kL}(x)$, $L = 1, \dots, Z$. That is to say, 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:

$$f_k(x) = \sum_{L=1}^Z f_{kL}(x) \quad (10)$$

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

Methods

Data Set for QSPR/QSAR studies

In order to illustrate the possibilities of the total and local (atom and atom-type) linear indices in the QSPR/QSAR studies, we have selected the following two series to be investigated: 1) boiling point of 28 alkyl-alcohols (see Table 2) firstly studied by Kier and Hall using E-state/biomolecular encounter parameters [13] and recently by Estrada and Molina [11] using the local spectral moments of the edge adjacency matrix, and 2) a set of 34 2-furylethylene derivatives previously studied using total and local spectral moments, 2D/3D connectivity indices (vertex and edge ones) and to quantum chemical descriptors to model their partition coefficient ($\log P$), specific rate constant ($\log k$) and antibacterial activity. These chemicals have different substituents at position 5 of the furan ring as well as at the β position of the exocyclic double bond [38,39]. The structures of these 34 furylethylene derivatives are given in Table 3.

The 2-furylethylene compounds have been well-known as antimicrobials, antitumoral, and cytotoxic during many years [40-42]. The values of the $\log k$ (for nucleophilic addition of the

mercaptoacetic acid) and *n*-octanol/water log *P* of these compounds have been experimentally determined and reported in the literature [38]. Tables 4 and 5 depict these values, respectively. The antibacterial activity of these compounds was determined as the inverse of the concentration *C* that produces 50% of growth inhibition in *E. coli* at six different times and reported as log (1/*C*) [38]. This antibacterial activity was used to classify furylethylenes in two groups by Estrada and Molina [39]. The group of active compounds is composed of those compounds having values of log (1/*C*) < 3, while the rest formed group of inactive compounds. Table 6 illustrates the classification of 2-furylethylene derivatives as antibacterial according to this experimental cutoff value. This Table also depicts the antibacterial activity of a series of nine new 2-furylethylenes using by Estrada and Molina [39] like external prediction (test) set. These compounds have a NO₂ group at position R₃ and a Br or I at positions R₁ and/or R₂. All these compounds were shown to have antibacterial activity in different assays [42,43]. The structures of these compounds are given at bottom of the Table 3.

Computational Methods: TOMOCOMD-CARDD Approach

TOMOCOMD is an interactive program for molecular design and bioinformatics research [14]. It is a composite by four subprograms, each one of them allows one to draw the structures (drawing mode) and to calculate molecular 2D and 3D structures (calculation mode). 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 this paper we outline the salient features of only one of these subprograms: CARDD. This subprogram was developed based on a user-friendly philosophy.

The calculation of total and local linear indices for any organic molecule was implemented in the **TOMOCOMD-CARDD** software [14]. The main steps for the application of this method in QSAR/QSPR can be briefly resumed as follows:

1. Draw the molecular pseudograph for each molecule of the data set, using the software drawing mode. This procedure is performed by selection of the active atom symbol belonging to different groups of the periodic table,
2. Use appropriate weights in order to differentiate the molecular atoms. In this work, we used as atomic property the Mulliken electronegativity [24] for each kind of atom,
3. Compute the total and local linear indices of the molecular pseudograph’s atom adjacency matrix. They can be carried out in the software calculation mode, where you can select the atomic properties and the family descriptor previously to calculate the molecular indices. This software generates a table in which the rows correspond to the compounds and columns correspond to the total and local linear indices or other family molecular descriptors implemented in this program,
4. Find a QSPR/QSAR equation by using mathematical 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 a property **P** and the linear indices having, for instance, the following appearance,

$$P = a_0f_0(x) + a_1f_1(x) + a_2f_2(x) + \dots + a_kf_k(x) + c \quad (11)$$

where P is the measurement of the property, $f_k(x)$ is the k^{th} total linear indices, and the a_k 's are the coefficients obtained by the linear regression analysis.

5. Test the robustness and predictive power of the QSPR/QSAR equation by using internal and external cross-validation techniques,
6. Develop a structural interpretation of obtained QSAR/QSPR model using total and local (atom and atom-type) linear indices as molecular descriptors.

Statistical Analysis

In describing Bp, $\log k$, and $\log P$ the multiple linear regression analysis was used as statistical method. This experiment was performed with STATISTICA software package [44]. The tolerance parameter (proportion of variance that is unique to the respective variable) used was the default value for minimum acceptable tolerance, which is 0.01. Forward stepwise was fixed as the strategy for variable selection. The principle of parsimony (Occam's razor) was taken into account as strategy for model selection. In this connection, we select the model with higher statistical significance but having as few parameters (a_k) as possible. The quality of the models was determined examining the regression's statistic parameters and of the cross-validation procedures [45,46]. In this sense, the quality of models was determined by examining the regression coefficients (R), determination coefficients or squared regression coefficient (R^2), Fisher-ratio's p -level [$p(F)$], standard deviations of the regression (s) and the leave-one-out (LOO) press statistics (q^2, s_{cv}).

On the other hand, linear discriminant analysis (LDA) was used to the classification of 34 2-furylethylene derivatives as antibacterial. This statistical analysis was performed using also STATISTICA software [44]. In order to test the quality of the discriminant function derived, we used the Wilks' λ (U -statistic) and the Mahalanobis distance (D^2). The Wilks' λ statistical helpful to value the total discrimination and can take values between 0 (perfect discrimination) and 1 (no discrimination). The D^2 indicates the separation of the respective groups. The statistical robustness and predictive power of the obtained model was assessed using an external prediction (test) set. In developing classification models the values of 1 and -1 were assigned to active and inactive compounds, respectively. To make the classification of compounds in both groups we preferred the use of the a posteriori probabilities instead of cutoff values. This is the probability that the respective case belongs to a particular group (active or inactive) and it is proportional to the Mahalanobis distance from that group centroid. In closing, the posterior probability is the probability, based on our knowledge of the values of others variables, that the respective case belongs to a particular group. An external test set of nine new compounds was used in order to assess the predictive ability of the obtained LDA model.

Results and Discussion

Describing boiling points of 28 alkyl alcohols

The first data set that will be studied here is composed by 28 alkyl alcohols (14 are primary, 6 secondary and 8 tertiary) for which the boiling point (Bp) has been reported previously [11]. The best

linear regression model obtained to describe the Bp of these compounds using total linear indices is given below:

$$\text{Bp } (^{\circ}\text{C}) = -145.135 (\pm 10.853) + 10.19656 (\pm 0.5680) f_2^{\text{H}}(x) - 3.63852 (\pm 0.21036) f_3^{\text{H}}(x) \quad (12)$$

$$N = 28 \quad R = 0.992 \quad R^2 = 0.984 \quad s = 3.78 \quad q^2 = 0.981 \quad s_{\text{cv}} = 3.91 \quad F(2,25) = 748.57$$

where, $f_2^{\text{H}}(x)$ and $f_3^{\text{H}}(x)$ are the total linear indices (2nd and 3rd order, respectively) calculated considering H-atoms in molecular pseudograph, and N is the number of compounds.

