Journal of the Argentine Chemical Society

QSAR STUDY FOR THE FISH TOXICITY OF BENZENE DERIVATIVES

Pablo R. Duchowicz[1]♥, Juan J. Marrugo H.[2] Erlinda V. Ortiz[3], Eduardo A. Castro[1] and Ricardo Vivas-Reyes[2]

[1]Instituto de Investigaciones Fisicoquímicas Teóricas y Aplicadas INIFTA (UNLP, CCT La Plata-CONICET), Diag. 113 y 64, C.C. 16, Suc.4, (1900) La Plata, Argentina.

[2]Group of Quantum and Theoretical Chemistry, Department of Chemistry, University of Cartagena, Cartagena-Colombia.

[3]Facultad de Tecnología y Ciencias Aplicadas, Universidad Nacional de Catamarca, Av. Maximio Victoria 55, (4700), Catamarca, Argentina.

Received June 6, 2009. In final form September 9, 2009.

Abstract

We searched Quantitative Structure-Toxicity models for predicting the fish toxicity against *Poecilia reticulata* elicited by a diverse set of 92 benzene derivatives. The simultaneous linear regression analyzes on 1176 constitutional, topological, geometrical, electronic, and lipophilic molecular descriptors derived from the software Dragon lead to a three-parameter relationship characterized with correlation coefficient of calibration of R=0.953, Leave-one-out Cross Validation of R_{loo}=0.947, and test set validation of R_{val}=0.889, and compares fairly well with a previously reported model based on extended topo-chemical atom (ETA) indices. Our developed QSAR involves a topological

♥Corresponding author. E-mail: pabloducho@gmail.com Tel.: (+54)(221)425-7430 Fax: (+54)221-425-4642;

descriptor as the most relevant variable for the set of chemicals, a 3D-MoRSE and a Radial Distribution Function descriptor that show low inter-correlations.

Keywords: QSAR theory; replacement method; benzene derivatives; *poecilia reticulata;* p*C*

Resumen

Hemos estudiado algunos modelos de relaciones cuantitativas estructura-toxicidad para predecir la toxicidad de los peces hacia *Poecilia reticulata* obtenida a través de una serie de 92 derivados bencénicos. El análisis de regresión lineal simultáneo de 1176 descriptores moleculares constitucionales, topológicos, geométricos, electrónicos y lipofílicos provenientes del software Dragon conduce a una relación de tres parámetros caracterizada por un coeficiente de correlación de calibración de R = 0.953, un Leave-one-out Cross Validation de R_{loo}= 0.947 y un test de validación de R_{val}= 0.889 los cuales se comparan bastante bien con los modelos informados previamente basados en los índices extendidos átomo topo-químico (ETA). Nuestro desarrollo QSAR involucra un descriptor topológico como la variable más relevante para una serie de descriptores químicos, para descriptores 3D-MoRSE y para descriptores tipo Función Radial de Distribución los que muestran bajas intercorrelaciones.

Palabras clave: teoría QSAR; método de reemplazo; derivados del benceno; *poecilia reticulata*; p*C*

Introduction

The accurate estimation of adverse environmental impact is considered of great interest to the scientific community, together with a convenient way to regulate the production of toxic chemical compounds [1, 2]. It is well known that performing a toxicological experiment for a given substance is not an easy task as it usually results expensive, requires time and, furthermore, an analysis of such dimensions should consider multiple environments and all biological interactions with the living organisms of the ecosystems, data that quite often are not available [3].

It is known that whenever it is not possible to perform intensive biological tests over complex systems, the application of semi-empirical or theoretical methodologies proves to be an adequate alternative for obtaining information about the eco-toxicological features of a given compound. In recent years, a generally accepted strategy for overcoming the absence of experimental measurements in biological phenomena is the analysis based on Quantitative Structure-Activity Relationships (QSAR) [4]. The ultimate role of formulating the QSAR is to suggest mathematical models that estimate the toxicities by relying on the assumption that these relationships are determined solely by the molecular structures of the compounds.

