Quantitative Structure-Antibacterial Activity Relationship Modeling Using a Combination of Piecewise Linear Regression-Discriminant Analysis (I): Quantum Chemical, Topographic, and Topological Descriptors

ENRIQUE MOLINA,[1,2,3] ERNESTO ESTRADA,[3] DELVIN NODARSE,[1,2,4] LUIS A. TORRES,[5] HUMBERTO GONZÁLEZ,[6] EUGENIO URIARTE[6]

[1]*Faculty of Chemistry, University of Camagüey, Camagüey 74650, Cuba*

[2]*Chem-Bio-Informatic Group, University of Camagüey, Camagüey 74650, Cuba*

[3]*Complex Systems Research Group RIAIDT, Edificio CACTUS, University of Santiago de Compostela, Santiago de Compostela 15782, Spain*

[4]*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*

[5]*Faculty of Chemistry and Pharmacy, Central University of Las Villas, Santa Clara 54830, Villa Clara, Cuba*

[6]*Department of Organic Chemistry, Faculty of Pharmacy, University of Santiago de Compostela 15782, Spain*

Received 31 October 2007; accepted 3 March 2008
Published online 15 April 2008 in Wiley InterScience (www.interscience.wiley.com).
DOI 10.1002/qua.21702

ABSTRACT: Time-dependent antibacterial activity of 2-furylethylenes using quantum chemical, topographic, and topological indices is described as inhibition of respiration in *E. coli*. A QSAR strategy based on the combination of the linear piecewise regression and the discriminant analysis is used to predict the biological activity values of strong and moderates antibacterial furylethylenes. The breakpoint in the values of the biological activity was detected. The biological activities of the compounds are described by two linear regression equations. A discriminant analysis is carried out to classify the compounds

Correspondence to: D. Nodarse; e-mail: delvin.nodarse@reduc.edu.cu

Dedicated to Prof. Ricardo Grau-Ávalos (Las Villas University) for his many contributions to the research and education activities in applied statistics.

in one of the biological activity two groups. The results showed using different kind of descriptors were compared. In all cases the piecewise linear regression—discriminant analysis (PLR-DA) method produced significantly better QSAR models than the linear regression analysis. The QSAR models were validated using an external validation previously extracted from the original data. A prediction of reported antibacterial activity analysis was carried out showing dependence between the probability of a good classification and the experimental antibacterial activity. Statistical parameters showed the quality of quantum-chemical descriptors based models prediction in LDA having an accuracy of 0.9 and a *C* of 0.9. The best PLR-DA model explains more than 92% of the variance of experimental activity. Models with best prediction results were those based on quantum-chemical descriptors. An interpretation of quantum-chemical descriptors entered in models was carried out. © 2008 Wiley Periodicals, Inc. Int J Quantum Chem 108: 1856–1871, 2008

Key words: QSAR; LMR; piecewise regression; LDA; topological indices; topographic indices; quantum chemical descriptors; antibacterial activity

1. Introduction

In the development of quantitative structure-activity relationships (QSAR) models three main aspects are involved. They are the experimental values of the biological activity to be described; the molecular descriptors used to model the chemical structure and the statistical approaches engaged in the developing of the model [1]. The developer of the QSAR model does (in general) not control the first aspect of the generation of a QSAR model. The description of the molecular structure through the so-called molecular descriptors is a more difficult but necessary task. Difficulties arise in the generation of such indices, given by the proper non-mathematical nature of the chemical structure [2–4]. In spite of this, a very large series of molecular descriptors are at the disposition of the researcher [5–9]. Finally, there is an important pool of statistical approaches that can be used to generate the mathematical relationships between the quantitative measurements of the biological activity and the molecular descriptors [10–12].

There are many statistical approaches that can be used to generate the quantitative models. Among them the linear multiple regression (LMR)[13, 14] continues being one of the most used techniques in QSAR analysis, especially in the 2D QSAR [15]. New improvements of LMR in QSAR and QSPR analysis have recently appeared in the chemical literature [16–22]. This fact is not only concerned to tradition, which plays an important role in human activities, but is also due to the facilities in the interpretation of the results that these simple linear models give. There have been many advances in the use of chemometric analysis in QSAR studies [12]; however, the simplicity of the linear models is in several cases preferred in detriment of the statistical quality of the QSAR model. This is not a justifiable practice, but it is understandable due to the impossibility to interpret many QSAR models developed by using some nonlinear regression techniques [21–23]. It is well known that the interpretation of these models in terms of the mechanism of actions, metabolic or toxicological routes, etc., is one of the main advantages that this theoretical approach brings to the rational manipulation of chemicals [1].

What we propose here is the use of a linear piecewise regression analysis combined with discriminant analysis as an useful strategy in the search of QSAR models. This novel strategy presents some of the advantages that the LMR brings to QSAR analysis but showing significant improvements in the statistical quality of the models compared to the multilinear ones. It has been briefly used in scientific research. This article is conceived as follows. First, the method is introduced by using an illustrative hypothetical example. Then the method is applied to the development of QSAR models for the time-dependent antibacterial activity of 2-furylethylene compounds, which are compared to the linear regression models.

2. Methods

2.1. DATA SET

A data set of 2-furylethylenes was used in the present study. This data set was previously used by two of present authors [24] and by Marrero-Ponce [25]. The concentration [as log $(1/C)$] that produces the 50% of the inhibition of respiration in *Escherichia* *coli* at different times for these compounds is given [26]. The antibacterial activities of these 34 compounds, expressed in logarithmic form as log $(1/C)$, were reported at six different times forming a data set of 204 values of biological activities. However, 82 values of the antibacterial activity were reported as log $(1/C) < 3.00$, i.e., compounds having no antibacterial activity at the corresponding time. Consequently, the data set used in the present work was composed by 122 values of antibacterial activity measured for the 34 furylethylenes at six different times. In those previous articles [24, 25] were used values of antibacterial activity measured for this furylethylenes only at time 1 min and were classified compounds having log $(1/C) < 3.00$ (here excluded) as not actives. The values of the biological activity as well as the numbering of the compounds included in the data set used in the present study are given in Table I.

The resulting data set of 122 values of antibacterial activity of furylethylenes measured at different times was divided in two subsets. One of them used as the training set consisted of 92 measurements of log $(1/C)$ and the other 30 measurements selected at random were used as an external prediction set. Data points in the external prediction set are asterisked in the Table I. As can be seen, those 30 prediction measurements are the active concentration, at different times, of nine random selected chemical compounds.

Present work was carried out using atom and bond descriptors. In this sense emphasizing in the activity just discriminating between strong and moderates actives ones while in previous works [24, 25] authors discriminates between actives and not actives furylethylenes. We think that differences, concerning to the quality of previous models and those showed in this work, is not in the nature of the indices used but in the complexity of what is discriminated in each case (strong-moderates actives vs. actives-not actives).

2.2. MOLECULAR DESCRIPTORS

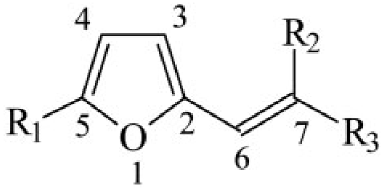
Three different kind of molecular descriptors will be used in the present study. They will be formally designed as topological [27–29], topographic [30–32], and quantum chemical [33]. The symbols of the molecular descriptors used throughout this work may be found in Table II. Calculations of molecular descriptors were performed by using the software MODEST 3.0 developed in our laboratory [34]. This computer program uses the results of the semiempirical quantum chemical calculations carried out by MOPAC 6.0 [35]. The calculations carried out here were performed by using the semiempirical AM1 method [36], using the following keywords for the MOPAC calculations: PRECISE, VECTORS, and BONDS. All descriptors entered (as variables) in a model were checked so avoiding two or more of them having intercorrelation over 80%. So there were never used descriptors codifying the same information (colinearity). From each pair of colinear variables the one having lower correlation with log $(1/C)$ and a more deficient direct physical interpretation was eliminated.