When a local (heteroatom) linear indices ($f_{L4}^{\text{H}}(x_0)$) were included in the statistical analysis an improvement in the description of Bp was obtained:

$$\text{Bp } (^{\circ}\text{C}) = 104.06 (\pm 9.9088) + 0.2464 (\pm 0.009) f_3^{\text{H}}(x) - 0.413 (\pm 0.044) f_4(x) - 0.024 (\pm 0.003) f_7(x) - 0.9526 (\pm 0.118) f_{L4}^{\text{H}}(x_0) \quad (13)$$

$$N = 28 \quad R = 0.997 \quad R^2 = 0.9934 \quad s = 2.48 \quad q^2 = 0.990 \quad s_{\text{cv}} = 2.79 \quad F(4,23) = 871.96$$

The values of experimental and calculated values of the Bp for the data set (both models) are given in Table 2 and the linear relationships between them are illustrated in Figures 1 and 2.

Table 2. Experimental and Predicted Values of the Boiling Point of Alcohols R-OH Used in This Study.

Alcohol-R	Bp ($^{\circ}\text{C}$) ^a	Predicted ^b	Predicted ^c	Predicted ^d	Predicted ^e
		Bp ($^{\circ}\text{C}$)	Bp ($^{\circ}\text{C}$)	Bp ($^{\circ}\text{C}$)	Bp ($^{\circ}\text{C}$)
(CH ₃) ₂ CH-	82.3	85.4	88,2	82.9	91.1
CH ₃ CH ₂ CH ₂ -	97.2	97.5	99,0	96.0	97.4
CH ₃ (CH ₂) ₃ -	117.7	117.1	116,8	115.2	113.6
CH ₃ CH(CH ₃)CH ₂ -	107.8	107.7	106,0	108.0	109.0
CH ₃ CH ₂ C(CH ₃) ₂ -	102.4	103.0	101,0	105.4	112.4
CH ₃ CH ₂ CH ₂ CH(CH ₃)-	119.3	124.6	119,9	114.4	120.3
CH ₃ CH(CH ₃)CH ₂ CH ₂ -	131.1	127.3	129,5	134.5	127.4
CH ₃ CH ₂ CH(CH ₃)CH ₂ -	128.0	127.3	125,6	127.3	125.2
CH ₃ (CH ₂) ₄ -	137.9	136.7	136,3	134.3	131.8
CH ₃ C(CH ₃) ₂ CH(CH ₃)-	120.4	116.0	125,2	129.3	123.0
CH ₃ (CH ₂) ₂ C(CH ₃) ₂ -	121.1	122.6	119,6	124.9	128.9
(CH ₃ CH ₂) ₂ C(CH ₃)-	122.4	122.6	120,2	121.9	126.3
CH ₃ CH ₂ C(CH ₃) ₂ CH ₂ -	136.5	128.1	134,2	142.5	138.4
CH ₃ CH(CH ₃)CH ₂ CH(CH ₃)-	131.6	134.8	133,3	133.9	133.4
CH ₃ CH(CH ₃)CH(CH ₃ CH ₂)-	126.5	134.8	128,0	121.9	128.7
CH ₃ CH(CH ₃)CH(CH ₃)CH ₂ -	144.5	137.5	143,1	146.7	138.3
CH ₃ CH ₂ CH ₂ CH(CH ₃)CH ₂ -	149.0	146.9	145,3	146.4	143.4
CH ₃ (CH ₂) ₅ -	157.6	156.3	155,9	153.4	169.8
(CH ₃ CH(CH ₃)) ₂ CH-	138.7	145.0	141,2	136.4	139.0
CH ₃ CH(CH ₃)CH ₂ CH(CH ₃)CH ₂ -	159.0	157.1	160,7	165.5	157.7
(CH ₃ CH ₂) ₃ C-	142.0	142.2	140,0	138.6	138.5

Table 2. Cont.

Alcohol-R	Bp (°C) ^a	Predicted ^b	Predicted ^c	Predicted ^d	Predicted ^e
		Bp (°C)	Bp (°C)	Bp (°C)	Bp (°C)
CH ₃ (CH ₂) ₆ -	176.4	175.9	175,7	172.5	172.2
(CH ₃ CH ₂ CH ₂) ₂ (CH ₃)C-	161.0	161.8	160,1	160.9	161.3
(CH ₃ (CH ₂) ₃)(CH ₃ CH ₂)(CH ₃)C-	163.0	161.8	162,6	160.5	162.7
CH ₃ CH(CH ₃)CH ₂ (CH ₂) ₄ -	188.0	186.1	190,6	191.6	188.3
CH ₃ (CH ₂) ₇ -	195.1	195.5	195,5	191.6	193.0
CH ₃ (CH ₂) ₅ C(CH ₃) ₂ -	178.0	183.3	180,5	182.2	188.4
(CH ₃ CH ₂ CH ₂) ₂ (CH ₃ CH ₂)C-	182.0	181.4	182,5	177.6	177.0

^aExperimental values of Bp. ^bPredicted values using total linear indices (Eq. 12).

^cPredicted values using total and local linear indices (Eq. 13). ^dPredicted values using spectral moments (Eq. 14). ^ePredicted values using E-state (Eq. 15).

These models (Eqs. 12 and 13) explain more than 98% and 99% of the variance of the experimental Bp values, respectively. Similar equations were reported by Estrada and Molina [11] and Kier and Hall [13] using spectral moment and E-state/ biomolecular encounter parameters as molecular descriptors, respectively. These equations are given bellow with their statistical parameters:

$$\text{Bp (°C)} = 96.555 + 19.093n - 42.428\mu_2(\text{C-O}) + 29.451\mu_3(\text{C-O}) - 7.755\mu_4(\text{C-O}) + 0.0018\mu_8(\text{C-O}) \quad (14)$$

$$N = 28 \quad R^2 = 0.982 \quad s = 4.2 \quad F = 23.8$$

$$\text{Bp (°C)} = 198.02(\pm 25.29) + 0.08(\pm 0.031) (\text{H_H}_2) + 4.602(\pm 0.49) (\text{H_O}) - 22.413(\pm 3.85) \text{S(-OH)} \quad (15)$$

$$N = 28 \quad R^2 = 0.926 \quad s = 5.8 \quad F = 204$$

where, n is a number of carbon atoms in the molecule, $\mu_k(\text{C-O})$ are the k^{th} spectral moment for C-O bond [11] and H_H₂, H_O, S(-OH) are values of biomolecular encounter parameters and E-state, respectively [13].

These models (Eqs. 14 and 15) explain more than 98% and 92% of the variance of the experimental Bp values, respectively.

Predictability and stability of the obtained models using linear indices (Eqs. 12 and 13) to data variation is carried out here by means of LOO cross-validation. These models showed a cross-validation regression coefficient of 0.981 and 0.990 respectively.

Unfortunately, the authors (Estrada and Molina [11], and Kier and Hall [13]) do not report the result of the cross-validation. It is remarkable that one of our models (Eq 12) uses three variables less than the model obtained by Estrada and Molina [11] (Eq 14) and one variable less than the model obtained by Kier and Hall [13] (Eq 15). However, Eq. 12 explains a greater percent of the variance of the experimental Bp values than that the previously developed models do [11, 13].

