J. Mex. Chem. Soc. **2013**, *57(1),* 61-68
© 2013, Sociedad Química de México
ISSN 1870-249X

QSPR/QSAR Studies of 2-Furylethylenes Using Bond-Level Quadratic Indices and Comparison with Other Computational Approaches

Eugenio R. Martinez Albelo,[1,2] Yovani Marrero Ponce,[2-4] Stephen J. Barigye,[2] Yunaimy Echeverría-Díaz,[2] and Facundo Pérez-Giménez[4]

[1] Faculty of Chemistry-Pharmacy. Central University of Las Villas, Santa Clara, 54830, Villa Clara, Cuba. 53-42-281164. yovanimp@uclv.edu.cu

[2] Unit of Computer-Aided Molecular 'Biosilico' Discovery and Bioinformatic Research (CAMD-BIR Unit), Faculty of Chemistry-Pharmacy. Central University of Las Villas, Santa Clara, 54830, Villa Clara, Cuba.

[3] Institut Universitari de Ciència Molecular, Universitat de València, Edifici d'Instituts de Paterna, P.O. Box 22085, E-46071, València, Spain.

[4] Unidad de Investigación de Diseño de Fármacos y Conectividad Molecular, Departamento de Química Física, Facultad de Farmacia, Universitat de València, Spain.

Received August 1, 2011; accepted April 1, 2013

Abstract. The recently introduced, non-stochastic and stochastic quadratic indices (Marrero-Ponce *et al. J. Comp. Aided Mol. Des.* **2006**, *20*, 685-701) were applied to QSAR/QSPR studies of heteroatomic molecules. These novel bond-based molecular descriptors (MDs) were used for the prediction of the partition coefficient (log P), and the antibacterial activity of 34 derivatives of 2-furylethylenes. Two statistically significant QSPR models using non-stochastic and stochastic bond-based quadratic indices were obtained ($R^2 = 0.971$, $s = 0.137$ and $R^2 = 0.986$, $s = 0.096$). These models showed good stability to data variation in leave-one-out (LOO) cross-validation experiment ($q^2 = 0.9975$, $s_{CV} = 0.164$ and $q^2 = 0.947$, $s_{CV} = 0.114$). The best two discriminant models computed using the non-stochastic and stochastic molecular descriptors had globally good classification of 94.12% in the training set. The external prediction sets had accuracies of 100% in both cases. The comparison with other approaches (edge- and vertex-based connectivity indices, total and local spectral moments, quantum chemical descriptors as well as with other TOMOCOMD-CARDD MDs) revealed the good performance of our method in this QSPR study. The obtained results suggest that it is possible to obtain a good estimation of physical, chemical and physicochemical properties for organic compounds with the present approach.

Key words: TOMOCOMD-CARDD Software, non-Stochastic and Stochastic Bond-Based Quadratic Indices, Edge-Adjacency Matrix, Stochastic Matrix, QSPR/QSAR Model, 2-furylethylene.

Resumen. En el presente reporte, se aplican los índices cuadráticos de relaciones de enlace introducidos recientemente (Marrero-Ponce *et al. J. Comp. Aided Mol. Des.* **2006**, *20*, 685-701) en estudios QSAR/QSPR de moléculas heteroatómicas. Estos descriptores moleculares de tipo enlace son usados en la predicción del coeficiente de partición (log P) y la actividad antibacterial de 34 derivados de los 2-furiletilenos. Dos modelos QSPR estadísticamente significativos fueron obtenidos usando índices no estocásticos y estocásticos ($R^2 = 0.971$, $s = 0.137$ y $R^2 = 0.986$, $s = 0.096$, respectivamente) en la modelación del Log P. Estos modelos mostraron una estabilidad adecuada en la validación interna LOO ($q^2 = 0.9975$, $s_{CV} = 0.164$ and $q^2 = 0.947$, $s_{CV} = 0.114$, respectivamente). Por otro lado, los dos mejores modelos discriminantes muestran un porcentaje de exactitud global de 94.12% en la serie de entrenamiento y de 100% en la data de predicción en la modelación de actividad bactericida. Finalmente, la comparación con otros enfoques computacionales (índices de conectividad de enlace y de átomo tanto 2D como 3D, momentos espectrales totales y locales, descriptores químicos cuánticos al igual que con otros índices implementados en el programa TOMOCOMD-CARDD) evidencia un buen comportamiento de nuestros nuevos índices. Los resultados obtenidos sugieren que el método propuesto permite obtener una adecuada estimación de propiedades fisicoquímicas y biológicas de moléculas orgánicas.

Palabras clave: Programa TOMOCOMD-CARDD, índices cuadráticos de enlace, matriz de adyacencia de enlace, matriz estocástica, modelo QSPR/QSAR, 2-furiletileno.

Background

During the past decade, a great explosion of molecular descriptors (MDs) has been observed. For instance, topological indices (TIs), surface areas, volume descriptors, charges, and quantum-chemical measures have been extensively enhanced and used as whole molecule MDs [1-3]. However, local MDs have received very little attention [4]. One exception in this sense is the electrotopological state (E-state) index [5]. Other "global" MDs such as spectral moments of the edge-adjacency matrix have been redefined to their local form [4]. In this sense, in a manner similar to that for the atom- and atom-type level E-State, an E-State index for bonds and bond-type has been proposed. The bond-based E-State indices provided an improvement of 25% with regard to the atom-based E-State indices in the description of the boiling point of 372 alkanes, alcohols, and chloroalkanes [5].