2.3. STATISTICAL ANALYSIS

Linear discriminant analysis (LDA), linear multiple regression (LMR), and the nonlinear estimation analysis, piecewise linear regression (PLR) were used to obtain quantitative models. These statistical analyses were carried out with the STATISTICA 4.3 software package [37]. Forward stepwise was fixed as the strategy for variable selection in the case of LDA and LMR analysis. 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. LDA is used in order to generate the classifier function on the basis of the simplicity of the method [38]. To test the quality of the discriminant functions derived we used the Wilks' λ and the Mahalanobis distance. The Wilks' λ-statistic for overall discrimination can take values in the range of 0 (perfect discrimination) to 1 (no discrimination). The Mahalanobis distance indicates the separation of the respective groups. It shows whether the model possesses an appropriate discriminatory power for differentiating between the two respective groups. The classification of cases was performed by means of the posterior classification probability, which is the probability that the respective case belongs to a particular group, i.e., compounds with moderate antibacterial activity or compounds with strong antibacterial activity. In developing this classification function, the values of 1 and -1 were assigned to compounds with moderate and strong antibacterial activity. The quality of the LDA-model was also determined by examining the percentage of good classification. Validation of the discriminant function was corroborated by the prediction of an external data set of nine chemicals (a total of 30 activity values) not included in the training set. In all cases for training and prediction data sets some others statistical indices were given, i.e., Speci-

TABLE I

Experimental values of the antibacterial activity at different times of exposition for the 2-furylethylene compounds studied.



			log(1/*C*)					
R_1	R_2	R_3	t = 1 min	t = 2 min	t = 4 min	t = 8 min	t = 16 min	t = 32 min
H	NO_2	$COOCH_3$	3.51 (1)	3.70 (15)	3.77 (29)	3.78 (46)	3.78 (66)	3.78 (91)
CH_3	NO_2	$COOCH_3$	4.00 (2)	4.11 (16)	4.20 (30)	4.44 (47)	4.44 (67)	4.44 (92)
Br	NO_2	$COOCH_3$	4.90 (3)*	5.00 (17)*	5.01 (31)*	5.01 (48)*	5.01 (68)*	5.01 (93)*
I	NO_2	$COOCH_3$	4.20 (4)	4.25 (18)	4.31 (32)	4.31 (49)	4.31 (69)	4.31 (94)
$COOCH_3$	NO_2	$COOCH_3$	3.95 (5)	3.95 (19)	3.95 (33)	3.95 (50)	3.95 (70)	3.95 (95)
NO_2	NO_2	$COOCH_3$	4.10 (6)	4.11 (20)	4.11 (34)	4.11 (51)	4.11 (71)	4.11 (96)
NO_2	$COOC_2H_5$	$COOC_2H_5$	3.00 (7)	3.21 (21)	3.43 (35)	3.78 (52)	4.10 (72)	4.31 (97)
NO_2	H	NO_2	4.90 (8)	4.92 (22)	4.95 (36)	4.95 (53)	4.95 (73)	4.95 (98)
H	H	NO_2	3.60 (9)	3.91 (23)	4.21 (37)	4.45 (54)	4.66 (74)	4.83 (99)
NO_2	H	$CONH_2$	3.72 (10)	4.02 (24)	4.20 (38)	4.60 (55)	4.88 (75)	5.21 (100)
NO_2	H	$CONHCH_3$	3.00 (11)	3.31 (25)	3.62 (39)	3.90 (56)	4.19 (76)	4.52 (101)
NO_2	H	$CON(CH_3)_2$	3.21 (12)	3.57 (26)	4.02 (40)	4.37 (57)	4.45 (77)	4.90 (102)
NO_2	H	$CONHC_2H_5$	—	—	—	3.17 (58)*	3.45 (78)*	3.81 (103)*
NO_2	H	$CONH(CH_2)_2CH_3$	—	—	—	—	—	3.00 (104)
NO_2	H	$CONHCH(CH_3)_2$	—	—	—	—	3.00 (79)*	3.21 (105)*
NO_2	H	$CONH(CH_2)_3CH_3$	—	—	—	—	—	3.03 (106)
NO_2	H	$CONHCH_2CH(CH_3)_2$	—	—	—	—	—	3.01 (107)
NO_2	H	$CONHCH(CH_3)C_2H_5$	—	—	—	—	—	3.01 (108)
NO_2	H	$CONHC(CH_3)_3$	—	—	—	—	—	3.00 (109)*
NO_2	H	$CONHCH_2C(CH_3)_3$	—	—	—	—	—	3.15 (110)
NO_2	H	$COOCH_3$	—	—	—	3.02 (59)*	3.25 (80)*	3.55 (111)*
NO_2	H	$COOC_2H_5$	—	—	—	—	3.00 (81)	3.15 (112)
NO_2	H	$COO(CH_2)_2CH_3$	—	—	—	—	3.10 (82)	3.41 (113)
NO_2	H	$COOCH(CH_3)_2$	—	—	—	—	3.00 (83)	3.35 (114)
NO_2	H	$COO(CH_2)_3CH_3$	—	—	—	—	3.20 (84)*	3.41 (115)*
NO_2	H	$COOCH_2CH(CH_3)_2$	—	—	3.10 (41)	3.45 (60)	3.79 (85)	4.00 (116)
NO_2	H	$COOCH(CH_3)C_2H_5$	—	—	3.01 (42)*	3.20 (61)*	3.41 (86)*	3.80 (117)*
NO_2	H	$COOC(CH_3)_3$	—	—	—	3.15 (62)*	3.55 (87)*	3.65 (118)*
NO_2	H	$COO(CH_2)_4CH_3$	3.00 (13)	3.31 (27)	3.65 (43)	3.87 (63)	4.27 (88)	4.55 (119)
NO_2	H	CN	—	—	3.00 (44)	3.25 (64)	3.70 (89)	3.91 (120)
NO_2	H	H	—	—	—	—	—	3.15 (121)
NO_2	CN	$COOCH_3$	3.70 (14)*	3.95 (28)*	4.25 (45)*	4.49 (65)*	4.70 (90)*	4.95 (122)*

The numbering of compounds used in the present work is given in parenthesis and compounds in the external prediction set are asterisked. The enumeration of atoms used in the calculation of quantum chemical descriptors is given together with the molecular graphic of the furylethylene framework.

ficity as True Negative rate, Sensitivity as True Positive Rate and Accuracy as precision.

A simple linear and other more complex nonlinear model was obtaining using LMR and PLR as statistic techniques, respectively. The quality of the models was determined examining the statistic parameters of multivariable comparison of regression. In this sense, the quality of models was determined by examining

TABLE II
Symbols for topological, topographic, and quantum chemical descriptors and their definitions.

${}^h\chi_p$	Path connectivity index of order $h = 0$–6
${}^h\chi_c$	Cluster connectivity index of order $h = 3$–6
${}^h\chi_{pc}$	Path-cluster connectivity index of order $h = 4$–6
${}^h\chi_p^v$	Valence path connectivity index of order $h = 0$–6
${}^h\chi_c^v$	Valence cluster connectivity index of order $h = 3$–6
${}^h\chi_{pc}^v$	Valence path-cluster connectivity index of order $h = 4$–6
${}^h\varepsilon_p$	Path bond connectivity index of order $h = 1$–6
${}^h\varepsilon_c$	Cluster bond connectivity index of order $h = 3$–6
${}^h\varepsilon_{pc}$	Path-cluster bond connectivity index of order $h = 4$–6
${}^h\Omega_p$	Path bond-order-based topographic connectivity index of order $h = 0$–6
${}^h\Omega_c$	Cluster bond-order-based topographic connectivity index of order $h = 3$–6
${}^h\Omega_{pc}$	Path-cluster bond-order-based topographic connectivity index of order $h = 4$–6
${}^h\Omega_p$ (q)	Path charge-based topographic connectivity index of order $h = 0$–6
${}^h\Omega_c$ (q)	Cluster charge-based topographic connectivity index of order $h = 3$–6
${}^h\Omega_{pc}$ (q)	Path-cluster charge-based topographic connectivity index of order $h = 4$–6
${}^h\Omega_p^c$ (q)	Path hydrogen-corrected charge-based topographic connectivity index of order $h = 0$–6
${}^h\Omega_c^c$ (q)	Cluster hydrogen-corrected charge-based topographic connectivity index of order $h = 3$–6
${}^h\Omega_{pc}^c$ (q)	Path-cluster hydrogen-corrected charge-based topographic connectivity index of order $h = 4$–6
Q_i	Electronic charge on atom i of the furylethylene framework
ES_σ (A_i)	σ electrophilic superdeslocalizability on atom A_i
ES_π (A_i)	π electrophilic superdeslocalizability on atom A_i
ES_T (A_i)	Total electrophilic superdeslocalizability on atom A_i
NS_σ (A_i)	σ nucleophilic superdeslocalizability on atom A_i
NS_π (A_i)	π nucleophilic superdeslocalizability on atom A_i
NS_T (A_i)	Total nucleophilic superdeslocalizability on atom A_i
E_{HOMO}	Highest occupied molecular orbital energy
E_{LUMO}	Lowest unoccupied molecular orbital energy
HB1	Hydrogen bonding potential of molecule

the regression coefficients (R), determination coefficients (R^2), standard deviations of the regression (s), and the leaveone-out (LOO) press statistic (Scv) [39]. To assess the robustness and predictive power of the found models, external prediction (test) sets were also used. This type of model validation is very important, if we take into consideration that the predictive ability of a QSAR model can only be estimated using an external test set of compounds that was not used for building the model [39]. Validations of this model were carried out using the same external prediction data set used validating the LDA models.

3. Results and Discussion

3.1. PIECEWISE LINEAR REGRESSION-DISCRIMINANT ANALYSIS

One of the most widespread and used approaches in QSAR studies is the extra-thermodynamic scheme of Hansch et al. [40–42]. To illustrate the theoretical basis of the piecewise linear regression—discriminant analysis (PLR-DA) strategy in QSAR we will use a hypothetical example based on this approach. This example consists of a data set of 14 substituted aromatic compounds for which a hypothetical biological activity has been determined. The list of the substituents in such compounds and their biological activities are given in Table III together with the values of the electronic and hydrophobic substituent constants [43].