Figure 1. Correlation between experimental and calculated (by Eq. 12) Boiling point of 28 alcohols of the data set.

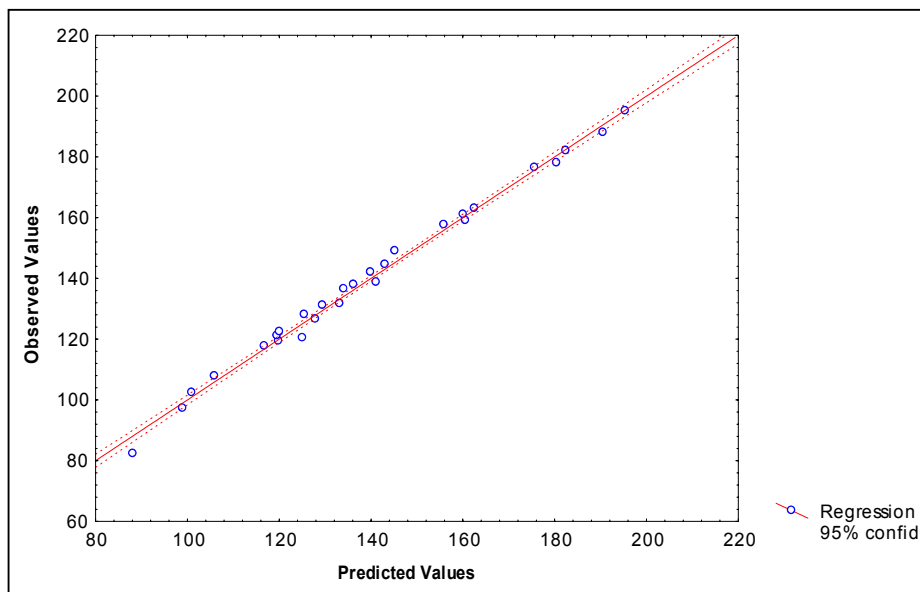
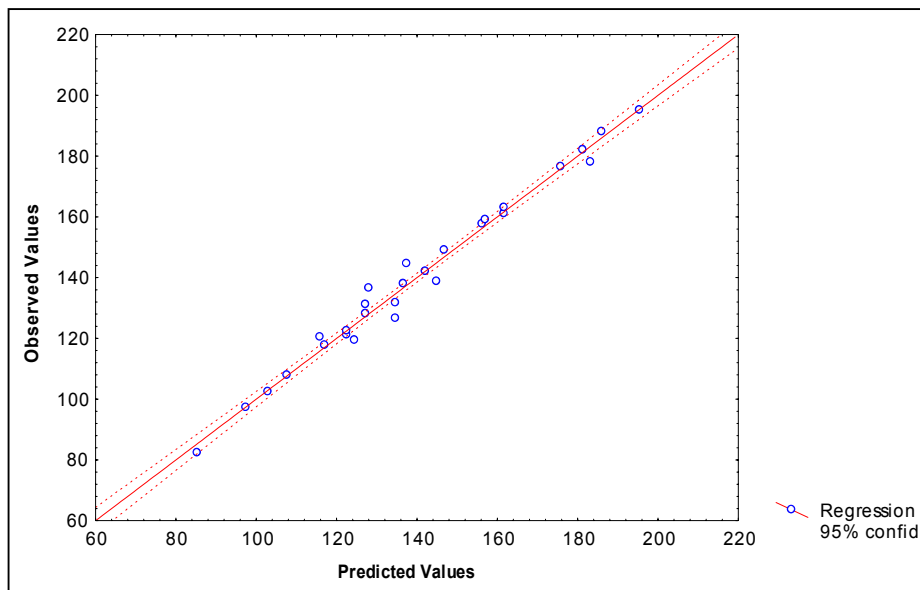


Figure 2. Correlation between experimental and calculated (by Eq. 13) Boiling point of 28 alcohols of the data set.



Modeling specific rate constants (log k) of 34 2-furylethylenes derivatives

Many topological descriptors are not useful to describe chemical reactions [11]. In order to prove the applicability of this new approach in the **QSR(Reactivity)R** studies, we select a data set of 34 derivatives of 2-furylethylene. The molecular structures of such compounds are depicted in Table 3.

Table 3. Chemical structures and numbering of atoms in the 2-furylethylene compounds used in this study.

no.	R ₁	R ₂	R ₃	no.	R ₁	R ₂	R ₃
1	H	NO ₂	COOCH ₃	18	NO ₂	H	CONHCH(CH ₃)C ₂ H ₅
2	CH ₃	NO ₂	COOCH ₃	19	NO ₂	H	CONHC(CH ₃) ₃
3	Br	NO ₂	COOCH ₃	20	NO ₂	H	CONHCH ₂ C(CH ₃) ₃
4	I	NO ₂	COOCH ₃	21	NO ₂	H	COOCH ₃
5	COOCH ₃	NO ₂	COOCH ₃	22	NO ₂	H	COOC ₂ H ₅
6	NO ₂	NO ₂	COOCH ₃	23	NO ₂	H	COO(CH ₂) ₂ CH ₃
7	NO ₂	COOC ₂ H ₅	COOC ₂ H ₅	24	NO ₂	H	COOCH(CH ₃) ₂
8	NO ₂	H	NO ₂	25	NO ₂	H	COO(CH ₂) ₃ CH ₃
9	H	H	NO ₂	26	NO ₂	H	COOCH ₂ CH(CH ₃) ₂
10	NO ₂	H	CONH ₂	27	NO ₂	H	COOCH(CH ₃)C ₂ H ₅
11	NO ₂	H	CONHCH ₃	28	NO ₂	H	COOC(CH ₃) ₃
12	NO ₂	H	CON(CH ₃) ₂	29	NO ₂	H	COO(CH ₂) ₄ CH ₃
13	NO ₂	H	CONHC ₂ H ₅	30	NO ₂	H	Br
14	NO ₂	H	CONH(CH ₂) ₂ CH ₃	31	NO ₂	H	CN
15	NO ₂	H	CONHCH(CH ₃) ₂	32	NO ₂	H	OCH ₃
16	NO ₂	H	CONH(CH ₂) ₃ CH ₃	33	NO ₂	H	H
17	NO ₂	H	CONHCH ₂ CH(CH ₃) ₂	34	NO ₂	CN	COOCH ₃

Novel R₁,R₂-Substituted 2-Furylethylenes (R₃ = NO₂) used as *external test set* to assess the predictive power of the classification model for antibacterial activity

1	Br	Br	NO ₂	6	H	I	NO ₂
2	I	I	NO ₂	7	H	CH ₃	NO ₂
3	Br	H	NO ₂	8	Br	CH ₃	NO ₂
4	H	Br	NO ₂	9	I	CH ₃	NO ₂
5	I	H	NO ₂				

These compounds were studied by Estrada y Molina [11] to describe the specific rate constant k of nucleophilic addition of the mercaptoacetic acid using their total and local spectral moments, connectivity indices and quantum chemical local descriptors. All developed models had seven variables.