The edge (bond)-adjacency relationships have also been used in the generation of new TIs [1-3]. Their matrix form has been considered and explicitly defined in the chemical graph theory literature, but has attracted little interest in both chemical and mathematical literature. Nevertheless, in the last decade Estrada rediscovered this matrix as an important source of graph theoretical invariants useful in the generation of new MDs [1]. For instance, first the edge-connectivity descriptor ϵ was defined by this author using the Randić-type graph-theoretical invariant [6]. That is to say, this new index is analogous to the Randić branching index but calculated by edge degrees instead of vertex degrees. In a second work, Estrada also extended the edge adjacency matrix **E** for a molecular graph to a 3D-**E** matrix in order to generate the so-called topographic edge-connectivity index $\epsilon(\rho)$ [7], also using the Randić-type graph-theoretical invariant. Later, this author used the same edge adjacency relationships in the generation of a new family of TIs, spectral moments of the E-matrix [7]. The analogous concept of spectral moments of vertex-adjacency matrix had also been previously discussed by different authors [8]. Afterward, Estrada et al. [9] introduced an extended set of edge connectivity indices, ${}^{m}\epsilon_{t}(\mathrm{G})$, using the same way in which the Randić branching index was extended to the series of molecular connectivity indices. Finally, a novel graph theoretical polynomial, $\mathrm{P}_{\epsilon}(G,x)$, counting the edge connectivity was introduced by the same researcher [10]. Such edge-adjacency relationships will be applied in the present report in order to generate a series of bond-based MDs to be used in drug design and chemoinformatic studies.

Recently one of the present authors, Y.M-P, has introduced a new set of atom-level molecular descriptors of relevance to QSAR/QSPR studies and 'rational' drug design, non-stochastic and stochastic quadratic indices [$\boldsymbol{q}_k(\bar{x})$ and ${}^{s}\boldsymbol{q}_k(\bar{x})$, respectively] [11, 12]. These local (atom, group and atom-type) and total chemical indices are based on the calculation of quadratic maps in $\Re^n$ in canonical basis set. The description of the significance-interpretation and the comparison with other molecular descriptors was also performed [12]. This approach describes changes in the electronic distribution with time throughout the molecular backbone. Specifically, the features of the *k*th total and local quadratic indices were illustrated by examples involving various molecular structural changes, such as chain lengthening and branching as well as the inclusion of heteroatoms and multiple bonds [12]. Additionally, the linear independence of the atom-type quadratic fingerprints to other 0D-3D molecular descriptors was demonstrated. In this sense, it was concluded that local (atom-based) quadratic fingerprints are independent indices, which contain important structural information to be used in QSPR/QSAR and drug design studies [12].

These MDs are easily and quickly calculated, thus being suitable for both QSAR/QSPR modeling and drug design studies of large chemical databases. This -*in silico*- method has been successfully applied to the prediction of several physical, physicochemical and chemical properties of organic compounds [11, 12]. These atom-level MDs, and their stochastic forms [13, 14], have also been useful for the selection of novel subsystems of compounds having a desired property/activity. In this sense, it was successfully applied to the virtual (computational) screening of novel anthelmintic compounds, which were then synthesized and evaluated *in vivo* on *Fasciola hepatica* [15]. Studies for the fast-track discovery of novel antibacterial [16], paramphistomicide [13], antimalarial [14, 17], trichomonicidal [16], and tripanocidals [18] lead-like chemicals were also conducted with this theoretical approach. In addition, the atom-based quadratic indices have been extended to consider three-dimensional features of small/medium-sized molecules based on the trigonometric-3D-chirality-correction factor approach [19, 20]. This approach has also been successfully employed in QSAR and *in silico* ADME studies of Caco-2 Permeability of Drugs [21-23].

The main aim of this paper is test the correlation ability of the new MDs, calculated as quadratic maps similar to those defined in linear algebra, in QSPR/QSAR studies to examine the partition coefficient (log *P*), as well as the antibacterial activity of 34 derivatives of 2-furylethylenes.

Material and Methods

Dataset Selection for QSPR/QSAR studies

The decisive criterion of quality for any MDs is its ability to describe structure-related properties of molecules. The QSPR/QSAR models were developed to examine the partition coefficient (log *P*), as well as the antibacterial activity of 34 derivatives of 2-furylethylenes [4, 24].

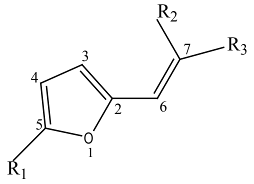
The heteromolecule-based database consisting of 34 2-furylethylene derivatives (see Table 1), was recently studied by using total and local spectral moments, 2D/3D (vertex- and edge-) connectivity indices and two quantum-chemical descriptors [4, 24]. These chemicals have different substituents at position 5 of the furan ring, as well as at the β position of the exocyclic double bond [25]. The values of the *n*-octanol/water partition coefficient (log *P*) of these compounds have been experimentally determined and reported in the literature [25]. 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*) [25]. This antibacterial activity was used to classify furylethylenes into two groups by Estrada and Molina [24]. The group of active compounds is composed of those substance having values of log(1/*C*) < 3, while the rest form the group of inactive molecules. In this study, we also took into account a series of nine new 2-furylethylenes, used by Estrada and Molina [24] as external prediction (test) set. These compounds have an NO_2 group at position R_3 and a Br or I at positions R_1 and/or R_2. All these compounds showed antibacterial activity in different assays [26].

Computational Strategies

The total and local (bond-type) bond-based quadratic indices used to search for the best regression of the selected physicochemical property of 2-furylethylenes were calculated by the TOMOCOMD-CARDD (acronym of TOpological MOlecular COMputational Design-Computer Aided "Rational" Drug Design) program [27]. This software is an interactive program for molecular design and bioinformatic research. The software was developed based on a user-friendly philosophy. That is to say, this computational program offers an interactive environment for the user and does not require prior programming skills. Computer Aided "Rational" Drug Design) subprogram allows drawing the structures (drawing mode) and calculating 2D (topologic), 3D-chiral (2.5D) and 3D (geometric and topographic) non-stochastic and stochastic MDs (calculation mode).

