QSAR Study for Carcinogenicity in a Large Set of Organic Compounds

Pablo R. Duchowicz[*,1], Nieves C. Comelli[2], Erlinda V. Ortiz[3] and Eduardo A. Castro[1]

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

[2]*Facultad de Ciencias Agrarias, Universidad Nacional de Catamarca, Av. Belgrano y Maestro Quiroga, 4700 Catamarca, Argentina*

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

Abstract: In our continuing efforts to find out acceptable Absorption, Distribution, Metabolization, Elimination and Toxicity (ADMET) properties of organic compounds, we establish linear QSAR models for the carcinogenic potential prediction of 1464 compounds taken from the "Galvez data set", that include many marketed drugs. More than a thousand of geometry-independent molecular descriptors are simultaneously analyzed, obtained with the softwares E-Dragon and Recon. The variable subset selection method employed is the Replacement Method, and also the improved version Enhanced Replacement Method. The established models are properly validated through an external test set of compounds, and by means of the Leave-Group-Out Cross Validation method. In addition, we apply the Y-Randomization strategy and analyze the Applicability Domain of the developed model. Finally, we compare the results obtained in present study with the previous ones from the literature. The novelty of present work relies on the development of an alternative predictive structure-carcinogenicity relationship in a large heterogeneous set of organic compounds, by only using a reduced number of geometry independent molecular descriptors.

Keywords: QSAR theory, ADMET, multivariable linear regression analysis, carcinogenicity, molecular descriptors.

1. INTRODUCTION

A growing area of modern pharmaceutical research is the application of the Quantitative Structure-Activity Relationships (QSAR) Theory [1-5] to predict ADMET properties of chemical compounds, such as Absorption, Distribution, Metabolization, Elimination and Toxicity. Rational QSAR researches on ADMET properties enable to reduce the failure rate of drug development programs due to ADMET issues at late stages of the research (i.e. clinical trials) [6-8]. Therefore, allow moving towards finding a balance between potency, selectivity, bioavailability and safety from the very beginning of the project. This modern paradigm may be synthesized under the expression "to fail early is to fail cheap", and is implemented by including at early stages parallel ADMET filters to discard structures with unfavorable pharmacokinetic and toxicity profiles [9].

The underlying hypothesis of the different formulations of the QSAR Theory is that the chemical structure is the only influential factor on the exhibited activity of the collection of interacting molecules. This hypothesis has been extensively applied during past decades to the study of many different physicochemical and biological properties of interest [1-5]. In the realms of QSAR, the molecular structure is translated into the so-called molecular descriptors, which are used to describe different structural characteristics/attributes of the molecules in order to model the property being studied [10, 11]. A descriptor is defined as an empirical or a theoretically defined numerical quantity, and the most elementary type of descriptor is the count of atom types and chemical bonds.

Every QSAR study is established for a set of chemical structures known as the molecular training set, and does not have to be limited to work correlatively on the training set, but also has to behave predictively on new structures (never seen by the model) which take part of the test set. This validation process [12-15] is the most import step during the model design, and it is considered the basis of the QSAR hypothesis. When the validation step fails, then there is no way to use QSAR to create new chemical information for the biological activity based on the known one from the training set.

The carcinogenic activity exhibited by chemical substances is a toxicological endpoint of high health concern. There are available many QSAR studies developed during past years by different research groups, which involve a limited number of class-specific compounds, most of them nitro-containing derivatives [16-32]. On the other hand, Galvez has gathered the carcinogenic activity in the Discriminant Function (DF) scale (DFcarc) for a wide set of 1815 organic compounds extracted from the Merck index, based on the annual report of carcinogenesis [33]. From this data set, different molecular subsets have been taken to establish QSAR models [34, 35].