Table 4. Experimental and calculated values of the specific rate constant for the reaction of nucleophilic addition of thiols ($\log k$) to the exocyclic double bond of the studied 2-furylethylenes

Number	Experimental	TIs	Total moments	QC	Local moments	Linear indices
1	6.591	7.532	6.281	6.780	6.679	6.441
2	6.518	6.983	5.583	6.646	6.563	6.540
3	6.914	6.352	6.893	7.021	6.979	7.075
4	6.982	6.997	6.554	6.935	7.069	7.015
5	7.176	6.750	7.390	7.137	7.291	7.188
6	7.602	6.685	7.410	7.466	7.572	7.589
7	5.255	5.533	5.002	5.212	4.934	5.063
8	6.763	4.965	5.962	6.737	6.662	6.353
9	5.623	4.997	4.839	5.454	5.599	5.292
10	3.813	4.831	4.455	3.695	4.318	4.468
11	3.840	3.636	4.876	3.616	4.034	4.005
12	3.874	4.162	2.911	3.537	3.448	3.610
13	3.825	3.916	4.255	3.562	3.875	3.722
14	3.623	3.798	4.032	3.550	3.841	3.698
15	3.751	3.528	3.255	3.522	3.721	3.461
16	3.784	3.114	3.437	3.566	3.842	3.693
17	3.697	4.175	3.468	3.570	3.826	3.684
18	3.705	3.934	3.157	3.541	3.711	3.443
19	3.697	3.264	3.686	3.482	3.798	3.227
20	3.650	3.526	4.269	3.581	3.851	3.683
21	4.000	3.693	5.184	3.985	3.735	4.259
22	3.920	3.974	4.202	3.920	3.617	3.993
23	3.790	3.856	4.409	3.926	3.596	3.970
24	3.763	3.869	3.438	3.849	3.552	3.749
25	3.623	3.172	3.893	3.921	3.600	3.965
26	3.650	4.233	4.178	3.917	3.610	3.957
27	3.592	4.193	3.613	3.873	3.566	3.732
28	3.584	4.059	3.596	3.819	3.881	3.532
29	3.590	3.478	3.204	3.918	3.600	3.968
30	2.987	3.463	3.534	3.281	2.785	2.997
31	3.273	3.070	3.779	4.089	3.651	3.370
32	2.140	3.070	2.794	2.287	2.751	1.941
33	3.553	3.780	2.716	3.338	3.485	3.558
34	5.557	5.114	5.474	4.969	5.524	5.465

The model obtained by these authors using the connectivity indices describes an 82% of the experimental values of $\log k$, with a standard deviation of 0.681. In addition, these researchers obtained similar results using the global spectral moments as molecular descriptors in QSRR equation ($R^2 = 84\%$ and $s = 0.655$) [11]. The use of local molecular descriptors such as quantum chemical or graph-theoretical (local spectral moments) produces a significant improvement in the statistical quality of the obtained models. In this sense, both models (quantum chemical and local spectral moments) explain more than 96% (96.8% and 96.4%) of the variance of the $\log k$, with a standard deviation of 0.288 and 0.320, respectively.

The molecular descriptors included in these equations clearly pointed to the identification of the reaction centers involved in the studied chemical interaction. That is to say, the molecular indices calculated for the atoms 2, 6 and 7 or for the bonds defined by these atoms (C₂-C₆ and C₆-C₇) were included in the obtained models. These atoms are those involved in the exocyclic double bond of the 2-furylethylene and these are the “target” of the nucleophilic attack by thiol (mercapto) group.

Taking into account this logical result, we calculated the k^{th} local linear indices for the atoms C₂, C₆ and C₇ (bonds C₂-C₆ and C₆-C₇). The best obtained model, using these local linear indices as molecular descriptors, together with its statistical parameters is given below:

$$\begin{aligned} \log k = & 231.464(\pm 14.586) - 0.00183(\pm 0.0008)f_6^{\text{H}}(x_{\text{C2-C6}}) - 4.6054(\pm 0.3103)f_2^{\text{H}}(x_{\text{C2-C6}}) \\ & - 0.00107(\pm 8.92 \times 10^{-5})f_8^{\text{H}}(x_{\text{C2-C6}}) + 0.225923(\pm 0.02621)f_4^{\text{H}}(x_{\text{C2-C6}}) \\ & + 3.85 \times 10^{-6}(\pm 4.53 \times 10^{-7})f_{12}(x_{\text{C2-C6}}) + 0.05(\pm 0.0076)f_2^{\text{H}}(x_{\text{C6-C7}}) \end{aligned} \quad (16)$$

N = 34 R = 0.986 R² = 0.973 s = 0.26 q² = 0.948 s_{cv} = 0.33 F(6,27) = 161.22

Note, that our model (Eq. 16) included only six variables (one less than the models object of comparison) and explained more than 97% of the variance ($s = 0.26$). These statistics are slightly better than those obtained previously.

Figure 3. Observed versus predicted log k of the specific rate constant for the reaction of nucleophilic addition of thiols to the exocyclic double bond of the 2-furylethylene derivatives.

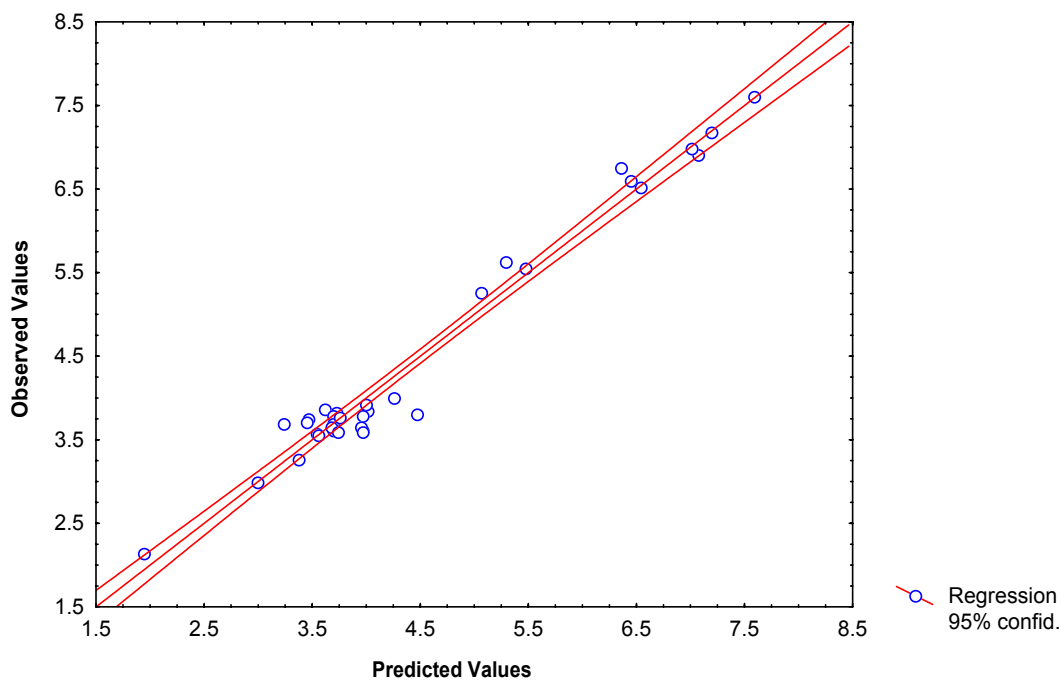


Table 4 depicts the experimental and calculated values of reactivity index (log k) from connectivity indices, total and local spectral moments, quantum chemical indices and local linear indices. Plots of observed versus calculated log k for data set of compounds are illustrated in Figure 3.