The bond–based TOMOCOMD-CARDD MDs computed in this study were the following:

Table 1. 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	$CONHCH(CH_3)C_2H_5$
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	$CONHCH_2C(CH_3)_3$
4	I	NO_2	$COOCH_3$	21	NO_2	H	$COOCH_3$
5	$COOCH_3$	NO_2	$COOCH_3$	22	NO_2	H	$COOC_2H_5$
6	NO_2	NO_2	$COOCH_3$	23	NO_2	H	$COO(CH_2)_2CH_3$
7	NO_2	$COOC_2H_5$	$COOC_2H_5$	24	NO_2	H	$COOCH(CH_3)_2$
8	NO_2	H	NO_2	25	NO_2	H	$COO(CH_2)_3CH_3$
9	H	H	NO_2	26	NO_2	H	$COOCH_2CH(CH_3)_2$
10	NO_2	H	$CONH_2$	27	NO_2	H	$COOCH(CH_3)C_2H_5$
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	$COO(CH_2)_4CH_3$
13	NO_2	H	$CONHC_2H_5$	30	NO_2	H	Br
14	NO_2	H	$CONH(CH_2)_2CH_3$	31	NO_2	H	CN
15	NO_2	H	$CONHCH(CH_3)_2$	32	NO_2	H	OCH_3
16	NO_2	H	$CONH(CH_2)_3CH_3$	33	NO_2	H	H
17	NO_2	H	$CONHCH_2CH(CH_3)_2$	34	NO_2	CN	$COOCH_3$

Novel R$_1$,R$_2$-Substituted 2-Furylethylenes (R$_3$ = NO_2) 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				

1) kth (k = 15) total non-stochastic bond-based quadratic indices not considering and considering H-atoms in the molecular graph (G) [$\boldsymbol{q}_{\boldsymbol{k}}(\overline{w})$ and $\boldsymbol{q}_{\boldsymbol{k}}^{\mathrm{H}}(\overline{w})$, respectively].
2) kth (k = 15) total stochastic bond-based quadratic indices not considering and considering H-atoms in the molecular graph (G) [$^{\mathrm{s}}\boldsymbol{q}_{\boldsymbol{k}}(\overline{w})$ and $^{\mathrm{s}}\boldsymbol{q}_{\boldsymbol{k}}^{\mathrm{H}}(\overline{w})$, correspondingly].
3) kth (k = 15) group (heteroatoms: O, N, S and halogens) non-stochastic quadratic indices considering and non-considering H-atoms in the molecular graph (G) [$\boldsymbol{q}_{\boldsymbol{k}\mathrm{L}}^{\mathrm{H}}(\overline{w}_{\mathrm{E}})$ and $\boldsymbol{q}_{\boldsymbol{k}\mathrm{L}}(\overline{w}_{\mathrm{E}})$, correspondingly]. These local MDs are putative molecular charge, dipole moment, and H-bonding acceptors.
4) kth (k = 15) group (heteroatoms: O, N, S and halogens) stochastic quadratic indices considering and non-considering H-atoms in the molecular graph (G) [$^{\mathrm{s}}\boldsymbol{q}_{\boldsymbol{k}\mathrm{L}}^{\mathrm{H}}(\overline{w}_{\mathrm{E}})$ and $^{\mathrm{s}}\boldsymbol{q}_{\boldsymbol{k}\mathrm{L}}(\overline{w}_{\mathrm{E}})$, respectively]. These local MDs are putative molecular charge, dipole moment, and H-bonding acceptors.
5) kth (k = 15) bond-type (C2-C6) non-stochastic and stochastic quadratic indices considering H-atoms in the molecular graph (G) [$\boldsymbol{q}_{\boldsymbol{k}\mathrm{L}}^{\mathrm{H}}(\overline{w}_{\mathrm{C2\text{-}C6}})$ and $^{\mathrm{s}}\boldsymbol{q}_{\boldsymbol{k}\mathrm{L}}^{\mathrm{H}}(\overline{w}_{\mathrm{C2\text{-}C6}})$,respectively].
6) kth (k = 15) bond-type (C2-C6) non-stochastic and stochastic quadratic indices not considering H-atoms in the molecular graph (G) [$\boldsymbol{q}_{\boldsymbol{k}\mathrm{L}}(\overline{w}_{\mathrm{C2\text{-}C6}})$ and $^{\mathrm{s}}\boldsymbol{q}_{\boldsymbol{k}\mathrm{L}}(\overline{w}_{\mathrm{C2\text{-}C6}})$, respectively].
7) kth (k = 15) bond-type (C6-C7) non-stochastic and stochastic quadratic indices considering H-atoms in the molecular graph (G) [$\boldsymbol{q}_{\boldsymbol{k}\mathrm{L}}^{\mathrm{H}}(\overline{w}_{\mathrm{C6\text{-}C7}})$ and $^{\mathrm{s}}\boldsymbol{q}_{\boldsymbol{k}\mathrm{L}}^{\mathrm{H}}(\overline{w}_{\mathrm{C6\text{-}C7}})$, respectively].
8) kth (k = 15) bond-type (C6-C7) non-stochastic and stochastic quadratic indices not considering H-atoms in the molecular graph (G) [$\boldsymbol{q}_{\boldsymbol{k}\mathrm{L}}(\overline{w}_{\mathrm{C6\text{-}C7}})$ and $^{\mathrm{s}}\boldsymbol{q}_{\boldsymbol{k}\mathrm{L}}(\overline{w}_{\mathrm{C6\text{-}C7}})$, respectively].