A recent study of Kar and Roy [36] employs for the first time a greater number of carcinogenic compounds, having

*Address correspondence to this author at the Instituto de Investigaciones Fisicoquímicas Teóricas y Aplicadas INIFTA (CCT La Plata-CONICET, UNLP), Diag. 113 y 64, C.C. 16, Sucursal 4, 1900 La Plata, Argentina;
Tel: (+54) (221) 4257430/(+54) (221) 4257291; Fax: (+54) (221) 4254642;
E-mail: pabloducho@gmail.com

1574-8863/12 $58.00+.00

1464 molecules from the Galvez data set involving many marketed drugs. These authors use 17 molecular descriptors having physical meaning, including topological indices, structural fragments, and hydrophobicity descriptors, and modeled the carcinogenicity by means of a predictive Partial Least Squares (PLS) model having 10 latent variables (732 molecules for the training set, 732 for the test set). They conclude that higher lipophilicity values and conjugated ring systems, thioketo and nitro groups contribute positively towards drug carcinogenicity. On the contrary, tertiary and secondary nitrogens, phenolic, enolic and carboxylic OH fragments and presence of three-membered rings reduce the carcinogenicity. Branching, size and shape are found to be crucial factors for drug-induced carcinogenicity.

In this work, we establish linear QSAR models on the same large set of 1464 organic compounds studied previously [36], with the purpose of improving the structure-carcinogenicity relationship found. For comparison purposes, we use the same training and test sets, and explore a pool containing more than a thousand of geometry-independent molecular descriptors. We decide to exclude the three dimensional aspects of the molecular structure, in order to avoid problems associated to numerical uncertainties. Such ambiguities result from an incorrect geometry optimization due to the existence of molecular structures in various conformational states. We consider that such kind of problem may lead to loosing of the predictive capability of the QSAR when applied for the prediction of the molecular test set. Therefore, the novelty of present work relies on the development of an alternative predictive structure-carcinogenicity relationship in a large heterogeneous set of organic compounds, by only using a reduced number of geometry independent molecular descriptors.

2. MATERIALS AND METHODS

2.1. Experimental Data Set

The carcinogenic activity of the organic compounds collected by Galvez in the DFcarc scale [33] is a discriminant function obtained by Linear Discriminant Analysis (LDA). The DF is performed in such a manner that values are positive for carcinogenic and negative for non-carcinogenic compounds. The DF values are not normalized and are in arbitrary units, but it is expected that when higher is the positive value, higher is the observed carcinogenicity, while the more negative value, the lower is the activity. The chemical domain analyzed involves hydrocarbons, aliphatic alcohols, phenols, ethers, and esters; anilines, amines, nitriles, nitroaromatics, amides, and carbamates; urea and thiourea derivatives, isothiocyanates, thiols, phosphate esters, and halogenated derivatives. The carcinogenicity values range in the interval [-9.91, 9.86], and the complete list of 1464 compounds studied here are included in Table **S1** as Supplementary Material.

2.2. Molecular Descriptors Calculation

The compounds are imported from Pubchem [37] in sdf format, with exception to few structures not available which are drawn with Hyperchem for Windows [38].

We compute 931 conformation-independent molecular descriptors using E-Dragon [39]. This well-known descriptor's database includes structural descriptors of thirteen different types, such as Constitutional, Topological, Walk and Path Count, Connectivity Index, Information Index, Edge Adjacency Index, Topological Charge Index, Burden Eigenvalues, Eigenvalue-Based Index, 2D-Autocorrelation, Functional Group Count, Atom-Centred Fragment, and Molecular Property [10].

In addition, atomic charge density-based descriptors are obtained, encoding electronic and structural information relevant to the chemistry of intermolecular interactions, by means of Recon 5.5 [40]. This sort of descriptors is not provided by E-Dragon, while the robustness of Recon has previously been demonstrated elsewhere [41, 42]. Recon is an algorithm for the reconstruction of molecular charge densities and charge density-based electronic properties of molecules, using atomic charge density fragments precomputed from ab initio wavefunctions. The method is based on the Quantum Theory of Atoms in Molecules [43]. A library of atomic charge density fragments has been built in a form that allows for the rapid retrieval of the fragments and molecular assembly. In present case, the smiles chemical notation is employed as input for the generation of 248 Transferable Atom Equivalent (TAE) descriptors, developed by Breneman *et al*. [44].