Modeling partition coefficients (log P) of 34 2-furylethylenes derivatives

It has been clear from structure-activity relationship studies that the lipophilicity of 2-furylethylenes derivatives is critical for the development of their antibacterial activity [38]. The partition coefficient *n*-octanol/water (log *P*) has an important role in the understanding of the biological behavior of these 2-furylethylene derivatives [38]. Consequently, we will study this parameter to compare the possibilities of molecular linear indices in QSPR and to compare this result to those obtained by Estrada and Molina [39] using 2D and 3D (topographic) connectivity indices (vertex and edge ones), and quantum chemical descriptors.

The best obtained model, using total and local linear indices as molecular descriptors, together with its statistical parameters is given below:

$$\begin{aligned} \log P = & -3.184(\pm 0.488) + 0.067(\pm 0.002) f_{0(x)}^H + 0.77659(\pm 0.058) f_{L0(x_E)}^H \\ & - 0.1576(\pm 0.012) f_{L2(x_E)}^H + 0.00915(\pm 0.0022) f_{L3(x_E)}^H - 0.069(\pm 0.007) f_{L2(x_{E-H})}^H \\ & + 0.0026(\pm 0.0007) f_{L3(x_{C6-C7})}^H + 3.511 \times 10^{-6} (\pm 6.23 \times 10^{-6}) f_{L3(x_{C6-C7})}^H \quad (17) \end{aligned}$$

$N = 34 \quad R = 0.984 \quad R^2 = 0.968 \quad s = 0.143 \quad q^2 = 0.938 \quad s_{cv} = 0.176 \quad F(7,26) = 113.38$

This equation explained 96.8% of the variance of both log *P*. This statistic is lightly better than those obtained previously [39]. The experimental and calculated values of log *P* obtained with 2D and 3D connectivity indices, quantum chemical descriptors, total and local spectral moments as well as molecular linear indices are show in Table 5. Plots of observed versus calculated log *P* according to the Eq. 17 are illustrated in Figure 4.

Finally, LOO cross-validation procedure was used in order to assess the predictive ability of developed model (17). Using this approach, the model 17 had a LOO q^2 of 0.938. This value of q^2 ($q^2 > 0.5$) can be considered as a proof of the high predictive ability of the models [44-47]. In this sense, the equations obtained with the vertex and edge connectivity indices, with the topographic descriptors, and with the quantum chemical descriptors (Eqs. 10, 11, and 13 in Ref. 39) showed a smaller predictive abilities (s_{cv} of 0.247, 0.176, and 0.370, respectively) that the equation 17 ($s_{cv} = 0.176$), achieved with the total and local linear indices.

Classification of 34 2-furylethylene derivatives as antibacterial

Linear discriminant analysis (LDA) will be used here to obtain a classification model of 2-furylethylene compounds according to their antibacterial activity. The classification model obtained is given below together with the statistical parameters of LDA:

$$\text{Class} = -107.731 + 0.5155 f_{L3(x_{C2-C6})}^H - 0.120 f_{1(x)}^H + 0.4099 f_{L2(x_{E-H})}^H \quad (18)$$

$N = 34 \quad \lambda = 0.304 \quad D^2 = 9.44 \quad F(3,30) = 22.862 \quad p < 0.0000$

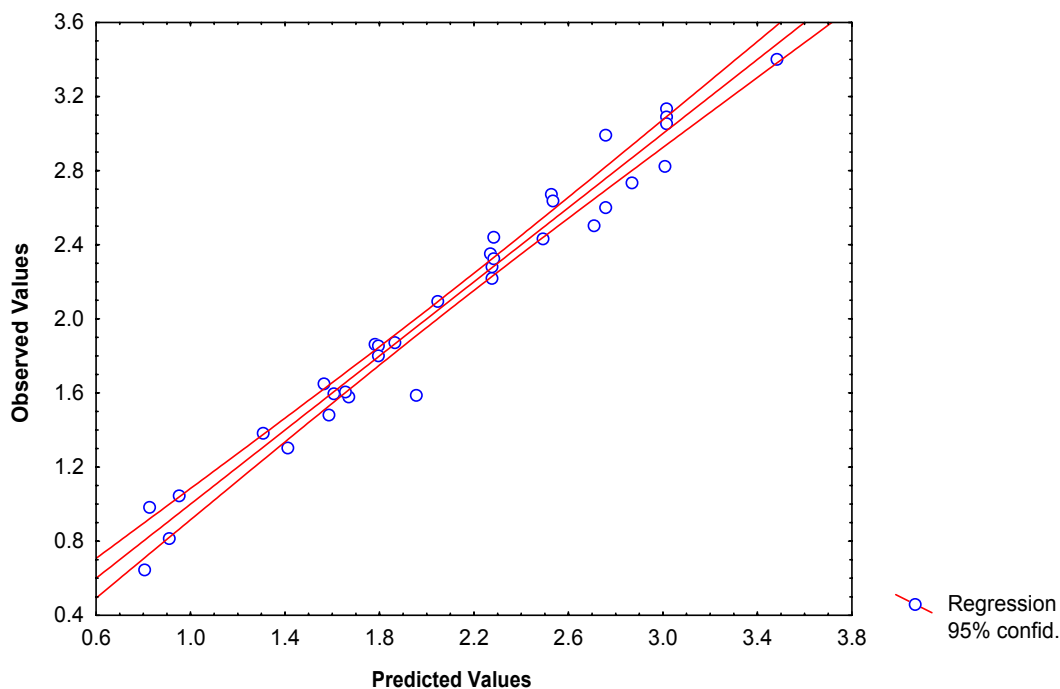
where, λ is Wilk's statistic, D^2 is the squares of Mahalanobis distances, and F is the Fisher ratio. The statistical analysis showed that exist appropriate discriminatory power for differentiating between the two respective groups. The calculation of percentages of good classification in the data set and external prediction set permitted us to carry out the assessment of the models.