Chemometric Analysis

These kth total and local bond-based quadratic indices were used as MDs for derived QSPRs. One of the difficulties with the large number of MDs is deciding which ones will provide the best regressions, considering both goodness of fit and the chemical meaning of the regression. In addition, as testing a large number of all possible combinations of variables would be a tedious task and time-consuming procedure, we have used a genetic algorithm (GA) input selection [28-35]. The GAs are a class of algorithms inspired by the process of biological evolution in which species having a high fitness under some conditions can prevail and survive to the next generation; the best species can be adapted by crossover and/or mutation in the search for better individuals.

The software BuildQSAR [36] was employed to perform variable selection and QSAR modeling. The mutation probability was specified as 35%. The size of the equations was set at three-four terms and a constant. The population size was established as 300. The GA with an initial population size of 300 rapidly converged (2000 generations) and reached an optimal QSAR model in a reasonable number of GA generations.

The search for the best model can be processed in terms of the highest correlation coefficient (*R*) or F-test equations (Fisher-ratio's *p*-level [*p*(F)]), and the lowest standard deviation equations (s) [36]. The quality of models was also determined by examining the Leave-One-Out (LOO) cross validation (CV) (q^2, s_{cv}) [73]. In recent years, the LOO press statistics (e.g., q^2) have been used as a means of indicating predictive ability. Many authors consider high q^2 values (for instance, $q^2 > 0.5$) as an indicator or even as the ultimate proof of the high predictive power of a QSAR model. The Hasse diagram technique may be used to rank the QSAR models in terms of their respective statistics [37].

On the other hand, linear discriminant analysis (LDA) was used in the classification of the 34 2-furylethylene derivatives according to their antibacterial activity. This statistical analysis was performed using STATISTICA software [38]. In order to test the quality of the discriminant function derived, we used the Wilk's λ (U-statistic) and the Mahalanobis distance (D^2). The Wilk's λ statistic is helpful to evaluate the group 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 classify the 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. The posterior probability is the probability, based on our knowledge of the values of other 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.

Applications in QSPR/QSAR Studies and Comparison with other Computacional Approaches

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

The partition coefficient *n*-octanol/water (log *P*) has an important role in the understanding of the biological behavior of the 34 2-furylethylenes derivatives [25], specifically for the development of their antibacterial activity [26]. The values of the *n*-octanol/water log *P* of these compounds have been experimentally determined and reported in the literature [25]. This experiment offers the possibility of comparing the present results with those achieved by using some atom-based TOMOCOMD-CARDD MDs [12, 39]. The best obtained models together with their statistical parameters using bond-based (non-stochastic and stochastic) quadratic indices respectively are given below:

$$\begin{aligned}\log P = {} & 1.937(\pm 0.285) + 5.40\text{e-}06(\pm 7.5\text{e-}07)^{M}\boldsymbol{q}_{6L}{}^{H}(\overline{w}_{E}) \\ & - 1.89\text{e-}03(\pm 5.3\text{e-}04)^{V}\boldsymbol{q}_{0}{}^{H}(\overline{w}) + 2.18\text{e-}03(\pm 5.5\text{e-} \\ & 04)^{V}\boldsymbol{q}_{0L}{}^{H}(\overline{w}_{E}) + 0.182(\pm 0.022)^{P}\boldsymbol{q}_{0}(\overline{w}) \\ & - 0.152(\pm 0.012)^{P}\boldsymbol{q}_{2L}{}^{H}(\overline{w}_{E}) + 0.047(\pm 0.005)^{K}\boldsymbol{q}_{1L}{}^{H}(\overline{w}_{E}) \\ & - 0.033(\pm 0.004)^{K}\boldsymbol{q}_{1L}(\overline{w}_{E}) \end{aligned} \tag{1}$$

$$N = 34 \quad R^2 = 0.971 \quad q^2 = 0.947 \quad s = 0.137 \quad s_{CV} = 0.164$$
$$F(7,26) = 124.72 \quad p < 0.0001$$

$$\begin{aligned}\log P = {} & 0.545(\pm 0.152) + 7.39\text{e-}03(\pm 3.3\text{e-}04)^{M}\boldsymbol{q}_{7L}{}^{H}(\overline{w}_{E}) \\ & + 0.051(\pm 0.005)^{V}\boldsymbol{q}_{4H}(\overline{w}) \\ & - 0.048(\pm 0.005)^{V}\boldsymbol{q}_{5}{}^{H}(\overline{w}) \\ & + 6.82\text{e-}04(\pm 5.0\text{e-}05)^{V}\boldsymbol{q}_{2}(\overline{w}) \\ & - 0.011(\pm 0.001)^{V}\boldsymbol{q}_{4L}{}^{H}(\overline{w}_{E}) \\ & - 2.267(\pm 0.128)^{P}\boldsymbol{q}_{5L}{}^{H}(\overline{w}_{E}) \\ & + 0.256(\pm 0.013)^{K}\boldsymbol{q}_{7L}{}^{H}(\overline{w}_{E}) \end{aligned} \tag{2}$$

$$N = 34 \quad R^2 = 0.986 \quad q^2 = 0.975 \quad s = 0.096 \quad S_{CV} = 0.114$$
$$F(7,26) = 257.58 \quad p < 0.0001$$

where, N is the number of compounds, R² is the determination coefficient, s is the standard deviation of the regression, q^2 (s_{CV}) is the square regression coefficient (standard deviation) obtained from the LOO cross validation procedure, and *F* is the Fisher ratio.