In this way, the total number of calculated molecular descriptors is 1179 variables.

2.3. Model Development

2.3.1. Molecular Descriptors Selection

In recent years theoretical and experimental researchers have focused an increasing attention on finding the most efficient tools for selecting molecular descriptors in QSAR studies. There is available a great number of feature selection methods are available to search for the best descriptors from a pool of variables, and the Replacement Method (RM) [45, 46], employed here, has been successfully applied elsewhere [47-51]. In brief, the RM is an efficient optimization tool which generates linear regression models on the training set by searching in a set having D descriptors for an optimal subset having $d<<D$ ones with smallest training set standard deviation (S). The quality of the results achieved with this technique approaches that obtained by performing an exact (combinatorial) full search of molecular descriptors although, of course, requires much less computational work. However, in some cases, the RM can get trapped in a local minimum of S. Although such local minima provide acceptable models, an improvement of the method has been developed in the Enhanced Replacement Method (ERM) [52, 53]. We use Matlab 7.0 in all our calculations [54].

2.3.2. Model Validation

The theoretical validation of the linear regression models is based on the popular validation criteria based on Cross Validation using Leave-One-Out (loo) and Leave-More-Out (ln%o, with n% being the percentile of molecules removed from the training set). The statistical parameters $R_{ln\%o}$ and

$S_{ln\%o}$ (correlation coefficient and standard deviation of Leave-More-Out) measure the stability of the QSAR upon inclusion/exclusion of molecules. The number of cases for random data removal analyzed in this study is 100000. According to the specialized literature, the loo explained variance (R^2_{loo}) should be greater than 0.5 for a validated model, although this is a necessary but not sufficient condition for its predictive power [55].

A more reliable validation is applied, that consists on using an external test set of structures. The same training set-test set partition from ref. [36] is used in present analysis, that is to say, the 1464 organic compounds are ranked according to their carcinogenicity values and every alternate compound is assigned to the test set. Each set thus includes 732 compounds.

We use Y-Randomization [56] as a way of checking that the model does not result from happenstance. This technique consists on scrambling the experimental property values in such a way that they do not correspond to the respective compounds. After analyzing 10000 cases of Y-Randomization, the smallest standard deviation value obtained using this procedure (S^{rand}) has to be a poorer value than the one found by considering the true calibration (S).

2.3.3. Applicability Domain

The applicability domain (AD) for the QSAR model is also explored, as not even a predictive QSAR model can be expected to reliably predict the modeled property for the entire universe of molecules. In fact, only the predictions for molecules falling within this AD can be considered reliable and not just model extrapolations. The AD is a theoretical region in the chemical space, and depends upon the molecular descriptor values and the experimental property analyzed [57]. The AD can be characterized in various ways, and one is the leverage approach [58], which allows to verify whether a new compound can be considered as interpolated (with reduced uncertainty, reliable prediction) or extrapolated outside the domain (unreliable prediction). The leverage for compound i (h_i) and its critical or warning value (h^*) are defined as follows:

$$h_i = x_i(\mathbf{X}^T\mathbf{X})^{-1}x_i^T \qquad h^* = 3(d+1)/N_{train} \tag{1}$$

where x_i is the descriptor vector for the compound, $\mathbf{X}$ is the model matrix for the training set, and N_{train} is the size of the training set. When $h_i > h^*$, then a warning should be given: for the training set, it means that the compound is highly influential in determining the model, while for the test set, it means that the prediction is the result of substantial extrapolation of the model and could not be treated as reliable.