Table 5. Experimental and calculated values of the partition coefficient *n*-octanol/water ($\log P$) for the furylethylenes studied.

no.	Obsd. ^a	Pred. ^b	Res. ^c	Res-CV ^d	topol. ^e	topog. ^f	QC ^g
1	1.879	1.86	0.02	0.02	1.894	1.955	1.836
2	2.439	2.49	-0.05	-0.07	2.482	2.398	2.239
3	2.739	2.86	-0.13	-0.20	2.753	2.748	2.405
4	2.999	2.75	0.25	0.39	2.905	2.898	2.510
5	1.869	1.78	0.09	0.13	1.763	1.930	1.976
6	1.599	1.60	0.00	-0.01	1.619	1.550	1.679
7	2.504	2.71	-0.20	-0.27	2.703	2.640	2.706
8	1.303	1.41	-0.11	-0.19	1.191	1.338	1.456
9	1.583	1.67	-0.09	-0.12	1.453	1.783	1.583
10	0.649	0.80	-0.15	-0.32	0.433	0.300	0.180
11	0.984	0.82	0.16	0.20	0.999	1.091	1.076
12	0.819	0.91	-0.09	-0.17	1.160	0.870	2.149
13	1.386	1.31	0.08	0.09	1.583	1.423	1.482
14	1.860	1.79	0.07	0.08	2.311	1.941	1.858
15	1.803	1.79	0.01	0.01	1.966	2.084	1.906
16	2.356	2.26	0.09	0.10	2.168	2.332	2.240
17	2.225	2.27	-0.05	-0.05	2.493	2.526	2.241
18	2.284	2.27	0.01	0.01	2.384	2.383	2.277
19	2.333	2.28	0.06	0.07	2.316	2.316	2.346
20	2.605	2.75	-0.15	-0.18	2.382	2.575	2.618
21	1.652	1.56	0.09	0.10	1.347	1.585	1.830
22	2.098	2.05	0.05	0.06	1.984	1.947	2.126
23	2.673	2.53	0.15	0.16	2.733	2.459	2.504
24	2.641	2.53	0.11	0.12	2.484	2.666	2.592
25	2.827	3.00	-0.18	-0.19	2.726	2.837	2.902
26	3.135	3.01	0.13	0.14	3.052	3.034	2.902
27	3.091	3.01	0.08	0.09	3.018	2.952	2.943
28	3.060	3.01	0.05	0.05	2.994	3.002	3.029
29	3.404	3.48	-0.07	-0.09	3.227	3.252	3.266
30	2.447	2.28	0.17	0.24	2.510	2.469	2.132
31	1.050	0.95	0.10	0.24	1.365	1.258	1.344
32	1.591	1.95	-0.36	-0.41	1.510	1.500	1.711
33	1.611	1.65	-0.04	-0.07	1.738	1.515	1.590
34	1.488	1.59	-0.10	-0.27	1.309	1.424	1.504

^aExperimental values taken of the Ref. [39]; ^bPredicted values using total and local (atom and atom-type linear indices (Eq. 17); ^cResidual values: $\log P(\text{Obsd}) - \log P(\text{Pred})$; ^dResidual values of the LOO cross-validation experiment (Deleted Residual); ^ePredicted values using topological indices (vertex and edge connectivity indices) [39]. ^fPredicted values using topographic descriptors [39]. ^gPredicted values using quantum chemical molecular descriptors [39].

Figure 4. Linear correlations of observed versus calculated log *P* according to the model obtained from molecular linear indices.



Model 18 classified correctly 94.12% of the compounds in the training data set (92.85% and 95.0% of good classification in active and inactive training data set, respectively), misclassifying only 2 compounds of a total of 34. The percentage of false actives as well as of the false inactive in this data set was only 2.94%.

The statistical analysis of three models obtained previously using 2D and 3D connectivity and quantum chemical descriptors showed quite similar results. In this case, the overall accuracy of the three models was 91.2%, 94.1%, and 88.2%, respectively [39].

The classification of all compounds in the complete training data set provides some assessment of the goodness of fit of the model, but it does not provide a thorough criterion of how the model can predict the biological properties of new compounds. To assess such predictive power, the use of an external test set is essential [45-47]. In this sense, the activity of the compounds in such set was predicted with the obtained discrimination function.

The overall accuracy for this group was 100.0%. Using this same external test set of nine new 2-furylethylenes, the QSAR models obtained with 2D and 3D connectivity and quantum chemical descriptors have also 100.0% of global good classification, including one NC (not-classified) compound [39]. The results of global classification of compounds in both, training and external prediction sets archived with all these approaches are shown in Table 6 (see also Table 7).

Table 6. Classification of 2-furylethylene derivatives as antibacterial according to the four models obtained with molecular linear indices, 2D and 3D connectivity as well as quantum chemical descriptors.

Compd.	Obsd.[39]	Linear Indices		2D Conn. [39]		3D Conn. [39]		Quantum [39]	
		Class.	Prob.	Class.	Prob.	Class.	Prob.	Class.	Prob.
1	+	+	99.99	+	95.43	+	99.49	+	99.72
2	+	+	99.97	+	91.67	+	95.83	+	99.86
3	+	+	100.00	+	84.95	+	96.22	+	98.31
4	+	+	100.00	+	79.65	+	95.78	+	97.67
5	+	+	99.58	+	99.72	+	99.63	+	99.66
6	+	+	99.96	+	99.85	+	99.98	+	98.91
7	+	+	88.62	+	94.29	+	91.77	+	98.57
8	+	+	90.32	+	74.81	+	57.81	+	92.50
9	+	+	98.47	-	9.86	-	8.11	+	77.08
10	+	+	99.82	+	99.13	+	99.28	-	32.29
11	+	+	85.96	+	88.24	+	57.28	-	9.46
12	+	-	4.58	+	66.00	+	86.94	-	4.26
13	+	+	50.52	+	57.89	+	71.56	-	12.79
14	-	-	14.55	-	6.25	-	46.43	-	12.96
15	-	-	14.55	-	28.14	-	36.15	-	8.72
16	-	-	2.76	-	0.92	-	1.10	-	11.78
17	-	-	2.76	-	2.35	-	6.19	-	11.05
18	-	-	2.76	-	37.62	-	4.56	-	9.96
19	-	-	2.76	-	8.96	-	2.97	-	9.96
20	-	-	0.47	-	1.14	-	0.08	-	9.75
21	-	-	49.07	+	55.73	+	88.95	-	8.59
22	-	-	13.84	-	22.77	-	18.96	-	7.60
23	-	-	2.61	-	1.36	-	8.23	-	7.66
24	-	-	2.61	-	7.01	-	4.83	-	6.44
25	-	-	0.44	-	0.19	-	0.09	-	8.56
26	-	-	0.44	-	0.33	-	0.58	-	8.49
27	-	-	0.44	-	7.59	-	0.43	-	7.94
28	-	-	0.44	-	1.41	-	0.27	-	7.02
29	-	-	0.07	-	0.02	-	0.04	-	7.21
30	-	-	0.55	-	4.65	-	7.56	-	0.32
31	-	+	52.57	-	29.58	-	37.61	-	3.04
32	-	-	1.05	-	23.67	-	14.96	-	5.30
33	-	-	0.33	+	58.87	-	14.08	-	0.42
34	+	+	99.69	+	97.13	+	97.50	+	62.36
Test set									
1	+	+	99.26	+	88.53	+	95.81	+	87.18
2	+	+	99.04	+	86.87	+	94.59	+	85.53
3	+	+	98.77	+	59.01	+	65.00	+	54.95
4	+	+	99.07	+	96.35	+	99.59	+	96.12
5	+	+	98.67	NC	50.07	+	53.00	NC	50.01
6	+	+	98.89	+	96.72	+	99.51	+	97.68
7	+	+	99.99	+	95.92	+	99.62	+	94.76
8	+	+	99.99	+	86.38	+	96.06	+	84.27
9	+	+	99.99	+	81.75	+	95.74	+	79.29