The structure is therefore translated into the so-called molecular descriptors, describing some relevant feature of the compounds, with mathematical formulae obtained from the Chemical Graph Theory, Information Theory, Quantum Mechanics, etc. [5, 6]. There are more than a thousand available descriptors in the literature, and one has to decide how to select those that characterize in the best possible manner the property under consideration. An obvious advantage of this sort of studies is to avoid animal testing.

Present research deals with the QSAR prediction of fish toxicity values for the same data set of aromatic chemicals analyzed previously [7], for comparison purposes. We explore a greater pool of variables composed of 1176 structural descriptors including definitions of all classes, and resort to the widely applied Replacement Method (RM) approach for performing the optimal variable subset selection [8-11]. RM is an algorithm proposed by our theoretical group some years ago, that efficiently generates multivariable linear regression QSAR models with minimized standard deviation.

As a next step, and with the main purpose of improving the statistical performance of the linear regression results found in the test set of compounds (validation data), we present a novel optimization algorithm. This is very convenient, as the search for new mathematical algorithms

usually lead to mathematical relationships displaying a better fit of the training data but simultaneously to a decreased performance on the test set.

Materials and Methods

Data Set and Molecular Descriptors Calculation

The observed fish toxicities of the benzene derivatives expressed as 96h LC_{50} data for both *Pimephales promelas* and *Poecilia reticulata* (**p*C***) were taken from [7] and originally from ref. [12]. The training set analyzed (denoted as *train*) is composed of the first sixty compounds shown in Table 1, whereas molecules 61-80 constitute a test set (*val*) that is employed for verifying the predictive capability of the QSAR models and not for calculating their calibration parameters. In addition, an external test set consisting on compounds 81-92 (*ext*), which does not influence the model design, is used for further validating the models. The members of these three molecular set were selected in such a way to share similar structural characteristics of the compounds.

The structures of the compounds are firstly pre-optimized with the Molecular Mechanics Force Field (MM+) procedure [13] included in the Hyperchem 6.03 package [14], and the resulting geometries are further refined by means of the semiempirical method PM3 (Parametric Method-3) [15] using the Polak-Ribiere algorithm [16] and a gradient norm limit of 0.01 kcal.$Å^{-1}$. We derived D=1176 molecular descriptors using the software Dragon 5.0 [17], including descriptors such as Constitutional, Topological, Geometrical, Charge, GETAWAY (Geometry, Topology and Atoms-Weighted AssemblY), WHIM (Weighted Holistic Invariant Molecular descriptors), 3D-MoRSE (3D-Molecular Representation of Structure based on Electron diffraction), Molecular Walk Counts, BCUT descriptors, 2D-Autocorrelations, Aromaticity Indices, Randic Molecular Profiles, Radial Distribution Functions, Functional Groups, Atom-Centred Fragments, Empirical and Properties [18]. Furthermore, four Quantum-Chemical descriptors (such as molecular dipole moments, total energies, and Homo-Lumo energies, etc), not provided by the program Dragon, were added to the pool.

Table 1. Experimental **p*C*** values and predicted residuals by the best QSAR models found.

No.	Compound	Exp.	Eq. (3)
Training set			
1	phenol	3.45	0.12
2	2-methylphenol	3.77	0.24
3	4-methylphenol	3.74	0.20
4	2,4-dimethylphenol	3.86	0.10
5	3,4-dimethylphenol	3.92	0.12
6	2,3,6-trimethylphenol	4.21	0.09
7	4-Ethylphenol	4.07	0.19
8	4-propylphenol	4.09	-0.10
9	4-tert-butylphenol	4.46	0.06
10	2-tert-butyl-4-methylphenol	4.90	0.04
11	4-pentylphenol	5.12	0.17
12	4-tert-pentylphenol	4.81	-0.10
13	2-allylphenol	3.96	0.13