These models show significant superiority from a statistic point of view with respect to other methods previously used for the same data set. Our models (non-stochastic and stochastic) explain 97.1% and 98.6% of the experimental variance of log *P* while the previous approaches describe less than 97% of the variance. In this sense, standard deviation between 0.142 and 0.319 were reported in previous studies while our models show the lowest values of standard deviation ($s = 0.137$ and $s = 0.096$ for Eqs. **1** and **2**, correspondingly). In addition, the statistics for the LOO CV procedure, achieved by our models were in general better than those obtained by the other methods, for a detailed comparison see Table 2.

The values of experimental and calculated values of the Log P for the data set (both models) are given in Table 3. Plots of the linear relationship between the observed and calculated log *P* values for the data set of compounds are illustrated in Figures 1 and 2.

Table 2. Statistical Parameters for the Models Describing the log P of 34 2-furylethylene derivatives by Using Different MDs.

Partition Coefficient *n*-Octanol/Water (log *P*) of 34 2-Furylethylenes						
Index	n	R^2	*s*	q^2	s_{cv}	F
Bond-based Non-stochastic QI (Eq. 1)	7	0.971	0.137	0.947	0.164	124.7
Bond-based stochastic QI (Eq. 2)	7	0.986	0.096	0.974	0.114	257.6
Atom-based Non-stochastic QI [12]	7	0.969	0.142	0.951	0.156	116.76
Atom-based Non-stochastic LI [39]	7	0.968	0.143	0.938	0.176	113.38
Vertex and edge conn. Indices [24]	7	0.939	0.199	*	0.247	56.90
Topographic descriptors [24]	7	0.964	0.155	*	0.176	84.60
Quantum chemical descriptors [24]	R&C[a]	0.875	0.319	*	0.370	45.50

[a]used the Rogers and Cammarata approach. *values not reported. QI (Quadratic Indices) and LI (Linear Indices)

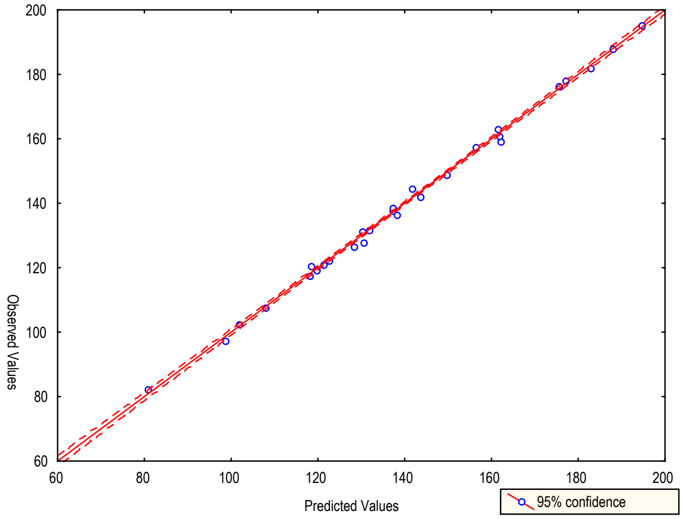


Fig. 1. Linear correlations of observed versus calculated log P of the 2-furylethylene derivatives according to the model obtained from non-stochastic bond-level quadratic indices (Eq. 1).

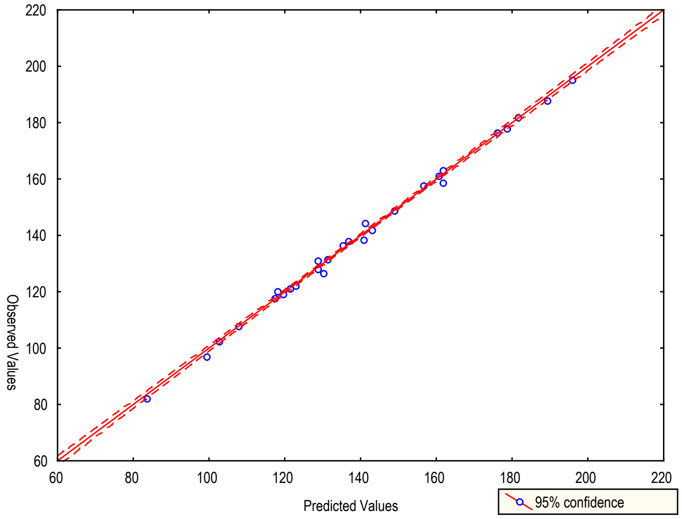


Fig. 2. Linear correlations of observed versus calculated Log P of the 2-furylethylene derivatives according to the model obtained from stochastic bond-level quadratic indices (Eq.2).

Classifying 34 2-furylethylene derivatives as antibacterial

Finally, Linear Discriminant Analysis (LDA) is used to obtain classification models for the 2-furylethylene compounds according to their antibacterial activity. The classification models obtained are given below together with their statistical parameters:

$$\begin{aligned} \mathit{Class}\ \mathrm{Act} = {} & -185.67 - 2.357\,{}^{\mathrm{K}}\boldsymbol{q}^{\mathbf{H}}{}_{4\mathrm{L}}(\overline{w}_{\mathrm{E}}) \\ & + 2.187\,{}^{\mathrm{K}}\boldsymbol{q}_{1\mathrm{L}}(\overline{w}_{\mathrm{E}}) + 185.73\ {}^{\mathrm{K}}\boldsymbol{q}^{\mathbf{H}}{}_{7\mathrm{L}}(\overline{w}_{\mathrm{C2\text{-}C6}}) \end{aligned} \qquad (3)$$