The best linear QSAR model obtained to describe this activity by using the substituent constants given in Table I is given below:

$$\mathrm{Act}=1.796-1.043\sigma_p + 0.788\pi$$

$$R = 0.8654 \quad s = 0.390 \quad F = 16.4$$

where R is the correlation coefficient, s is the standard deviation of the regression, and F is the Fisher ratio. The use of quadratic models to describe this

TABLE III

Values of electronic and hydrophobic constants for the different substituents in the hypothetical aromatic compounds, as well as their hypothetical biological activities.

Substituent	σ_p	π	Act
NH_2	−0.57	−1.23	1.07
OH	−0.38	−0.67	1.35
OCH_3	−0.28	−0.02	2.15
CH_3	−0.14	0.50	2.35
F	0.15	0.13	2.33
Cl	0.24	0.76	2.42
Br	0.26	0.94	2.48
I	0.28	1.15	2.62
COOH	0.44	−0.28	1.15
CF_3	0.53	1.07	1.18
CN	0.70	−0.57	0.93
NO_2	0.81	−0.28	0.53
$COOCH_3$	0.44	−0.01	1.09
$COCH_3$	0.47	−0.55	0.98

activity with the use of these two substituent constants does not produce a significant improvement in the statistical quality of the model. If we plot the biological activity of these 14 compounds as a function of the two substituent constants, we obtain the three-dimensional graphic given in Figure 1.

As can be seen in this figure, there is a bilinear behavior of the biological activity when it is expressed in terms of the two descriptors used. There is a point in the graphic that we have designed as Act_0, which represents a breakpoint in the linear behavior of the biological activity. Below this point, the biological activity increases when the values of the Hammett σ_p constant decreases. However, over this breakpoint the activity is increased when the Hammett constant also increases.

It is obvious that the determination of this breakpoint cannot be carried out by a simple inspection of the plot of the biological activity versus the molecular descriptors. This is especially true when many molecular descriptors, i.e., more than two, are used in the QSAR model. Because of this reason, it is necessary to use a mathematical procedure to determine, for a specific data set, the exact value of the breakpoint. Consequently, we propose to utilize an efficient optimization method for the determination of the breakpoint. Here and in subsequent studies, we will use the quasi-Newton optimization algorithm implemented in STATISTICA [37] for the determination of this point in the hyperspace of variables that describe the biological activity under study. The mathematical principles of the quasi-Newton algorithms, the computational implementation as well as its applications to chemical problems have been described in details in the literature [44, 45]. By these reasons, we will be not extended here in explaining the method.

After the application of the optimization algorithm to the data set of 14 aromatic compounds under study, we obtain the value of the biological activity that corresponds to the breakpoint for this specific problem, $Act_0 = 1.6293$. This value is used then to obtain two linear models describing the biological activity in terms of the substituent constants of the form:

$$Act(< Act_0) = 1.272 - 0.509\sigma_p + 0.189\pi \quad (1)$$

$$Act(> Act_0) = 2.232 + 0.204\sigma_p + 0.246\pi \quad (2)$$

The correlation coefficient of the plot between experimental and calculated values of the biological

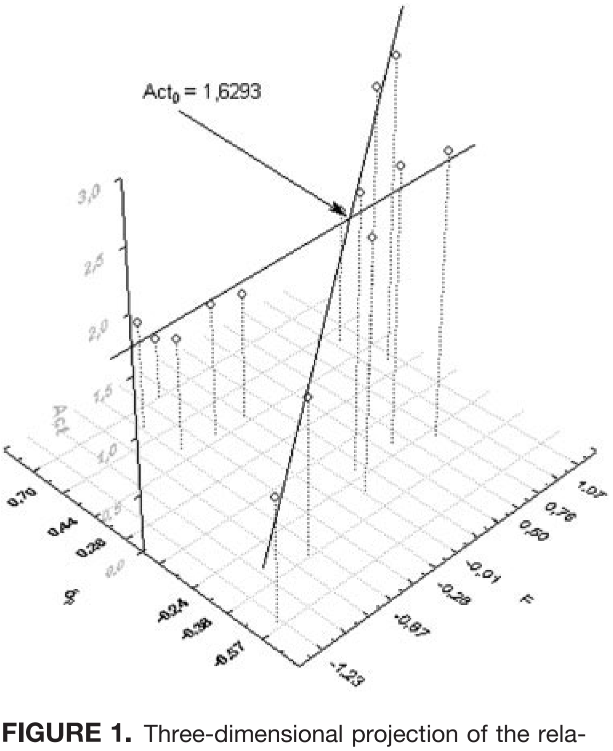


FIGURE 1. Three-dimensional projection of the relationships between the biological activity and the substituent constants for the hypothetical aromatic compounds. The linear piecewise nature of this relationship can be observed.

activity is 0.9898 and the standard deviation of this regression is only 0.106. By this way, the use of the piecewise linear regression (PLR) produced an improvement of 72.8% in the standard deviation respect to the linear QSAR model. This PLR-QSAR model explains almost 98% of the variance in the experimental values of the biological activity versus less than 75% explained by the linear regression model.

At this point, only one problem needs to be solved in order to demonstrate the usefulness of the PLR approach in QSAR studies. This point is concerned to the possibility of predicting the biological activity for compounds not included in the training series used in the development of the PLR-QSAR model. The problem arises because the breakpoint is a value of the experimental property, i.e., a value of the biological activity, and this value needs to be known in order to apply one of the two equations developed by the PLR approach. To solve this problem, we propose the use of the linear discriminant analysis (LDA)[14, 46, 47] to obtain a mathematical function that permits the classification of chemicals in two classes: those having biological activities under the breakpoint and those having activities greater than the value of the breakpoint. In the hypothetical example given here, it is very easy to determine by simple inspection that compounds having biological activities greater than the breakpoint have values of the σ_p constant which are in the following range: $-\ 0.28 \leq \sigma_p \leq 0.28$. On the other hand, compounds having activities under the breakpoint have values of the σ_p constant in the range: $-\ 0.28 > \sigma_p > 0.28$. By this way, it is not difficult to realize that substituent such as $N(CH_3)_2$ ($\sigma_p = -0.63$) and SO_2CH_3 ($\sigma_p = 0.73$) will have biological activities under the breakpoint and their activities should be estimated from model (1), while the activities of substituent such as $C(CH_3)_3$ ($\sigma_p = -0.15$) and Ph ($\sigma_p = -0.01$) should be estimated from model (2). This classification permits an easy interpretation of the results in terms of the structural features that influence the biological activity as well as from the point of view of action mechanism that are occurring in the biological media. For instance, in this example we can observe that two different mechanisms are involved in this hypothetical activity depending on the nature of the substituents. Those substituents that are activating of the electrophilic aromatic substitution (EAS)[48] have values of the activity greater than the breakpoint with the only exceptions of NH_2 and OH, two substituents with a great capacity to form hydrogen bonds. On the other hand, deactivating substituents of the EAS always have biological activities under the breakpoint, i.e., they are less active than the activating ones.

This combination of piecewise linear regression and discriminant analysis (PLR-DA) is the strategy that we propose here in order to solve some QSAR problems in which the linear regression approach fails [21–23]. This approach will be used now in a real QSAR problem related to the time-dependent antibacterial activity of 2-furylethylene derivatives.

3.2 ANTIBACTERIAL ACTIVITY OF 2-FURYLETHYLENE DERIVATIVES

3.2.1. Linear Regression QSAR Models

The development of linear regression models has a twofold objective. First, they serve as a way to select the independent variables that will be used in the development of the PLR-LDA QSAR models and second, they will be used as a way to compare the quality of the models obtained with the novel QSAR approach to be used here. These multivariate linear regression (MLR) models to describe the log (1/*C*) were obtained with the forward stepwise method and using the different sets of molecular descriptors, topological, topographic, and quantum chemical [27–33], (see Table II for definitions) as well as the time of exposition, *t*, as independent variables.

The best linear regression model obtained with the use of the topological indices studied here is given below together with the statistical parameters of the regression.

$$\begin{aligned}\log(1/C) = 4.422 &+ 0.459\,{}^{1}\chi_{p}^{v} - 52.699\,{}^{5}\chi_{p}^{v} + 5.119\,{}^{6}\chi_{p}^{v}\\ &+ 0.022t + 26.444\,{}^{4}\chi_{p}^{v} - 34.042\,{}^{5}\chi_{pc}^{v} - 3.604\,{}^{3}\varepsilon_{c}\\ &+ 1.930\,{}^{5}\varepsilon_{p} + 53.402\,{}^{6}\chi_{pc}^{v} - 12.127\,{}^{4}\varepsilon_{p} - 8.889\,{}^{4}\chi_{pc}^{v}\\ &+ 4.292\,{}^{2}\varepsilon_{p} \qquad (3)\end{aligned}$$

$$R = 0.8612 \quad s = 0.3194 \quad s_{CV} = 0.555\ F(12{,}79) = 18.9$$

Here and henceforth, *R* is the correlation coefficient, *s* is the standard deviation of the regression, s_{CV} is the standard deviation of the cross-validation, that is the standard deviation obtained by predicting the log (1/*C*) for data points in the external prediction set, and *F* is the Fisher ratio. The correlation between observed and predicted values of log (1/*C*) for the prediction set fits in less than 50% (*R* = 0.6907) which illustrate the poor predictive quality of the linear regression QSAR model previously shown.