Table 1. *Cont.*

14	2-phenylphenol	4.76	-0.02
15	1-naphthol	4.50	0.23
16	4-chlorophenol	4.18	0.27
17	4-chloro-3-methylphenol	4.33	0.14
18	4-chloro-3,5-dimethylphenol	4.66	0.15
19	3-methoxyphenol	3.22	0.01
20	4-methoxyphenol	3.05	-0.10
21	4-phenoxyphenol	4.58	-0.49
22	quinoline	3.63	-0.40
23	chlorobenzene	3.77	-0.01
24	1,3-dichlorobenzene	4.28	-0.06
25	1,4-dichlorobenzene	4.56	0.23
26	1,2,3-trichlorobenzene	4.89	0.04
27	1,2,4-trichlorobenzene	4.83	-0.03
28	1,2,3,4-tetrachlorobenzene	5.35	-0.02
29	1,2,3,5-tetrachlorobenzene	5.43	0.05
30	3-chlorotoluene	3.84	-0.11
31	4-chlorotoluene	4.33	0.38
32	2,4-dichlorotoluene	4.54	-0.05
33	2,4,5-trichlorotoluene	5.06	-0.08
34	3,4,5-trichlorotoluene	4.60	-0.43
35	pentachlorotoluene	6.15	0.06
36	benzene	3.09	-0.11
37	toluene	3.13	-0.25
38	2-xylene	3.48	-0.15
39	4-xylene	3.48	-0.06
40	nitrobenzene	2.97	-0.59
41	2-nitrotoluene	3.59	-0.13
42	3-nitrotoluene	3.65	-0.07
43	2,3-dimethylnitrobenzene	4.39	0.37
44	3,4-dimethylnitrobenzene	4.21	0.07
45	2-chloronitrobenzene	3.72	-0.29
46	3-chloronitrobenzene	4.01	-0.07
47	4-chloronitrobenzene	4.42	0.32
48	2,3-dichloronitrobenzene	4.66	-0.03
49	2,5-dichloronitrobenzene	4.59	-0.09
50	3,5-dichloronitrobenzene	4.58	-0.04
51	2-chloro-6-nitrotiluene	4.52	0.19
52	4-chloro-2-nitrotoluene	4.44	0.27
53	aniline	2.91	-0.23
54	2-methylaniline	3.12	-0.20
55	3-methylaniline	3.47	0.27
56	N,N-dimethylaniline	3.33	-0.18
57	2-ethylaniline	3.21	-0.33
58	3-ethylaniline	3.65	0.15

Table 1. *Cont.*

59	4-ethylaniline	3.52	0.07
60	4-butylaniline	4.16	0.08
Test set *val*			
61	2,6-diisopropylaniline	4.06	-0.57
62	3-chloroaniline	3.98	0.27
63	4-chloroaniline	3.67	-0.05
64	2,5-dichloroaniline	4.99	0.76
65	3,4-dichloroaniline	4.39	0.12
66	3,5-dichloroaniline	4.62	0.33
67	2,3,4-trichloroaniline	5.15	0.39
68	2,3,6-trichloroaniline	4.73	-0.05
69	2,4,5-trichloroaniline	4.92	0.14
70	RRR-4-tetrafluoro-3-methylaniline	3.77	0.11
71	RRR-4-tetrafluoro-2-methylaniline	3.78	0.13
72	pentafluoroaniline	3.69	-0.16
73	2-nitroaniline	3.15	-0.21
74	3-nitroaniline	3.24	-0.09
75	4-nitroaniline	3.23	-0.18
76	2-chloro-4-nitroaniline	3.93	0.06
77	4-bromoaniline	3.56	-0.28
78	3-benzyloxyaniline	4.34	-0.24
79	4-hexyloxyaniline	4.78	0.02
80	4-ethoxy-2-nitroaniline	3.85	0.01
Test set *ext*			
81	3-methylphenol	3.48	-0.09
82	2,6-dimethylphenol	3.75	-0.11
83	4-butylphenol	4.47	-0.06
84	1,2-dichlorobenzene	4.40	0.11
85	1,3,5-trichlorobenzene	4.74	-0.15
86	1,2,4,5-tetrachlorobenzene	5.85	0.45
87	3-xylene	3.45	-0.11
88	4-nitrotoluene	3.67	0.00
89	2,4-dichloronitrobenzene	4.46	-0.20
90	4-methylaniline	3.72	0.61
91	2-chloroaniline	4.31	0.67
92	2,4-dichloroaniline	4.41	0.33