$$N = 34 \quad \lambda = 0.262 \quad D^2 = 10.92 \quad F(3,30) = 28.10$$
$$p < 0.0001$$
$$Q_{Total} = 94.12\% \quad MCC = 0.89 \quad Sen = 100 \quad Spec = 87.5$$

$$\begin{aligned} \mathit{Class}\ \mathrm{Act} = {} & -20.806 - 0.0191\,{}^{V}\boldsymbol{q}^{\mathbf{H}}{}_{8}(\overline{w}) + 0.0130\,{}^{\mathrm{V}}\boldsymbol{q}_{1\mathrm{L}}(\overline{w}_{\mathrm{E}}) \\ & + 0.3028\ {}^{V}\boldsymbol{q}^{\mathbf{H}}{}_{10L}(\overline{w}_{\mathrm{C6\text{-}C7}}) \end{aligned} \qquad (4)$$

$$N = 34 \quad \lambda = 0.282 \quad D^2 = 9.91 \quad F(3,30) = 25.51$$
$$p < 0.0000$$
$$Q_{Total} = 94.12\% \quad MCC = 0.89 \quad Sen = 100 \quad Spec = 87.5$$

where, λ is Wilk's statistic, D^2 is the squares of Mahalanobis distances, and F is the Fisher ratio, Q_{Total} is the accuracy of the model for the training set, MCC is the Matthews' correlation coefficient, *Sen* and *Spec* are the sensibility and specificity of the model, respectively. The statistical analysis showed that there exists appropriate discriminatory power for differentiating between the two respective groups.

The non-stochastic model (Eq. **3**), has an accuracy of 94.12% for the training set, misclassifying only 2 compounds of a total of 34. This model showed a high MCC of 0.89; MCC quantifies the strength of the linear relation between the molecular descriptors and the classifications, and it may often

Table 3. Experimental and calculated values of the partition coefficient *n*-octanol/water (log *P*) for the 2-Furylethylenes derivatives.

No.	Obsd.[a]	Pred.[b]	$\mathrm{Res}_{\mathrm{VC\text{-}LOO}}$[c]	Pred.[d]	$\mathrm{Res}_{\mathrm{VC\text{-}LOO}}$[e]
1	1.879	1.761	−0.209	1.755	−0.209
2	2.439	2.240	0.351	2.548	0.351
3	2.739	2.721	0.651	2.760	0.651
4	2.999	2.994	0.911	2.997	0.911
5	1.869	1.751	−0.219	1.846	−0.219
6	1.599	1.667	−0.489	1.529	−0.489
7	2.504	2.687	0.416	2.666	0.416
8	1.303	1.386	−0.785	1.359	−0.785
9	1.583	1.483	−0.505	1.600	−0.505
10	0.649	0.816	−1.439	0.756	−1.439
11	0.984	1.038	−1.104	1.022	−1.104
12	0.819	0.900	−1.269	0.833	−1.269
13	1.386	1.358	−0.702	1.380	−0.702
14	1.860	1.873	−0.228	1.790	−0.228
15	1.803	1.740	−0.285	1.812	−0.285
16	2.356	2.262	0.268	2.300	0.268
17	2.225	2.311	0.137	2.235	0.137
18	2.284	2.335	0.196	2.308	0.196
19	2.333	2.173	0.245	2.257	0.245
20	2.605	2.741	0.517	2.685	0.517
21	1.652	1.572	−0.436	1.748	−0.436
22	2.098	2.030	0.010	2.138	0.010
23	2.673	2.587	0.585	2.489	0.585
24	2.641	2.535	0.553	2.585	0.553
25	2.827	2.996	0.739	2.986	0.739
26	3.135	3.067	1.047	2.927	1.047
27	3.091	3.176	1.003	3.077	1.003
28	3.060	3.106	0.972	3.014	0.972
29	3.404	3.346	1.316	3.505	1.316
30	2.447	2.482	0.359	2.429	0.359
31	1.050	1.332	−1.038	1.105	−1.038
32	1.591	1.780	−0.497	1.592	−0.497
33	1.611	1.484	−0.477	1.554	−0.477
34	1.488	1.257	−0.600	1.398	−0.600

[a]Experimental values of log *P*. [b,d]Predicted values using non-stochastic (Eq. 1) and stochastic (Eq. 2) bond-based linear indices, respectively. [c,e]Residual values of LOO cross-validation process using non-stochastic and stochastic bond-based linear indices, respectively [$\mathrm{Res}_{\mathrm{CV\text{-}LOO}}$ = Bp(Obsd.) - Bp($\mathrm{Pred.}_{\mathrm{\text{-}CV\text{-}LOO}}$)].

provide a much more balanced evaluation of the prediction than, for instance, the percentages (accuracy) [40]. Nevertheless, the most important criterion, for the acceptance or not of a discriminant model, is based on the statistic for the external prediction set. The non-stochastic model showed an accuracy of 100% (MCC = 1.00) for the compounds in the test set.

A rather similar behavior was obtained with the stochastic quadratic indices (Eq. **4**), but with a greater λ and a shorter Mahalanobis distance (λ = 0.282 D^2 = 9.91). Quite similar results were shown by the model previously obtained with atom-based (non-stochastic and stochastic) bilinear indices and atom-based non stochastic linear indices, which as well as our model used three parameters. The model developed with atom-based non stochastic quadratic indices had similar statistical parameters, but it was obtained with five descriptors, however our indices showed better results than the models derived previously by Estrada and Molina, using 2D and 3D connectivity and quantum chemical descriptors; for a detailed comparison see Table 4. The overall accuracy of these models in both, training and external prediction sets achieved with all these approaches is shown in Table 5.