The results obtained by using the topographic molecular descriptors are given below:

$$\begin{aligned}\log(1/C) = 16.117 - 0.085\,{}^{1}\Omega_{\mathrm{p}} - 43.138\,{}^{5}\Omega_{\mathrm{p}} \\ + 8.962\,{}^{6}\varepsilon_{(\mathrm{p})\mathrm{p}} + 0.022t - 6.829\,{}^{1}\Omega^{\mathrm{c}}_{(\mathrm{q})\mathrm{p}} + 4.107\,{}^{1}\varepsilon_{(\mathrm{p})\mathrm{p}} \\ + 6.639\,{}^{4}\Omega_{\mathrm{p}} + 7.953\,{}^{5}\Omega^{\mathrm{c}}_{(\mathrm{q})\mathrm{p}} - 0.233\mathrm{HB1} + 33.701\,{}^{5}\Omega_{(\mathrm{q})\mathrm{p}} \\ + 3.024\,{}^{6}\Omega^{\mathrm{c}}_{(\mathrm{q})\mathrm{pc}} + 2.453\,{}^{3}\Omega^{\mathrm{c}}_{(\mathrm{q})\mathrm{p}} \qquad (4)\end{aligned}$$

$$R = 0.8954 \quad s = 0.280 \quad s_{\mathrm{CV}} = 0.502\ F(12{,}79) = 26.6$$

In this case, a slight improvement is observed in the statistical parameters of the regression model. However, while the model explains 80% of the variance in the experimental values of log (1/*C*) in the training set, it explains only 57% of the variance in the prediction set. The standard deviation of the prediction set is almost twice the value of it for the training set, which indicates the poor predictive ability of this model too.

The final linear regression model obtained here is that relating the biological activity to the quantum chemical descriptors. This model is given below:

$$\begin{aligned}\log(1/C) = 79.620 - 7.250Q_7 + 51702.955\mathrm{ES_T}A_7 \\ - 38.306Q_6 + 1102.855\mathrm{ES}_{\sigma}A_5 + 0.012t \\ - 2.817\mathrm{NS}_{\pi}A_5 + 3.146\mathrm{NS}_{\pi}A_6 - 51816.510\mathrm{ES}_{\pi}A_7 \\ + 13.977Q_5 - 51642.544\mathrm{ES}_{\sigma}A_7 - 6.076\mathrm{NS}_{\pi}A_7 \\ + 0.104\mathrm{HB1} \qquad (5)\end{aligned}$$

$$R = 0.7718 \quad s = 0.400 \quad c_{\mathrm{v}} = 0.536 \quad F(12{,}79) = 9.7$$

As can be seen from the statistical parameters of this regression, it is the worse of the three linear QSAR models developed here. This model explains only 60% of the variance in the experimental antibacterial activity for the training set and 61% of it in the prediction set, i.e., the observed versus predicted values of log (1/*C*) in the prediction set correlate with *R* = 0.7168. Curiously, this model produce better results for the prediction set than the model using topological indices in spite of the fact that the last model shows better statistical results for the training set. This is probably because the compounds in the prediction set have some particular structural features that are well described by these quantum chemical descriptors. However, we have maintained the same training and prediction sets in the development of the three models in order to permit the comparative analysis between them in an appropriate way.

In the following section, we will use the molecular descriptors included in the MLR models as the independent variables in the development of the PLR-DA QSAR models.

3.2.2. PLR-DA QSAR Models

The first step in the development of the PLR models is the selection of the variables that will be included in the models. At present, there are several efficient ways for the automated selection of molecular descriptors to be used in the development of QSAR models [49–55]. However, we will use the same sets of molecular descriptors selected in the precedent section by the MLR analysis. By using these descriptors, we will proceed to the search of the breakpoint as before explained. The use of the quasi-Newton algorithm [37, 44, 45] for the training data set of 92 points used in the present study detected a breakpoint of log $(1/C)_0$ = 3.92. The breakpoint found was the same independently of the pool of molecular descriptors used, i.e., topological, topographic, or quantum chemical. This breakpoint divides the antibacterial furylethylenes in two subclasses: moderate antibacterial compounds having values of log (1/*C*) ≤ log $(1/C)_0$ and strong antibacterial compounds if log (1/*C*) > log $(1/C)_0$ at a determined time of exposition.

Now we can obtain two linear regression models using the same variables which were selected from the MLR, one describing the activity of moderate antibacterial furylethylenes, which will be designed by log $(1/C)_<$, and the other describing that for strong antibacterials, designed by log $(1/C)_>$. Both quantitative models obtained with the use of topological indices are given below:

$$\begin{aligned}\log(1/C)_< = 2.582 + 3.742\,{}^{2}\varepsilon_{\mathrm{p}} - 7.934\,{}^{4}\varepsilon_{\mathrm{p}} + 0.558\,{}^{5}\varepsilon_{\mathrm{p}} \\ - 4.102\,{}^{3}\varepsilon_{\mathrm{c}} - 0.167\,{}^{1}\chi^{\mathrm{v}}_{\mathrm{p}} + 14.444\,{}^{4}\chi^{\mathrm{v}}_{\mathrm{p}} - 28.052\,{}^{5}\chi^{\mathrm{v}}_{\mathrm{p}} \\ - 1.099\,{}^{6}\chi^{\mathrm{v}}_{\mathrm{p}} - 6.247\,{}^{4}\chi^{\mathrm{v}}_{\mathrm{pc}} - 17.849\,{}^{5}\chi^{\mathrm{v}}_{\mathrm{pc}} + 30.761\,{}^{6}\chi^{\mathrm{v}}_{\mathrm{pc}} \\ + 0.013t \qquad (6)\end{aligned}$$

and

$$\log(1/C)_{>} = 10.934 - 6.827\,{}^{2}\varepsilon_{\mathrm{p}} + 5.469\,{}^{4}\varepsilon_{\mathrm{p}} + 8.028\,{}^{5}\varepsilon_{\mathrm{p}} + 7.293\,{}^{3}\varepsilon_{\mathrm{c}} + 0.493\,{}^{1}\chi^{\mathrm{v}}_{\mathrm{p}} + 7.698\,{}^{4}\chi^{\mathrm{v}}_{\mathrm{p}} - 7.648\,{}^{5}\chi^{\mathrm{v}}_{\mathrm{p}} - 3.807\,{}^{6}\chi^{\mathrm{v}}_{\mathrm{p}} + 9.999\,{}^{4}\chi^{\mathrm{v}}_{\mathrm{pc}} - 9.415\,{}^{5}\chi^{\mathrm{v}}_{\mathrm{pc}} - 1.092\,{}^{6}\chi^{\mathrm{v}}_{\mathrm{pc}} + 0.011t \quad (7)$$

$$R = 0.9340 \quad s = 0.210 \quad s_{\mathrm{CV}} = 0.305$$

Here, the statistical parameters correspond to the general regression obtained by the combination of both quantitative models, i.e., the statistical parameters were obtained by fitting the observed values of biological activity to those computed by using models (6) and (7).

To carry out the prediction of the antibacterial activity of compounds in the external prediction set, it is necessary to find a discriminant function able to classify these compounds in one of the two groups of activities defined by the breakpoint: moderate and strong antibacterial compounds. This step is necessary because we do not know the values of the biological activities of compounds external to the training series, and consequently we do not know what of the two linear models, model (6) or model (7), should be applied to predict their quantitative activities. By this way we employ a LDA approach [14, 46, 47] to find this classification function by using the pool of topological indices under study; here it is not necessary, of course, to use the same variables selected from the MLR analysis. The classification function found with the use of these descriptors is given below:

$$\mathrm{Class} = -22.5 - 60.07\,{}^{5}\chi^{\mathrm{v}}_{\mathrm{pc}} - 22.48\,{}^{1}\chi^{\mathrm{v}}_{\mathrm{p}} + 205.17\,{}^{6}\chi^{\mathrm{v}}_{\mathrm{p}} - 72.48\,{}^{5}\chi^{\mathrm{v}}_{\mathrm{p}} + 28.34\,{}^{2}\chi_{\mathrm{p}} + 107.73\,{}^{6}\varepsilon_{\mathrm{p}} - 42.11\,{}^{3}\chi^{\mathrm{v}}_{\mathrm{c}} + 0.11t - 18.95\,{}^{5}\varepsilon_{\mathrm{pc}} - 117.02\,{}^{6}\chi_{\mathrm{p}} + 170.22\,{}^{6}\chi^{\mathrm{v}}_{\mathrm{pc}} - 11.05\,{}^{4}\varepsilon_{\mathrm{p}} \quad (8)$$

For the development of this discriminant function, compounds with moderate antibacterial activity were designed as −1 and the strong ones as 1. By this way, if the application of the classification function to any compound gives a value of Class < 0 it will be classified as moderate antibacterial, and if Class > 0, the compound is classified as strong antibacterial. This classification criterion will also be use in the development of the forthcoming discriminant functions.