Model Search

In our calculations we employ the computer system Matlab 5.0 [19]. It is our purpose to search the set **D**, containing D descriptors, for an optimal subset **d** of $d<<D$ ones with minimum standard deviation S:

$$S = \frac{1}{(N-d-1)} \sum_{i=1}^{N} res_i^2 \qquad (1)$$

with N being the number of molecules in the training set, and res_i the residual for molecule i (difference between the experimental and predicted property **p**). More precisely, we want to obtain the global minimum of $S(\mathbf{d})$ where **d** is a point in a space of $D!/[d!(D-d)!]$ ones. A full search (FS) of optimal variables is impractical because it requires $D!/[d!(D-d)!]$ linear regressions. Some time ago we proposed the Replacement Method (RM) [8-11] that produces linear QSPR-QSAR models that are quite close the FS ones with much less computational work. This technique approaches the minimum of S by judiciously taking into account the relative errors of the coefficients of the least-squares model given by a set of d descriptors $\mathbf{d}=\{X_1,X_2,\ldots,X_d\}$. The RM gives models with better statistical parameters than the Forward Stepwise Regression procedure [20] and similar ones to the more elaborated Genetic Algorithms [21].

The Kubinyi function (*FIT*) [22, 23] is a statistical parameter that closely relates to the Fisher ratio (F), but avoids the main disadvantage of the latter that is too sensitive to changes in small d values and poorly sensitive to changes in large d values. The $FIT(\mathbf{d})$ criterion has a low sensitivity to changes in small d values and a substantially increasing sensitivity for large d values. The greater the *FIT* value the better the linear equation. It is given by the following equation, where $R(\mathbf{d})$ is the correlation coefficient:

$$FIT = \frac{R^2(N-d-1)}{(N+d^2)(1-R^2)} \qquad (2)$$

In present study, the optimal number of molecular descriptors (d_{opt}) to be included in the linear regression equation is deduced from two criteria: (i) the plot of *FIT* vs. d and (ii) the performance of the model on the test set. For case (i), as the Kubinyi function achieves a maximum value at $d_{\max}$, it is possible to calculate d_{opt} in the following way:

1. calculate $d_1 = \left[\frac{d_{\max}}{2}\right] + 1$, where $[x]$ denotes the integer part of x.

2. if the slope of *FIT* at d_1 is greater than at d_1+1, then $d_{opt} = d_1$, otherwise, $d_{opt} = d_1 + 1$.

Therefore, the d_{opt} value reflects a "breaking point" beyond which the *FIT* improvement can be considered negligible. In case that the predictive performance of the model on the test set *val* is better for a smaller value of d_{opt} than that given by criterion (i), then the smaller d_{opt} value is adopted.

Results and Discussion

We begin the QSAR analysis by applying the RM algorithm on the training set of 60 benzene derivatives in order to find out a suitable model size (d). Table 2 shows the best linear models found with 1-7 molecular descriptors, while the specific details of the numerical variables involved throughout the whole article are provided in Table 3. Figure 1 (a) shows that the *FIT* parameter improves with d up to a certain "breaking point", which according to the criterion mentioned in section 2.2 corresponds to the value d_{opt}=4. Despite of this, Table 2 demonstrates that the model exhibiting a better predictive power on the test set *val* (compounds 61-80) involves one descriptor less (d_{opt}=3), leading to the validation parameters R_{val}=0.889 and S_{val}=0.312. Therefore, the best linear QSAR found includes the following three molecular descriptors:

$$pC = 1.824(\pm 0.2) - 0.296(\pm 0.04) \cdot RDF020e + 0.939(\pm 0.05) \cdot X1v + 0.584(\pm 0.1) \cdot Mor21e$$