Concluding remarks

We have shown here that total and local bond-based quadratic indices are useful MDs for modeling physicochemical properties of heteroatomic-organic chemicals. The obtained QSPR/QSAR models for the description and prediction of log *P* of 34 2-furylethylene derivatives were statistically significant and better than those obtained previously using recognized methods, such as topological [total and local spectral moment and 2D (edge- and vertex-) connectivity indices], topographic and quantum chemical descriptors as well as some atom-level TOMOCOMD-CARDD MDs. This point is important because of the well-known broad applicability of these MDs in QSPR/QSAR studies. As a consequence, the bond-based quadratic indices represent a novel source for successful structure/activity-property models and drug design strategies.

Table 4. Statistical parameters for the classification of 34 2-furylethylene derivatives as antibacterial by Using Different MDs.
Classification of 34 2-Furylethylene Derivatives as Antibacterial

Index	n	λ	D^2	Accuracy (Training)	Accuracy (Test)	F
Bond-based Non-stochastic QI (Eq. 3)	3	0.262	10.92	94.12%	100%	28.10
Bond-based stochastic QI (Eq. 4)	3	0.282	9.91	94.12%	100%	25.51
Atom-based Non-stochastic BI[41]	3	0.289	9.54	97.06%	100%	24.56
Atom-based Stochastic BI[41]	3	0.297	8.87	94.12%	100%	22.83
Atom-based Non-stochastic LI [39]	3	0.300	9.44	94.12%	100%	22.90
Atom-based Non-stochastic QI[12]	5	0.259	11.78	97.06%	100%	15.98
Vertex and edge conn. Indices [24]	5	0.43	5.7	91.2%	100%	7.70
Topographic descriptors[24]	5	0.38	6.7	94.1%	100%	9.10
Quantum chemical descriptors [24]	5	0.44	5.2	88.2%	100%	7.10

QI (Quadratic Indices), LI (Linear Indices), and BI (Bilinear Indices).

Table 5. Classification of 2-furylethylene derivatives as antibacterial according to the models obtained with non-stochastic and stochastic bond-based quadratic indices.

		(Eq.3)		(Eq.4)	
		Class.	Prob. %	Class.	Prob. %
		Training set			
1	+	+	99.93	+	99.97
2	+	+	100.00	+	99.60
3	+	+	100.00	+	99.97
4	+	+	100.00	+	99.96
5	+	+	98.65	+	99.97
6	+	+	99.86	+	99.82
7	+	+	97.22	+	99.61
8	+	+	99.90	+	86.90
9	+	+	99.95	+	97.38
10	+	+	99.11	+	50.74
11	+	+	74.04	+	73.34
12	+	+	98.61	+	99.78
13	+	+	62.63	+	55.94
14	-	-	2.89	-	0.75
15	-	-	19.31	-	1.18
16	-	-	1.86	-	0.21
17	-	-	0.23	-	0.05
18	-	-	0.59	-	0.08
19	-	-	2.25	-	0.04
20	-	-	0.03	-	0.01
21	-	+	**81.17**	+	73.31
22	-	+	**50.43**	-	44.90
23	-	-	0.81	-	0.55
24	-	-	5.22	-	0.71
25	-	-	0.42	-	0.14
26	-	-	0.03	-	0.04
27	-	-	0.06	-	0.05
28	-	-	0.17	-	0.02
29	-	-	0.39	-	0.04
30	-	-	0.01	-	40.67
31	-	-	0.07	-	1.82
32	-	-	16.21	+	**61.58**
33	-	-	0.00	-	0.09
34	+	+	95.90	+	87.14
		Test set			
1	+	+	97.06	+	99.93
2	+	+	98.76	+	99.93
3	+	+	100.00	+	98.50
4	+	+	58.24	+	99.94
5	+	+	100.00	+	98.32
6	+	+	79.00	+	99.97
7	+	+	100.00	+	81.24
8	+	+	100.00	+	80.33
9	+	+	100.00	+	76.30

Outlook

The development of more powerful MDs carries sustained interest in the drug discovery process. That is to say, although there have been many discoveries in the recent years in the field of theoretical drug-design it is necessary to continue developing new MDs that can represent, by means of QSAR (or similar theoretical works) studies, different physicochemical properties and biological activities of chemical substances. Therefore, our research group is working towards the definition of novel 2D/3D MDs based on algebra and group theory, geometric properties, discrete mathematics, etc. We are also interested in developing new (standard) rules and doubly stochastic indices.

Applications of these new bond (edge)-level MDs in molecular property/activity modeling, similarity/diversity analysis and *biosilico* drug discovery will be published in forthcoming papers.

Acknowledgments

One of the authors (M-P. Y) thanks the program '*Estades Temporals per a Investigadors Convidats*' for a fellowship to work at Valencia University (2012). The authors acknowledge also the partial financial support from Spanish Ministry of Science and Innovation (MICINN, project reference: SAF2009-10399). Finally, but not least, this work was supported in part by VLIR (Vlaamse InterUniversitaire Raad, Flemish Interuniversity Council, Belgium) under the IUC Program VLIR-UCLV.