Table 7. Statistical parameters of the QSPR/QSAR models obtained using different molecular descriptors.

index	n	R ²	s	q ²	s _{cv}	F
Boiling Point of 28 Alkyl-Alcohols						
Linear indices (Eq. 12)	2	0.984	3.78	0.981	3.91	748.57
Linear indices (Eq. 13)	4	0.993	2.48	0.99	2.79	871.96
Local spectral moments [11]	5	0.982	4.2	*	*	23.8
E-State/encounter parameters [13]	3	0.926	5.8	*	*	204
Reactivity (log <i>k</i>) of 34 2-Furylethylenes						
Linear indices	6	0.973	0.260	0.948	0.33	161.2
Conn. Indices [11]	7	0.821	0.681	*	*	17.1
Global spectral moments [11]	7	0.843	0.655	*	*	18.8
Local spectral moments [11]	7	0.964	0.320	*	*	70.4
Quantum chemical descriptors [11]	7	0.968	0.288	*	*	112.2
Partition Coefficient <i>n</i> -Octanol/Water (log <i>P</i>) of 34 2-Furylethylenes						
Linear indices	7	0.968	0.143	0.938	0.176	113.38
Vertex and edge conn. Indices [39]	7	0.939	0.199	*	0.247	56.9
Topographic descriptors [39]	7	0.964	0.155	*	0.176	84.6
quantum chemical descriptors [39]	used the Rogers and Cammarata approach	0.875	0.319	*	0.370	45.5
index	n	λ	D ²	Accuracy (Training)	Accuracy (Test)	F
Classification of 34 2-Furylethylene Derivatives as Antibacterial						
Linear indices	3	0.30	9.44	94.12%	100%	22.9
Vertex and edge conn. Indices [39]	5	0.43	5.7	91.2%	100%	7.7
Topographic descriptors [39]	5	0.38	6.7	94.1%	100%	9.1
quantum chemical descriptors [39]	5	0.44	5.2	88.2%	100%	7.1

*Values are not reported in the literature

Finally, the improvement in the statistical parameters of our model (Eq. 18) compared to that using 2D and 3D connectivity indices as well as quantum chemical descriptors is easily detected by the decrease in the Wilk's λ parameter and an increase in the Mahalanobis square distance (see Table 7).

Concluding Remarks

Although there have been many discoveries in the last years in the field of theoretical drug-design it is necessary to continue developing new molecular descriptors that can explain, by means of QSAR studies, different pharmacological properties of these substances. In this sense, the definition of molecular descriptors based on graph-theoretical invariants that contain important information on atoms (or bonds) in an explicit way is not only possible but also necessary [11]. In this context, total

and local (atom and atom-type) linear indices of the molecular pseudograph's atom adjacency matrix are promising total and local-level molecular descriptors.

We have shown here that total and local linear indices are useful molecular descriptors for modeling physicochemical and biological properties of organic compounds. The obtained models were statistically significant and better than other obtained previously using recognized methods (see Table 7). Taking into consideration those total and local spectral moments, connectivity indices, quantum chemical descriptors and E-state, which have been successfully applied in the QSAR/QSPR studies and drug design, the satisfactory comparative result showed that linear indices used here will be a novel chem-bioinformatic tool for the computer aided "rational" drug design (*TOMOCOMD-CARDD*).

Acknowledgements

Authors would like to offer our sincere thanks to the two anonymous referees for their critical opinions about the manuscript, which have significantly contributed to improving its presentation and quality. F. T. acknowledges financial support from the Spanish MCT (Plan Nacional I+D+I, Project No. BQU2001-2935-C02-01) and Generalitat Valenciana (DGEUI INF01-051, INFRA03-047 and OCYT GRUPOS03-173). Dr. Ivis Fernandez, Cuba, is acknowledged for language correction.

References

1. Julián-Ortiz, J. V. Virtual Darwinian Drug Design: QSAR Inverse Problem, Virtual Combinatorial Chemistry, and Computational Screening. *Comb. Chem. High Throughput Screen.* **2001**, *4*, 295-310
2. Dean, P.M.; Lewis, R.A. (Eds.), *Molecular Diversity in Drug Design*, Kluwer Academic Publishers: Amsterdam, **1999**
3. Walters, W.P.; Stahl, M.T.; Murcko, M.A. Virtual screening: An overview. *Drug Discovery Today* **1998**, *3*, 160-178
4. Karelson, M. *Molecular Descriptors in QSAR/ QSPR*; John Wiley & Sons: New York, **2000**
5. Katritzky, A. R.; Gordeeva, E. V. Traditional Topological Indexes vs Electronic, Geometrical, and Combined Molecular Descriptors in QSAR/QSPR Research. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 835
6. Kier, L. B.; Hall, L. H. *Molecular Connectivity in Chemistry and Drug Research*; Academic Press: London, **1976**
7. Balaban, A. Topological and Stereochemical Molecular Descriptors for Databases Useful in QSAR, Similarity/Dissimilarity and Drug Design. *SAR QSAR Environ. Res.* **1998**, *8*, 1-21
8. Estrada, E. On the Topological Sub-Structural Molecular Design (TOSS-MODE) in QSPR/QSAR and Drug Design Research. *SAR QSAR Environ. Res.* **2000**, *11*, 55-73
9. Julian-Ortiz, J. V.; Gálvez, J.; Muños-Collado, C.; García- Domenech, R.; Gimeno-Cardona, C. Virtual Combinatorial Syntheses and Computational Screening of New Potential Anti-Herpes Compounds. *J. Med. Chem.* **1999**, *42*, 3308-3314
10. Estrada, E.; Uriarte, E. Recent Advances on the Role of Topological Indices in Drug Discovery Research. *Curr. Med. Chem.* **2001**, *8*, 1573-1588