N=60, R=0.953, S=0.212, FIT=8.086, R_{loo}=0.947, S_{loo}=0.224, $p<10^{-4}$ (3)

N=20, R_{val}=0.889, S_{val}=0.312

Here, the absolute errors of the regression coefficients are given in parentheses, p is the significance of the model, and *loo* stands for the Leave-One-Out Cross Validation technique [24].

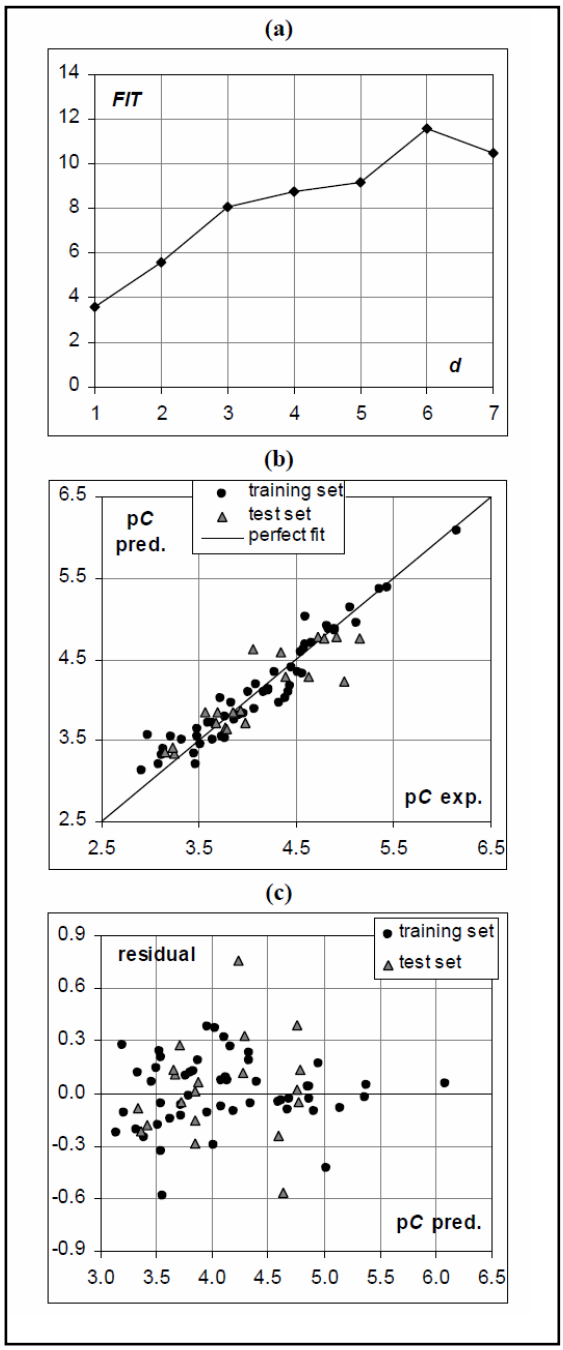

Figure 1. (a) *FIT* parameter versus number of descriptors (N=60).
(b) Predicted (Eq. 3) versus experimental **p*C*** (N=80). Circles: training set. Triangles: test set.
(c) Dispersion plot of the residuals for Eq. (3) (N=80). Circles: training set. Triangles: test set.

Table 2. Linear QSAR models established for the training set (N=60). The best relationship found is in bold.