This quantitative model classifies correctly 85.7% (36/42) of the moderate antibacterial compounds and 86% (43/50) of the compounds in the group of strongly active 2-furylethylenes. The total percentage of good classification obtained for this model is 85.9% (79/92). Other statistical parameters of model (8) are the Wilks' $\lambda = 0.40$, and the squared Mahalanobis distance between the group centroids, which is $D^2 = 5.97$. The Wilks' λ-statistic for the overall discrimination is computed as the ratio of the determinant of the within-groups variance/covariance matrix over the determinant of the total variance covariance matrix. We recall that Wilks' λ can takes values in the range of 0 (perfect discrimination) to 1 (no discrimination) [14, 37, 46, 47]. On the other hand, great values of the Mahalanobis distance indicate that the respective groups are significantly apart from each other and consequently the model posses an appropriate discriminatory power for differentiating between the respective two groups [14, 37, 46, 47].

Now, we use discriminant function (8) to classify the compounds in the external prediction set. Then, we use models (6) and (7) to predict the antibacterial activities of the compounds which were classified as moderate or strong antibacterials, respectively.

Previous works [24, 25] used LDA classifying this data set of furylethylenes as actives or inactives but having a different data conception so we cannot compare them with our results here. We used only compounds that those authors classified as actives but they used its reported activity at time 1 min and we took into consideration six different exposition times.

Even if statistical parameters of LDA reported here are not the best, in this work is shown that developed models described the experimental behavior. This can be observed in an additional external prediction data set of four compounds described in Section 3.2.3. Just as wrote Menger, "Models are to be used, not believed" [56].

The PLR model explains more than 87% of the variance in the antibacterial action of the compounds in the training series, which can be compared with the 74% explained by the linear regression model (3) developed by using the same variables. However, the most important question related to the novel PLR model is that concerned to its predictive power compared to that of the MLR model. The PLR-DA QSAR model explains 84% ($R = 0.9180$) of the variance in the external prediction set compared to only 48% explained by the linear regression model (3). The standard deviation of the prediction set by using the PLR-DA approach is reduced almost to a half of that obtained by using MLR approach. In Figure 2, we illustrate the plot of

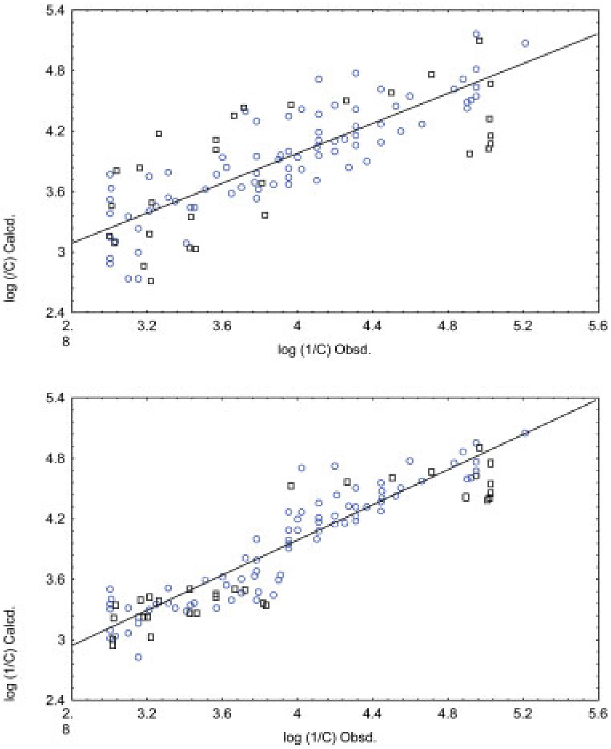


FIGURE 2. Graphical comparison of the results obtained by LMR and PLR-DA QSAR models in describing the antibacterial activity of 2-furylethylenes with topological indices for the training (○) and prediction (□) sets. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

the observed versus the predicted antibacterial activity for compounds in both training and external prediction sets. In this figure, we can observe the contrast existing between the results obtained with the use of the MLR analysis and those obtained by the PLR-DA method.

By following a procedure similar to that previously used with topological indices, we obtained the PLR-DA model with topographic molecular descriptors. The PLR equations are given below together with the statistical parameters:

$$\begin{aligned}\log(1/C)_{<} &= 12.223 + 3.141\,{}^{1}\varepsilon_{(\rho)\mathrm{p}} + 5.801\,{}^{6}\varepsilon_{(\rho)\mathrm{p}} \\ &\quad - 0.908\,{}^{1}\Omega_{\mathrm{p}} + 4.291\,{}^{4}\Omega_{\mathrm{p}} - 6.077\,{}^{5}\Omega_{\mathrm{p}} - 3.751\,{}^{5}\Omega_{(\mathrm{q})\mathrm{p}} \\ &\quad - 4.525\,{}^{1}\Omega^{\mathrm{c}}_{(\mathrm{q})\mathrm{p}} + 1.551\,{}^{3}\Omega^{\mathrm{c}}_{(\mathrm{q})\mathrm{p}} + 6.307\,{}^{5}\Omega^{\mathrm{c}}_{(\mathrm{q})\mathrm{p}} \\ &\quad + 2.594\,{}^{6}\Omega^{\mathrm{c}}_{(\mathrm{q})\mathrm{pc}} - 0.179\mathrm{HB1} + 0.016t \qquad (9)\end{aligned}$$

and

$$\begin{aligned}\log(1/C)_{>} &= 9.780 + 2.123\,{}^{1}\varepsilon_{(\rho)\mathrm{p}} + 6.644\,{}^{6}\varepsilon_{(\rho)\mathrm{p}} \\ &\quad - 1.601\,{}^{1}\Omega_{\mathrm{p}} + 3.989\,{}^{4}\Omega_{\mathrm{p}} - 0.668\,{}^{5}\Omega_{\mathrm{p}} + 0.731\,{}^{5}\Omega_{(\mathrm{q})\mathrm{p}} \\ &\quad - 2.329\,{}^{1}\Omega^{\mathrm{c}}_{(\mathrm{q})\mathrm{p}} + 0.150\,{}^{3}\Omega^{\mathrm{c}}_{(\mathrm{q})\mathrm{p}} + 2.128\,{}^{5}\Omega^{\mathrm{c}}_{(\mathrm{q})\mathrm{p}} \\ &\quad + 0.045\,{}^{6}\Omega^{\mathrm{c}}_{(\mathrm{q})\mathrm{pc}} - 0.132\mathrm{HB1} + 0.011t \qquad (10)\end{aligned}$$

$$R = 0.9501 \quad s = 0.184 \quad s_{\mathrm{CV}} = 0.292$$

The classification function obtained by using the LDA analysis is given below:

$$\begin{aligned}\mathrm{Class} &= 99.83 + 145.11\,{}^{5}\Omega^{\mathrm{c}}_{(\mathrm{q})\mathrm{c}} - 78.1\,{}^{1}\Omega^{\mathrm{c}}_{(\mathrm{q})\mathrm{p}} \\ &\quad + 66.81\,{}^{5}\varepsilon_{(\rho)\mathrm{p}} + 193.27\,{}^{3}\Omega^{\mathrm{c}}_{(\mathrm{q})\mathrm{p}} + 231.48\,{}^{6}\varepsilon_{(\rho)\mathrm{p}} \\ &\quad + 64.88\,{}^{4}\varepsilon_{(\rho)\mathrm{p}} + 0.16t + 88.06\,{}^{3}\Omega^{\mathrm{c}}_{(\mathrm{q})\mathrm{c}} + 0.9\,{}^{4}\varepsilon_{(\rho)\mathrm{pc}} \\ &\quad - 206.06\,{}^{6}\Omega^{\mathrm{c}}_{(\mathrm{q})\mathrm{p}} - 37.62\,{}^{6}\varepsilon_{(\rho)\mathrm{pc}} - 115.44\,{}^{3}\varepsilon_{(\rho)\mathrm{p}} \qquad (11)\end{aligned}$$

This model discriminates well 92.9% of data points in the group of moderate antibacterials (39/42) and 88.0% of data points in the other group (44/50), giving a total percentage of good classification of 90.2% (83/92), which is 5% greater than that reported by the previous model. This model also produces a better differentiation of the respective groups than that produced by the model using topological indices. For instance, the value of Wilks' λ is only 0.33 and the squared Mahalanobis distance is 8.05.