11. Estrada, E.; Molina, E. Novel Local (Fragment-Based) Topological Molecular Descriptors for QSPR/QSAR and Molecular Design. *J. Mol. Graphics Mod.* **2001**, *20*, 54-64
12. Kier, L.B.; Hall, L.H. The electrotopological state: structure modeling for QSAR and data base analysis, In *Topological Indices and Related Descriptors in QSAR and QSPR*; Devillers, J.; Balaban, A.T. (eds.); Gordon and Breach: Amsterdam, 1999; pp. 491-562
13. Kier, L.B.; Hall, L.H. *Molecular Structure Description. The Electrotopological State*; Academic Press: New York, **1999**
14. Marrero-Ponce, Y.; Romero, V. **TOMOCOMD** software. Central University of Las Villas. **2002**. **TOMOCOMD** (**T**opological **M**olecular **C**OMputer **D**esign) for Windows, version 1.0 is a preliminary experimental version; in the future a professional version will be available upon request from Y. Marrero: yovanimp@qf.uclv.edu.cu; ymarrero77@yahoo.es
15. Marrero-Ponce, Y. Total and Local Quadratic Indices of the Molecular Pseudograph's Atom Adjacency Matrix: Applications to the Prediction of Physical Properties of Organic Compounds. *Molecules* **2003**, *8*, 687-726, <http://www.mdpi.org>
16. 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.* DOI: 10.1021/ci049950k
17. Marrero-Ponce, Y.; Cabrera, M., A.; Romero, V.; Ofori, E.; Montero, L. A. Total and Local Quadratic Indices of the "Molecular Pseudograph's Atom Adjacency Matrix". Application to Prediction of Caco-2 Permeability of Drugs. *Int. J. Mol. Sci.* **2003**, *4*, 512-536
18. Marrero-Ponce, Y.; Cabrera, M. A.; Romero, V.; González, D. H.; Torrens, F. A New Topological Descriptors Based Model for Predicting Intestinal Epithelial Transport of Drugs in Caco-2 Cell Culture. *J. Pharm. & Pharm. Sci.* **2004**, *7*, 186-199
19. 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. *CPS: physchem/0401004*
20. Marrero-Ponce, Y.; Castillo-Garit, J.A.; Olazabal, E.; Serrano, H. S.; Morales, A.; Castañedo, N.; Ibarra-Velarde, F.; Huesca-Guillen, A.; Jorge, E.; del Valle, A.; Torrens, F.; Castro, E.A. **TOMOCOMD-CARDD**, a Novel Approach for Computer-Aided "Rational" Drug Design: I. Theoretical and Experimental Assessment of a Promising Method for Computational Screening and *in silico* Design of New Anthelmintic Compounds. *J. Comput. Aided Mol. Des.* Submitted for publication
21. Marrero-Ponce, Y.; Iyarreta-Veitía, M.; Montero-Torres, A.; Romero-Zaldivar, C.; Brandt, C. A.; Ávila, P. E.; Kirchgatter, K. A Novel Approach for Computer-Aided "Rational" Drug Design: Theoretical and Experimental Assessment of a Promising Method for Virtual Screening and *in silico* Design of New Antimalarial Compounds. *J. Med. Chem.* Submitted for publication
22. Marrero-Ponce, Y.; González-Díaz, H.; Romero-Zaldivar, V.; Torrens, F.; Castro, E. A. 3D-Chiral Quadratic Indices of the "Molecular Pseudograph's Atom Adjacency Matrix" and their Application to Central Chirality Codification: Classification of ACE Inhibitors and Prediction of σ -Receptor Antagonist Activities. *Bioorg. Med. Chem.* **2004**, *12*, 5331-5342

23. 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.* DOI: 10.1016/j.bmc.2004.09.034
24. Cotton, F. A. *Advanced Inorganic Chemistry*; Ed. Revolucionaria: Havana (Cuba), **1970**; p 103
25. Wang, R.; Gao, Y.; Lai, L. Calculating Partition Coefficient by Atom-Additive Method. *Perspect. Drug Discov. Des.* **2000**, *19*, 47-66
26. Ertl, P.; Rohde, B.; Selzer, P. Fast Calculation of Molecular Polar Surface Area as a Sum of Fragment-Based Contributions and Its Application to the Prediction of Drug Transport Properties. *J. Med. Chem.* **2000**, *43*, 3714-3717
27. Ghose, A. K.; Crippen, G. M. Atomic Physicochemical Parameters for Three-Dimensional-Structure-Directed Quantitative Structure-Activity Relationships. 2. Modeling Dispersive and Hydrophobic Interactions. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 21-35
28. Millar, K. J. Additivity Methods in Molecular Polarizability. *J. Am. Chem. Soc.* **1990**, *112*, 8533-8542
29. Gasteiger, J.; Marsilli, M. A New Model for Calculating Atomic Charge in Molecules. *Tetrahedron Lett.* **1978**, *34*, 3181-3184
30. Browder, A. *Mathematical Analysis. An Introduction*; Springer-Verlag: New York, **1996**; pp. 176-296.
31. Axler, S. *Linear Algebra Done Right*; Springer-Verlag: New York, **1996**; pp 37-70
32. Ross, K. A.; Wright, C. R. B. *Matemáticas discretas*; Prentice Hall Hispanoamericana: Mexico D.F., **1990**
33. Maltsev, A. I. *Fundamentos del Álgebra Lineal*; Mir: Moscow, **1976**; p 68-262
34. Harary, F. *Graph Theory*; Addison-Wesley, Reading, MA, **1969**; p. 10
35. G. Chartrand, *Graph as Mathematical Models*; Prindle, Weber & Schmidt: Boston, MA, **1977**; p. 22
36. R.J. Wilson, *Introduction to Graph Theory*; Oliver & Boyd: Edinburgh, **1972**; p. 10
37. N. Trinajstić, *Chemical Graph Theory*; CRC Press: Boca Raton, FL, **1992**; 2nd edition, pp. 6-7
38. Balaz, S.; Sturdik, E.; Rosenberg, M.; Augustin, J.; Skara, B. Kinetics of Drug Activities as Influenced by Their Physicochemical Properties: Antibacterial Effects of Alkylating 2-furylethylenes. *J. Theor. Biol.* **1988**, *131*, 115-134
39. Estrada, E.; Molina, E. 3D Conectivity Indices in QSPR/QSAR Studies. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 791-797
40. Dore, J. Ch.; Viel, C. Antitumoral Chemoterapy. X. Cytotoxic and Antitumoral Activity of β -Nitrostyrenes and Nitrovinyl Derivatives. *Farmaco.* **1975**, *30*, 81-109
41. Sturdik, E.; Drobica, L.; Balaz, S. Reaction of 2-Furylethylenes with Thiols in Vivo. *Coll. Czech. Chem. Comm.* **1985**, *50*, 470-480
42. Blondeau, J. M.; Castañedo, N.; Gonzalez, O.; Medina, R.; Silveira, E. In Vitro Evaluation of G-1: A Novel Antimicrobial Compound. *Antimicrob. Agents Chemother.* **1999**, *11*, 1663-1669
43. Castañedo, N.; Goizueta, R.; Perez, J.; Gonzalez, J.; Silveira, E. Cuesta, M.; Martinez, A.; Lugo, E.; Estrada, E.; Carta, A.; Navia, O.; Delgado, M. *Cuban Pat. 22446*, **1994**; *Can. Pat. 2,147,594*, **1999**.
44. STATISTICA ver. 5.5, Statsoft, Inc. **1999**.

45. Wold, S.; Erikson, L. Statistical Validation of QSAR Results. Validation Tools; In *Chemometric Methods in Molecular Design*, van de Waterbeemd, H., Ed.; VCH Publishers: New York, **1995**; pp. 309-318
46. Belsey, D. A.; Kuh, E.; Welsch, R. E. *Regression Diagnostics*; Wiley: New York, **1980**
47. Cronin, M. T. D.; Schultz, T. W. Pitfalls in QSAR. *J. Mol. Struct. (Theochem)*. **2003**, 622, 39-51

© 2004 by MDPI (<http://www.mdpi.org>). Reproduction is permitted for noncommercial purposes.