Model	Descriptors involved	R	S	FIT	S_{loo}	R_{val}	S_{val}
M1	*MW*	0.889	0.316	3.567	0.324	0.600	0.686
M2	*MW, nN*	0.929	0.257	5.602	0.268	0.692	0.487
M3	***RDF020e, X1v, Mor21e***	**0.953**	**0.212**	**8.086**	**0.224**	**0.889**	**0.312**
M4	*RDF020e, DP02, ATS2p, BEHe2*	0.961	0.195	8.753	0.212	0.824	0.411
M5	*MAXDP, nCL, BEHe7, RDF020e, Sp*	0.967	0.181	9.196	0.201	0.896	0.363
M6	*MW, nN, nRORPh, Mor32u, Kp, L2u*	0.977	0.153	11.605	0.174	0.585	0.636
M7	*MW, nN, nRORPh, H6m, R2u$^+$, HATS0u, E2m*	0.978	0.152	10.486	0.176	0.645	0.643

Table 1 includes the predicted residuals via equation (3) for the training and test sets, while the plot of predicted vs. experimental toxicities shown in Figure 1 (b) suggests that the 60 training and 20 test set compounds of *val* follow a straight line. The behavior of the residuals in terms of the predictions of Figure 1 (c) leads to a normal distribution. This figure includes three calibration outliers with a residual exceeding the value 2S=0.424: compounds **21** (4-phenoxyphenol, -0.491), **34** (3,4,5-trichlorotoluene, -0.426), and **40** (nitrobenzene, -0.586), while no-one of the training compounds exceeds the value 3S=0.636; the presence of these outliers can be attributed to be a purely consequence of the limited number of optimal descriptors participating in equation (3).

The correlation matrix in Table 4 reveals that the descriptors of the linear model are not seriously inter-correlated (R<0.5), and thus substantiate the presence of all the variables in the model. The predictive power of the linear model is satisfactory as revealed by its stability upon the inclusion or exclusion of compounds, as measured by the *loo* parameters R_{loo}=0.889 and S_{loo}=0.312, and especially by means of the predictive ability in the test set *val* of R_{val}=0.889 and S_{val}=0.312. Present linear QSAR is of better quality and involves fewer descriptors (three) than the one reported previously [7] employing seven extended topochemical atom (ETA) indices that lead to R=0.941 and S=0.230. It has to be mentioned, however, that the reported model employed all the 92 **p*C*** values for training the model and did not employ a validation set for testing its predictive potential.

The three structural variables appearing in equation (3) can be classified as follows: (i) a topological descriptor: *X1v*, the valence connectivity index chi-1, (ii) a 3D-MoRSE descriptor: *Mor21e*, 3D-MoRSE–signal 21 / weighted by atomic Sanderson electronegativities; and (iii) a radial distribution function: *RDF020e*, Radial distribution function – 2.0 / weighted by atomic Sanderson electronegativities.

Table 3. Symbols for molecular descriptors involved in different models.

Molecular descriptor	Type	Description
RDF020e	Radial Distribution Function	Radial distribution function – 2.0 / weighted by atomic Sanderson electronegativities
X1v	Topological	valence connectivity index chi-1
Mor21e	3D-MoRSE	3D-MoRSE – signal 21 / weighted by atomic Sanderson electronegativities
RDF045m	Radial Distribution Function	Radial distribution function – 4.5 / weighted by atomic masses
Mor05p	3D-MoRSE	3D-MoRSE – signal 05 / weighted by atomic polarizabilities
L3v	WHIM	3rd component size directional WHIM index / weighted by atomic van der Waals volumes
MW	Constitutional	Molecular weight
nN	Constitutional	Number of nitrogen atoms
DP02	Randic Molecular Profiles	Molecular profile number 02
ATS2p	2D-Autocorrelations	Broto-Moreau autocorrelation of a topological structure – lag 2 / weighted by atomic polarizabilities
BEHe2	BCUT	Highest eigenvalue n. 2 of Burden matrix / weighted by atomic Sanderson electronegativities
MAXDP	Topological	Maximal electrotopological positive variation
nCL	Constitutional	Number of chlorine atoms
BEHe7	BCUT	Highest eigenvalue n. 7 of Burden matrix / weighted by atomic Sanderson electronegativities
Sp	Constitutional	Sum of atomic polarizabilities (scaled on carbon atom)
nRORPh	Functional Groups	Number of ethers (aromatic)
Mor32u	3D-MoRSE	3D-MoRSE – signal 32 / unweighted
Kp	WHIM	K global shape index / weighted by atomic polarizabilities
L2u	WHIM	2nd component size directional WHIM index / unweighted
H6m	GETAWAY	H autocorrelation of lag 6 / weighted by atomic masses
$R2u^+$	GETAWAY	R maximal autocorrelation of lag 2 / unweighted
HATS0u	GETAWAY	Leverage-weighted autocorrelation of lag 0 / unweighted
E2m	WHIM	2nd component accessibility directional WHIM index / weighted by atomic masses