This PLR-DA QSAR model explains 90% of the variance of the biological activity in the training series, which represents 10% more than that explained by the MLR model (4). This model explains more than 85% ($R = 0.9250$) of the variance in the prediction set compared to that explained by the linear regression which is only 57% ($R = 0.7570$). The standard deviation for the prediction set is again almost a half of that obtained from model (4): 0.292 and 0.502, respectively. In Figure 3, we show a graphical comparison of the fits obtained between observed and calculated log (1/*C*) with the PLR-DA approach and those previously obtained from the MLR analysis.

The results obtained by using the PLR-DA approach in the developing of a QSAR model to describe antibacterial action of 2-furylethylenes employing quantum chemical molecular descriptors represent a very significant improvement compared to that reported with MLR analysis. The quantitative expressions of the PLR model together with the statistical parameters are given below:

$$\log(1/C)_{<} = 1.865 + 1.149\mathrm{ES}_{\sigma}A_{5} + 52.290\mathrm{ES}_{\sigma}A_{7}$$

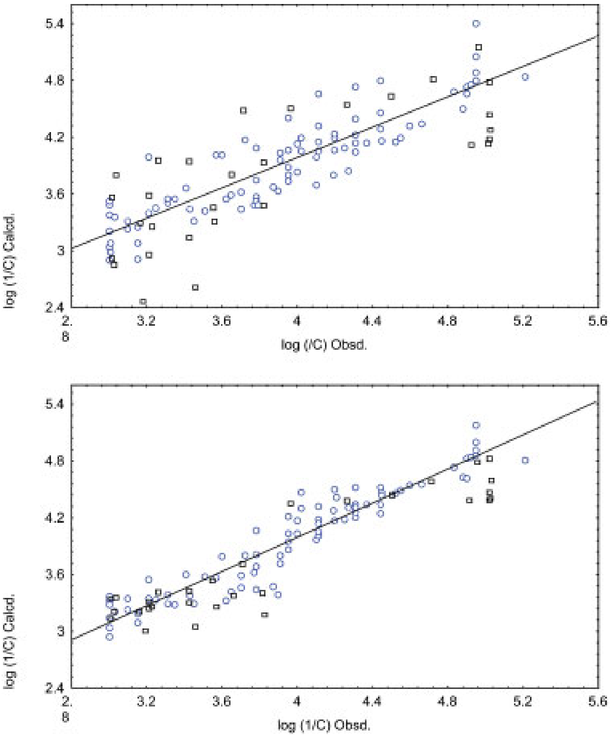


FIGURE 3. Graphical comparison of the results obtained by LMR and PLR-DA QSAR models in describing the antibacterial activity of 2-furylethylenes with topographic indices for the training (○) and prediction (□) sets. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

$$- 47.973\mathrm{ES}_{\pi}A_7 + 3.375\mathrm{ES}_{\mathrm{T}}A_7 + 0.915\mathrm{NS}_{\pi}A_5 + 1.340\mathrm{NS}_{\pi}A_6 + 0.894\mathrm{NS}_{\pi}A_7 - 3.761Q_5 - 21.371Q_6 - 2.416Q_7 + 0.125\mathrm{HB1} + 0.003t \qquad (12)$$

and

$$\log(1/C)_{>} = 3.103 + 154.260\mathrm{ES}_{\sigma}A_5 - 155.512\mathrm{ES}_{\sigma}A_7 + 211.276\mathrm{ES}_{\pi}A_7 + 54.850\mathrm{ES}_{\mathrm{T}}A_7 - 1.061\mathrm{NS}_{\pi}A_5 - 1.762\mathrm{NS}_{\pi}A_6 - 7.638\mathrm{NS}_{\pi}A_7 + 31.965Q_5 - 28.905Q_6 - 95.196Q_7 + 0.343\mathrm{HB1} + 0.011t \qquad (13)$$

$$R = 0.9260 \quad s = 0.222 \quad s_{\mathrm{CV}} = 0.227$$

The classification function obtained from the pool of quantum chemical descriptors is the following:

$$\mathrm{Class} = -191.88 - 1169.12\mathrm{ES}_{\pi}A_6 + 44.61\mathrm{NS}_{\pi}A_3 - 93.57Q_7 + 0.09t + 6.74\mathrm{NS}_{\mathrm{T}}A_7 + 115.53Q_5 - 26.4\mathrm{ES}_{\mathrm{T}}A_5 - 3670.4\mathrm{ES}_{\mathrm{T}}A_3 + 4818.81\mathrm{ES}_{\sigma}A_5 + 125.11\mathrm{NS}_{\sigma}A_2 - 155.88\mathrm{NS}_{\pi}A_7 - 8.42\mathrm{NS}_{\sigma}A_6 \qquad (14)$$

The discrimination model (14) classifies correctly 85.7% (36/42) of the points in group one, that is, moderately active furylethylenes, and 88% (44/50) of data points in group of strongly active compounds, classifying correctly 86.9% (80/92) of the total data points. The other statistical parameters of the discrimination model are as follows: Wilks' $\lambda = 0.40$ and $D^2 = 6.10$.

This PLR-DA QSAR model represents a significant improvement compared to the MLR models. For instance, this model explains 25% more of the variance in the biological activity of the training series than the linear regression model, and the standard deviation obtained from this model is 44.5% smaller than that obtained by the MLR. However, the most significant improvements are obtained in the predictive power of the model found. The PLR-DA model explains more than 92% of the variance of log $(1/C)$ in the external prediction set versus only 42% explained by the MLR, which can be considered as a nonpredictive model. The standard deviation of the prediction set is 61% smaller than that obtained for the MLR model. These improvements can be better seen in Figure 4, where observed versus predicted biological activities in training and prediction sets are plotted for both linear regression and PLR-DA approaches. It can be seen the great dispersion of data points obtained by linear regression, especially for prediction set, and the significant improvement reached for the same data sets when the PLR-DA method is employed.

A comparative table of others statistical parameters for reported LDA models using topologic, topographic and quantum-chemical descriptors is shown in Table IV. As can be clearly seen, the model showing best prediction results is that using quantum-chemical descriptors.

The values of the log $(1/C)$ that were calculated by MLR and PLR-DA models by using the three series of molecular descriptors used here are not reported in the present work. They may be obtained from us upon request.

3.2.3. Chemical Interpretation of Obtained Results

One of the essential topics of any SAR study is the interpretation of the role played by different

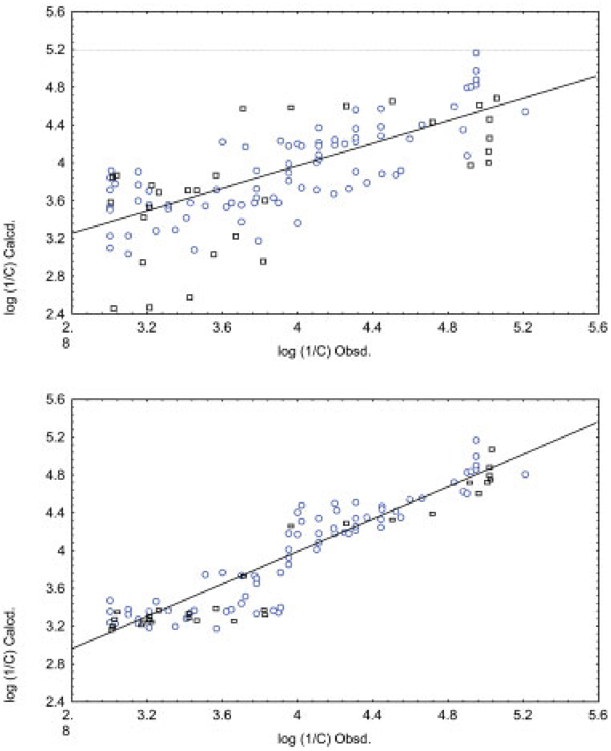


FIGURE 4. Graphical comparison of the results obtained by LMR and PLR-DA QSAR models in describing the antibacterial activity of 2-furylethylenes with quantum chemical indices for the training (○) and prediction (□) sets. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

substituents in the sense of the determination of biological activity for a chemical data. In this case the study of these relationships is not simple.

Is known the importance of *n*-octanol/water partition coefficient, log *P* (global property), and of the specific reaction rate constant for nucleophilic addition to the double bond, log *k* (local property) aiming understanding the biological behavior of this substances [57]. So, if is considered that the mentioned local property have not influence about the topological nature of involved molecules [57], then we can expect that the topological indices used in this work describe poorly (compared with the others two kinds of descriptors) the biological activity of the studied furylethylenes. This can be checked by the statistical parameters reported here.

On the other hand, local chemical reactivity has been well described using quantum-chemical descriptors. These descriptors give some light over the chemical nature of the interaction, modifying electronic aspects (bond orders and charge density) of these mentioned molecules.