References

1. Todeschini, R.; Consonni, V. *J. Chem. Inf. Comput. Sci.* **2000**.
2. Consonni, V.; Todeschini, R.; Pavan, M. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 682.
3. Seybold, P. G.; May, M.; Bagal, U.A. *J. Chem. Ed.* **1987**, *64*, 575.
4. Estrada, E.; Molina, E. *J. Mol. Graphics Mod.* **2001**, *20*, 54-64.
5. Kier, L. B.; Hall, L. H. *Academic Press* **1999**.
6. Estrada, E. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 31.
7. Estrada, E.; Ramírez, A. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 837.
8. Marković, S.; Gutman, I. *J. Mol. Struct. (Theochem)* **1991**, *235*, 81.
9. Estrada, E.; Guevara, N.; Gutman I. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 428.
10. Estrada, E.; Rodríguez, L. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1037.
11. Marrero-Ponce, Y. *Molecules*, **2003**, *8*, 687-726.
12. Marrero-Ponce, Y. *Bioorg. Med. Chem.* **2004**, *12*, 6351.
13. Marrero-Ponce, Y.; Huesca-Guillen, A.; Ibarra-Velarde, F. *J. Mol. Struct. (Theochem)* **2005**, *717*, 67.
14. Marrero-Ponce, Y.; Medina-Marrero, R.; Torrens, F.; Martinez, Y.; Romero-Zaldivar, V.; Castro, E. A. *Bioorg. Med. Chem.* **2005**, *13*, 2881.
15. Marrero-Ponce, Y.; Castillo-Garit, J. A.; Olazabal, E.; Serrano, H. S.; Castañedo, N.; Ibarra-Velarde, F.; Huesca-Guillen, A.; Valle, A.; Torrens, F.; Castro, E. *J. Comput.-Aided Mol. Design*, **2004**, *18*, 615.

16. Meneses-Marcel, A.; Marrero-Ponce, Y.; Machado-Tugores, Y.; Montero-Torres, A.; Montero Pereira, D.; Escario, J. A.; Nogal-Ruiz, J. J.; Ochoa, C.; Arán, V. J.; Martínez-Fernández, A. R.; García Sánchez, R. N. *Bioorg. Med. Chem. Lett.* **2005**, *17*, 3838.
17. Montero-Torres, A.; García-Sánchez, R. N.; Marrero-Ponce, Y.; Machado-Tugores, Y.; Nogal-Ruiz, J. J.; Martínez-Fernández, A. R.; Arán, V. J.; Ochoa, C.; Meneses-Marcel, A.; Torrens, F. *J. Med. Chem.* **2006**, *41*, 483-493.
18. Montero-Torres, A.; Celeste Vega, M.; Marrero-Ponce, Y.; Rolón, M.; Gómez-Barrio, A.; Escario, J. A.; Arán, V. J.; Martínez-Fernández, A. R.; Meneses-Marcel, A. *Bioorg. Med. Chem.* **2005**, *13*, 6264-6275.
19. Marrero-Ponce, Y.; Díaz, H. G.; Romero, V.; Torrens, F.; Castro, E. A. *Bioorg. Med. Chem.* **2004**, *12*, 5331.
20. Castillo-Garit, J. A.; Marrero-Ponce, Y.; Torrens, F. *Bioorg. Med. Chem.* **2006**, *14*, 2398-2408.
21. Marrero-Ponce, Y.; Cabrera, M. A.; Romero, V.; Ofori, E.; Montero, L. A. *Int. J. Mol. Sci.* **2003**, *4*, 512.
22. Marrero-Ponce, Y.; Cabrera, M. A.; Romero, V.; González, D. H.; Torrens, F. *J. Pharm. Pharmaceut. Sci.* **2004**, *7*, 186.
23. Marrero-Ponce, Y.; Cabrera, M. A.; Romero-Zaldivar, V.; Bermejo, M.; Siverio, D.; Torrens, F. *J. Mol. Des.* **2005**, *4*, 124.
24. Estrada, E.; Molina, E. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 791.
25. Balaz, S.; Sturdik, E.; Rosenberg, M.; Augustin, J.; Skara, B. *J. Theor. Biol.* **1988**, *131*, 115.
26. Blondeau, J.; Castañedo, N.; Gonzalez, O.; Medina, R.; Silveira, E. *Antimicrob. Agents Chemother* **1999**, *11*, 163.
27. Marrero-Ponce, Y.; Romero, V. Central University of Las Villas. Santa Clara, **2002**.
28. Senese, C. L.; Hopfinger, A. J. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1297.
29. Goldberg, D. E. *Genetic Algorithms*, Addison Wesley, MA, **1989**.
30. Willet, P. *Trends Biotechnol.* **1995**, *13*, 516.
31. So, S. S.; Karplus, M. *J. Med. Chem.* **1996**, *39*, 1521.
32. So, S. S.; Karplus, M. *J. Med. Chem.* **1997**, *40*, 4347.
33. Rogers, D.; Hopfinger, A. J. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 854.
34. Hopfinger, A. J.; Wang, S.; Tokarski, J. S.; Jin, B.; Albuquerque, M.; Madhav, P. J. *J. Am. Chem. Soc.* **1997**, *119*, 10509.
35. Senese, C. L.; Hopfinger, A. J. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 2180.
36. De Oliveira, D. B.; Gaudio, A. C. *Act. Relat.* **2000**, *19*, 599.
37. Restrepo, G.; Basak, S. C.; Mills, D. *Curr. Comput-Aid. Drug.* **2011**, *7*, 109-121.
38. in: StatSoft (Ed.) STATISTICA version. 6.0, Tulsa, **2001**.
39. Marrero-Ponce, Y.; Castillo-Garit, J. A.; Torrens, F.; Romero-Zaldivar, V.; Castro, E. *Molecules* **2004**, *9*, 1100-1123.
40. Penney, K. B.; Smith, C. J.; Allen, J. C. *J. Invest. Dermatol,* **1984**, *82*, 308-310.
41. Castillo-Garit, J. A.; Martinez-Santiago, O.; Marrero-Ponce, Y.; Casañola-Martín, G.; Torrens, F. *Chem. Phys. Lett.*, **2008**, *464*, 107-112.