The connectivity index *X1v* was proposed by Kier and Hall with the purpose of taking into account the nature of atoms symbolized by vertices [25]. This is readily calculated with a formula similar to that of Randic's molecular connectivity index, but considers products of valence delta values (δ_i) instead of vertex degrees:

$$X1v = \sum_{i,j} \delta_i \, \delta_j \qquad \delta_i = \frac{Z_i^v - H_i}{Z_i - Z_i^v - 1} \tag{5}$$

where Z_i^v indicates the number of valence electrons in atom *i*, Z is its atomic number, and H_i is the number of hydrogens attached to atom *i*. Thus, it is expected that index *X1v* reflects the molecular size. The 3D-MoRSE type of descriptor is obtained considering a molecular transform derived from an equation used in electron diffraction studies [26]. The electron diffraction does not directly yield atomic coordinates, but provides diffraction patterns from which the atomic coordinates are derived by mathematical transformations. These codes are defined in order to reflect the contribution to the property under investigation, at a prescribed scattering angle, of an atomic property such as mass (*m*), polarizability (*p*), electronegativity (*e*) or volume (*v*), and so enable to differentiate the nature of atoms. In example, for the case of *Mor21e*, the scattering angle is of 21 $Å^{-1}$ and the atomic Sanderson electronegativities are employed as weighting scheme. The Radial Distribution Function (RDF) [27] of an ensemble of atoms can be interpreted as the probability distribution of finding an atom in a spherical volume of certain radius, also incorporating different atomic properties in order to differentiate the contribution of atoms to the property being analyzed. For the case of *RDF020e*, the sphere radius is of 2.0 angstroms and the atomic Sanderson electronegativities are employed to distinguish their nature.

The standardization of the regression coefficients [20] in Equation (3) allows assigning a greater importance to the molecular descriptors that exhibit larger absolute standardized coefficients (shown in parentheses):

$$X1v\ (0.844)\ >\ RDF020e\ (0.346)\ >\ Mor21e\ (0.222) \tag{4}$$

and it is seen that the topological descriptor is the most relevant variable in present set of chemicals. It is mentioned that *X1v* and *RDF020e* descriptors take positive numerical values for all the compounds analyzed here, while *Mor21e* has negative values. Therefore, considering the sign of the regression coefficients a benzene derivative would tend to be more toxic (exhibiting a higher value of **p*C***) the higher the value of *X1v* index and the lower the values of *RDF020e* and *Mor21e* molecular descriptors in equation (3). Of course, compensating effects among the three variables would also lead to high toxicity of the compounds.

Table 4. Correlation matrix for descriptors of equation (3) (*N*=60).

	RDF020e	*X1v*	*Mor21e*
RDF020e	1	0.057	0.477
X1v		1	0.211
Mor21e			1

Conclusions

Present QSAR analysis established a linear regression model over sixty fish toxicity values exhibited by benzene derivatives, by means of three molecular descriptors that were rescued from a pool containing more than a thousand of variables through the Replacement Method variable subset selection procedure. The predictive performance of this model was assessed with two different test sets, one partially employed for guiding the model performance (*val*) and the other not employed at all during the training stage (*ext*), leading in both cases to satisfactory predictions of the toxicological behavior.