For instance, molecular descriptor included in Eq. (14) clearly pointed toward the identification of the reaction centers involved in the studied interaction. Atoms 6 and 7 are involved in the exocyclic double bond of furylethylene (it refers to the 6–7 bond, which is directly involved in the nucleophilic attack) which is attacked by the thiol group. As a consequence, this QSAR model has a good statistical and chemical quality. It shown that thiol addition to exocyclic double bonds, enhanced by electro-acceptors groups, run by α carbon. It is supposed that average contribution of variables $ES_{ó}$ (6) and $ES_{ð}$ (6) to log *k* should be the greatest ones.

Indice ES_{π} (6) is the one that show the best correlation with the biological activity of these compounds ($R = 0.701$). This can be explained considering the important contribution of the electrophilic superdelocalizability of carbon 6 to the two mediums with different polarity partition and the reactivity in front of the thiol groups presents at enzymes. These two processes are much related with the antibacterial activity of these compounds.

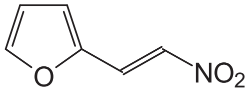
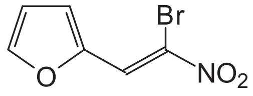
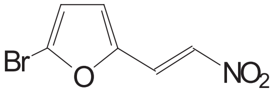
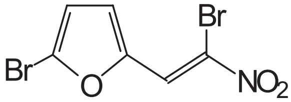
TABLE IV

Statistical parameters of obtained LDA models.

Descriptors	Data	FAR	*C*	Sensitivity	Specificity	Accuracy
Topological	Training	14.286	0.716	0.860	0.857	0.877
	Prediction	15.789	0.699	0.818	0.842	0.750
Topographic	Training	7.143	0.806	0.880	0.929	0.936
	Prediction	10.526	0.791	0.909	0.895	0.833
Quantum	Training	14.286	0.737	0.880	0.857	0.880
Chemical	Prediction	5.263	0.856	0.909	0.948	0.909

FAR, False Active Rate; C, Mathew's Correlation Coefficient.

TABLE V
Reported furylethylene derivatives antibacterial activity.

Compounds	MIC (μg/mL)[a]	Compounds	MIC (μg/mL)[a]
 I	50	 III	12.5
 II	25	 IV	3.125

I, 1-(fur-2-yl)-2-nitroethylene;
II, 1-(5-bromofur-2-yl)-2-nitroethylene;
III, 1-(fur-2-yl)-2-bromo-2-nitroethylene;
IV, 1-(5-bromofur-2-yl)-2-bromo-2-nitroethylene.
[a] *E. coli.*

The negative value of ES$_\pi$ (6) in Eq. (14) indicates a negative contribution to the drug power, which should be less active while ES$_\pi$ (6) arise higher values. However, it can be seen that electro-acceptors substituents, occupying different positions in the furylethylenic skeleton, produce a decrease of ES$_\pi$ (6) value. This bring into being a less negative contribution to log $(1/C)$. For example, the lowest values of ES$_\pi$ (6) are exposed by compounds III(3), IV(4), VI(6), and VIII(8), which presents two electro-acceptors groups (NO_2, Br, I) as substituents, when $t = 1$ min. Compounds III(3) and VIII(8) are two of the more actives in the data having a value of log $(1/C)$ equal to 4.90.

Is remarkable the lack of inclusion of E_{LUMO} or the contribution of α carbon to this orbital as variables in the obtained equation. After all this is a reaction of nucleophilic addition to the cited carbon in furylethylenes. They were both included in this study and declared having not significance. We think it is necessary express that the Klopmann-Salem equation for the chemical reactivity is ruled by a charge factor and an orbitalic factor [58]. In Eq. (14), charges are included (as $Q(7)$) and others belong to electrophilic (ES) and nucleophilic (NS) superdelocalizabilities. This superdelocalizabilities includes the p orbital coefficient to the atom A of the ith orbital ($C^{\pi}_{\mathrm{A}i}$) and also its energy (ε_i) so describing in a better way the orbitalic factor than E_{LUMO} along.

3.2.4. Prediction of Reported Furylethylene Derivatives Antibacterial Activity Analysis

We took into consideration results from previous published work reporting antibacterial activity of four furylethylene derivatives [57, 59] given in Table V. Compound I is included in the data set used here for development of all models reported in this study. The three remaining compounds, in this sense external ones, has been processed by the method proposed in the present work and the prediction of all four furylethylene derivatives antibacterial activity are given in Table VI.

The experimental value of the compound I (included in this study) antibacterial activity is 3.60 and prediction results (Table VI) show the PLR as the best prediction method reporting an activity of 3.64 for this molecule using quantum chemical descriptors. In a general way the three PLR models give better prediction results than MLR models. This support the validation of using the method proposed here (PLR-LDA) as an alternative to MLR.

As can be seen in Table V antibacterial activity value is increased from compound I to IV. Similar behavior is observed in the prediction analysis, no matters which molecular descriptor is employed, predicted activity value is increased from compound I to IV. The increment of LDA good results is also observed. The good classification percent probability by LDA models value is increased from

TABLE VI
Prediction of reported furylethylene derivatives antibacterial activity (time: 1 min) using models developed in this article.

Compound	Molecular descriptor	MLR	PLR	LDA
I	Topol.	3.00	3.54	73.82
	Topog.	3.05	3.51	72.51
	Q-C	3.15	3.64	80.64
II	Topol.	3.12	4.67	82.03
	Topog.	3.25	4.59	80.11
	Q-C	3.42	4.77	88.14
III	Topol.	3.27	4.79	84.55
	Topog.	3.40	4.68	82.33
	Q-C	3.63	4.90	91.38
IV	Topol.	3.61	5.21	92.43
	Topog.	3.94	5.09	90.18
	Q-C	4.10	5.53	97.11

compound I to IV. This clearly shows the dependence between the probability of a good classification and the experimental antibacterial activity.

4. Conclusions

In spite of the long continued use of the LDA in QSAR and of the implementation of piecewise regression techniques in some professional statistical packages, such as STATISTICA, there is an insignificant number of scientific publications using a combination of both methods to develop a predictive statistical tool for modeling. By this way, this work constitutes one of the first attempts of using this strategy not only in QSAR studies but also in mathematical modeling as a whole [see Ref. 60]. The present approach has significant differences to the bilinear regression, previously used in QSAR and chemometrics [61, 62]. The bilinear regression consists on the development of two linear models based on a breakpoint obtained for the independent variable. This approach is easy to apply when only one (or very few) independent variable(s) are used, but it is not applicable for the general multivariate case as the PLR-DA is.

The results presented in this study clearly indicate the applicability of the PLR-DA approach for the development of QSAR models in some cases in which MLR fails. This approach may be considered as a useful tool for the development of many different types of QSAR models. Here, we have illustrated the development of such models by using three different kind of molecular descriptors: topological, topographic, and quantum chemical. We have shown that for the three cases have been significant improvements compared to the linear regression models. However, this approach can also be applied when other types of molecular descriptors and QSAR approaches are used.

A comparison between the predictive capacity of all models obtained here are carried out using a four furylethylene derivatives series showing the better adjusted to experimental results given by the PLR models and the dependence between the probability of a good classification and the experimental antibacterial activity.

Another interesting point that should be recognized here is the possibility of combination of different series of molecular descriptors in the generation of the PLR-DA QSAR models. For instance, in the present study the best results from the PLR analysis are those obtained when quantum chemical molecular descriptors are employed. However, the best classification function is that developed from topographic indices. Consequently, significant improvements can be reached if both series of molecular descriptors are used; ones for the development of the PLR equations and the others for the generation of the discriminant function. On the other hand, the use of the strategies for the automatic selection of molecular descriptors previously mentioned [49–55] appears to be an attractive option in the future development of PLR-DA QSAR models.

The PLR-DA analysis has wide perspectives in QSAR studies not only because it can be applied to many different nonlinear situations but also because it can be modified to produce more powerful chemometrics techniques. For instance, the use of principal components piecewise linear regression (PCPLR) combined to discriminant analysis is one of these strategies to the development of novel QSAR models.

References

1. Franke, R. Theoretical Drug Design Methods; Elsevier: Amsterdam, 1984; p 1.
2. Randic, M. J Chem Educ 1992, 69, 713.
3. Randic, M. J Math Chem 1990, 4, 157.
4. Randic, M. J Chem Inf Comput Sci 1997, 37, 672.
5. Stuper, A. J.; Brugger, W. E.; Jurs, P. C. Computer-Assisted Studies of Chemical Structure and Biological Function; Wiley-Interscience: New York, 1979.

6. Cartier, A.; Rivail, J. L. Chemometrics Intell Lab Syst 1987, 1, 335.
7. Van de Waterbeemd, H.; El Tayar, N.; Carrupt, P. A.; Testa, B. J Comput-Aided Mol Des 1989, 3, 111.
8. Katritzky, A. R.; Lobanov, V. S.; Karelson, M. Chem Soc Rev 1995, 279.
9. Basak. S. C.; Grunwald, G. D.; Niemi, G. J. In From Topology to Three-Dimensional Geometry; Balaban, A. T., Ed.; Plenum Press: New York, 1997; p 73.
10. Sharaf, M. A.; Illman, D. L.; Kowalski, B. R. Chemometrics; Wiley: New York, 1986.
11. Haswell, S. J., Ed. Practical Guide to Chemometrics; Marcel Dekker: New York, 1992.
12. Van de Waterbeemd, H., Ed. Chemometric Methods in Molecular Design (Methods and Principles in Medicinal Chemistry, Vol. II); VCH: Weinheim, 1995.
13. Draper, N.; Smith, H. Applied Regression Analysis, 2nd ed.; Wiley: New York, 1981.
14. Thorndike, R. M. Correlational Procedures for Research; Gardner Press: New York, 1978.
15. Loew, G. H.; Villar, H. O.; Alkorta, I. Pharm Res 1993, 10, 475.
16. Ponce, Y. M.; Perez, M. A. C.; Zaldivar, V. R.; Sanz, M. B.; Mota, D. S.; Torrens, F. Internet Electrón J Mol Des 2005, 4, 124.
17. Ponce, Y. M.; Castillo-Garit, J. A.; Nodarse, D. Bioorg Med Chem 2005, 13, 3397.
18. Ponce, Y. M.; Nodarse, D.; Gonzalez-Diaz, H.; Ramos de Armas, R.; Zaldivar, V. R.; Torrens, F.; Castro E. A. Int J Mol Sci 2004, 5, 276.
19. Ponce, Y. M.; Perez, M. A. C.; Zaldivar, V. R.; Gonzalez-Diaz, H.; Torrens, F. J Pharm Pharma Sci 2004, 7, 186.
20. Pogliani, L. Med Chem Res 1997, 7, 380.
21. Marrero-Ponce, Y.; Medina-Marrero, R.; Castillo-Garit, J. A.; Romero-Zaldivar, V.; Torrens, F.; Castro, E. A. Bioorg Med Chem 2005, 13, 3003.
22. Marrero-Ponce, Y.; Medina-Marrero, R.; Romero-Zaldivar, V.; Castro, E. A.; Torrens, F.; Gonzalez-Diaz, H.; Ramos de Armas, R. Molecules 2004, 9, 1124.
23. Gonzalez-Diaz, H.; Ramos de Armas, R.; Uriarte, E. Online J Bioinformatics 2002, 1, 83.
24. Estrada, E.; Molina, E. J Chem Inf Comput Sci 2001, 41, 791.
25. Marrero-Ponce, Y. Bioorg Med Chem 2004, 12, 6351.
26. Balaz, S.; Sturdik, E.; Rosenberg, M.; Augustin, J.; Skara, B. J Theor Biol 1988, 131, 115.
27. Kier, L. B.; Hall, L. H. Molecular Connectivity in Structure-Activity Analysis; Wiley: New York, 1976.
28. Estrada, E. J Chem Inf Comput Sci 1995, 35, 31.
29. Estrada, E.; Guevara, N.; Gutman, I. J Chem Inf Comput Sci 1998, 38, 428.
30. Estrada, E.; Montero, L. A. Mol Eng 1993, 2, 363.
31. Estrada, E. J Chem Inf Comput Sci 1995, 35, 708.
32. Estrada, E.; Ramírez, A. J Chem Inf Comput Sci 1995, 36, 837.
33. Kikuchi, O. Quant Struct-Act Relat 1987, 6, 179.
34. Rodríguez, L.; Estrada, E.; Gutierrez, Y.; Muñoz, I. MODEST 3.0 for Windows (MOlecular DESign Tools), Central University of Las Villas, Santa Clara, 1996.
35. Stewart, J. J. P. MOPAC 6.0, Quantum Chemistry Program Exchange, Indiana University, Bloomington, IN, Program 455.
36. Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewar, J. J. P. J Am Chem Soc 1985, 107, 3902.
37. STATISTICA 4.3, StatSoft Inc., 1993.
38. Mc Farland, J. W.; Gans, D. J. In Comprehensive Medicinal Chemistry; Hansch, C.; Sammes, P. G.; Taylor, J. B.; Ramdsen, C. A., Eds.; Pergamon: New York, 1990; p 667.
39. Wold, S.; Erikson, L. In Chemometric Methods in Molecular Design; van der Waterbeemd, H., Ed.; VCH: New York, 1995; p 309.
40. Gould, R. G., Ed. Biological Correlations-The Hansch Approach, Advances in Chemistry 114; American Chemical Society: Washington, DC, 1972.
41. Hansch, C.; Klein, T. E. Acc Chem Res 1986, 19, 392.
42. Fujita, T. In Comprehensive Medicinal Chemistry; Hansch, C.; Sammes, P. G.; Taylor, J. B.; Ramdsen, C. A., Eds.; Pergamon: New York, 1990; Vol. 4, p 497.
43. Hansch, C.; Leo, A. Substituent Constants for Correlation Analysis in Chemistry and Biology; Wiley: New York, 1979.
44. Fletcher, R. Practical Methods of Optimization, Vol. 1: Unconstrained Optimization; Wiley: New York, 1980.
45. Schlick, T. In Reviews in Computational Chemistry; Lipkowitz, B.; Boyd, B., Eds.; VCH: New York, 1991; Vol. 3, Chapter 1.
46. Collus, T. Discriminant Analysis and Applications; Academic Press: New York, 1973.
47. Johnson, R. A.; Wichern, D. W. Applied Multivariate Statistical Analysis; Prentice Hall: New Jersey, 1988.
48. Morrison, R. T.; Boyd, R. N. Organic Chemistry; Ed. Rev.: La Habana, 1970.
49. Kubinyi, H. Quant Struct-Act Relat 1994, 13, 285.
50. Baroni, M.; Constantino, G.; Cruciani, G.; Riganelli, D.; Valigi, R. Quant Struct-Act Relat 1994, 12, 9.
51. Lindgren, F.; Geladi, P.; Rannar, S.; Wold, S. J Chemom 1994, 8, 349.
52. Rogers, D.; Hopfinger, A. J. J Chem Inf Comput Sci 1994, 34, 854.
53. Sutter, J. M.; Dixon, S. L.; Jurs, P. C. J Chem Inf Comput Sci 1995, 35, 77.
54. Hasegawa, K.; Miyashita, Y.; Funatsu, K. J Chem Inf Comput Sci 1997, 37, 306.
55. Pajeva, I. K.; Wiese, M. Quant Struct-Act Relat 1997, 16, 1.
56. Menger, F. M. J Am Chem Soc 1985, 107, 3165.
57. Molina, E. Desarrollo de nuevos compuestos antitumorales y modelación de la actividad antimicrobiana a través del empleo de descriptores moleculares novedosos. Tesis de doctorado, 2002.
58. Salem, L. J Am Chem Soc 1968, 90, 3.
59. (a) Castañedo, N.; Goizueta, R.; Pérez, J.; González, J.; Silveira, E.; Cuesta, M.; Martínez, A.; Lugo, E.; Estrada, E.; Carta, A.; Navia, O.; Delgado, M. S. Cuban Pat. 22446 (1995); (b) Castañedo, N.; Goizueta, R.; Pérez, J.; González, J.; Silveira, E.; Cuesta, M.; Martínez, A.; Lugo, E.; Estrada, E.; Carta, A.; Navia, O.; Delgado, M. S. Eur. Pat. Appl. 95,500,056-7 (1995); (c) Castañedo, N.; Goizueta, R.; Pérez, J.; González, J.; Silveira, E.; Cuesta, M.; Martínez, A.; Lugo, E.; Estrada, E.; Carta, A.; Navia, O.; Delgado, M. S. Canadian Pat. Appl. 2,147,594, (1995); (d) Castañedo, N.; Goizueta, R.; Pérez, J.; González, J.; Silveira, E.; Cuesta, M.; Martínez, A.;

Lugo, E.; Estrada, E.; Carta, A.; Navia, O.; Delgado, M. S. Jpn. Pat. Appl. 222,002, (1995); (e) Castañedo, N.; Goizueta, R.; Pérez, J.; González, J.; Silveira, E.; Cuesta, M.; Martínez, A.; Lugo, E.; Estrada, E.; Carta, A.; Navia, O.; Delgado, M. S. U.S. Pat. Appl. 6,008,011 (1995).

60. (a) Brown, S. D. Anal Chem 1990, 62, 84R; (b) Brown, S. D.; Bear, R. S. Jr.; Blank, T. B. Anal Chem 1992, 64, 22R; (c) Brown, S. D.; Blank, T. B.; Sum, S. T.; Weyer, L. G. Anal Chem 1994, 66, 315R; (d) Brown, S. D.; Sum, S. T.; Despagne, F. Anal Chem 1996, 68, 21R; (e) Lavine, B. K. Anal Chem 1998, 70, 209R.

61. Franke, R. Monthly Weather Review 1985, 2, 260.

62. San Roman, E.; Gonzalez, M. C. J Phys Chem 1989, 93, 3532.