Acknowledgements. PRD and EAC are researchers from the Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET) of Argentina, RVR and JJM thanks the University of Cartagena and Colciencias for supporting this work.

References

[1] M. L. Hanson, K. R. Solomon, *Environ. Sci. Technol.,* **2002**, *36*, 3257.

[2] S. Smith, V. J. Furay, P. J. Layiwola, J. A. Menezes-Filho, *Chemosphere*, **1994**, *28*, 825.

[3] S. P. Bradbury, *Toxicol. Lett.,* **1995**, *79*, 229.

[4] C. Hansch, A. Leo, *Exploring QSAR. Fundamentals and Applications in Chemistry and Biology.* American Chemical Society, Washington D. C., 1995.

[5] A. R. Katritzky, V. S. Lobanov, M. Karelson, *Chem. Soc. Rev.*, **1995**, *24*, 279.

[6] N. Trinajstic, *Chemical Graph Theory*, CRC Press, Boca Raton (FL), 1992.

[7] K. Roy, G. Ghosh, *J. Chem. Inf. Model.*, **2004**, *44*, 559.

[8] P. R. Duchowicz, E. A. Castro, F. M. Fernández, M. P. González, *Chem. Phys. Lett.*, **2005**, *412*, 376.

[9] P. R. Duchowicz, E. A. Castro, F. M. Fernández, *MATCH Commun. Math.* Comput. *Chem.*, **2006**, *55*, 179.

[10] P. R. Duchowicz, M. Fernández, J. Caballero, E. A. Castro, F. M. Fernández, *Bioorg. Med. Chem.*, **2006**, *14*, 5876-5889.

[11] A. M. Helguera, P. R. Duchowicz, M. A. C. Pérez, E. A. Castro, M. N. D. S. Cordeiro, M. P. González, *Chemometr. Intell. Lab.,* **2006**, *81*, 180.

[12] K. Rose, L. H. Hall, *SAR&QSAR Environ. Res.*, **2003**, *14*, 113.

[13] A. K. Rappe, C. J. Casewit, *Molecular Mechanics Across Chemistry*. University Science Books, California, 1997.

[14] HYPERCHEM 6.03 (Hypercube) http://www.hyper.com.

[15] J. J. P. Stewart, *J. Comput. Chem.*, **1989**, *10*, 221.

[16] M. Krizek, P. Neittaanmäki, R. Glowinski, S. Korotov (Eds.), *Conjugate Gradient Algorithms and Finite Element Methods*. Springer-Verlag, Berlin, 2004.

[17] Dragon, Milano Chemometrics and QSAR Research Group, http://michem.disat.unimib.it/chm

[18] R. Todeschini, V. Consonni, *Molecular Descriptors for Chemoinformatics.* Wiley-VCH, Weinheim, 2009, Vol. 2.

[19] Matlab 7.0, The MathWorks Inc.

[20] N. R. Draper, H. Smith, *Applied Regression Analysis*. John Wiley&Sons, New York, 1981.

[21] S. S. So, M. Karplus, *J. Med. Chem.*, **1996**, *39*, 1521.

[22] H. Kubinyi, *Quant.- Struct.-Act. Relat.*, **1994**, *13*, 393.

[23] H. Kubinyi, *Quant. Struct.-Act. Relat.*, **1994**, *13*, 285.
[24] D. M. Hawkins, S. C. Basak, D. Mills, *J. Chem. Inf. Model.*, **2003**, *43*, 579.
[25] L. B. Kier, L. H. Hall, *J. Pharm. Sci.*, **1976**, *65*, 1806.
[26] J. Schuur, P. Selzer, J. Gasteiger, *J. Chem. Inf. Model.*, **1996**, *36*, 334.
[27] V. Consonni, R. Todeschini, M. Pavan, *J. Chem. Inf. Model*., **2002**, *42*, 693.