2.3.4. Degree of Contribution of Selected Descriptors

In order to determine the relative importance of each descriptor in the linear regression model, we standardize the regression coefficients (b_j^s) as:

$$b_j^s = s_j b_j / s_Y \tag{2}$$

where b_j is the regression coefficient for the descriptor j-th descriptor, and s_j and s_Y are the standard deviations for such descriptor and for the experimental activity, respectively. The larger the absolute value of b_j^s, the greater the importance of the descriptor [59].

3. RESULTS AND DISCUSSION

We apply the RM and ERM variable subset selection methods on the training set and explore the pool containing D=1179 molecular descriptors. In this way, we try to identify the most representative structural features of the organic compounds that lead to their carcinogenicity behavior. Table **1** includes the best molecular descriptors found in present analysis, while a brief description for the meaning of each descriptor is given in Table **2**. It is clear from Table **1** that the use of six descriptors has an acceptable predictive power on the test set, and that the explained variance of the training set (R^2_{train}) does not improve so much beyond such a number of variables. We also follow the common practice of keeping the model's size as small as possible, in order to avoid any possible fortuitous correlation. Thus, we select the following structure-carcinogenicity relationship:

$$\begin{aligned}\mathrm{DF}_{\mathrm{carc}} = {} & -0.163(\pm0.01)Se + 3.079(\pm0.2)nR09 + \\ & 3.753(\pm0.3)DECC - 3.409(\pm0.1)IC2 + \\ & 1.198(\pm0.09)C - 003 + 5.784(\pm0.4)S - \\ & 108 + 12.506(\pm0.4)\end{aligned}$$

$$\begin{aligned}&N_{train} = 732,\ d = 6,\ N_{train}/d = 122,\ R^2_{train} = 0.74,\\ &S_{train} = 2.04,\ F = 345.83,\ R^2_{ij\max} = 0.66\\ &o(3S) = 6,\ R^2_{loo} = 0.76,\ S_{loo} = 2.08,\\ &S^{rand} = 3.93,\ R^2_{l20\%o} = 0.69,\ S_{l20\%o} = 2.23,\\ &N_{test} = 732,\ R^2_{test} = 0.77,\ S_{test} = 1.91,\end{aligned} \tag{3}$$

Here, F is the Fisher parameter, $R_{ij\max}$ denotes the maximum correlation coefficient between descriptors, $o(3S)$ indicates the number of outlier compounds having a residual (difference between experimental and calculated activity) greater than three times S.

The statistical quality achieved in the QSAR of Eq. 3 compares fairly well with the one found in the study of Kar and Roy [36], who used 17 molecular descriptors searched with the Stepwise Regression technique, leading to a 10 latent variables-PLS model. According to this model, the explained variance in the training set is 74.5% while for the test set is 73.1%; for Eq. 3 it results in 74% and 77% for the training and test sets, respectively.

The approval of the internal validation process of Eq. 3 is evidenced by the stability of this equation upon the inclusion/exclusion of compounds from the training set, measured *via* the exclusion of one molecule at a time (loo) and also by excluding 20% of the observations (147

Table 1. The Best Linear QSAR Models Obtained from a Pool of 1179 Geometry Independent Descriptors. The Selected Model Appears in Bold

d	R^2	S	R^2_{test}	S_{test}	R^2_{ijmax}	Descriptors
1	0.38	3.16	0.35	3.21	0.00	*IC2*
2	0.53	2.75	0.55	2.67	0.04	*nR09 IC2*
3	0.60	2.54	0.64	2.41	0.46	*nR09 ATS5e ALOGPS_logS*
4	0.65	2.37	0.70	2.22	0.46	*nR09 ATS5e S-108 ALOGPS_logS*
5	0.69	2.24	0.74	2.05	0.55	*nR09 IC1 ATS5v S-108 ALOGPS_logS*
6	**0.74**	**2.04**	**0.77**	**1.91**	**0.66**	***Se nR09 DECC IC2 C-003 S-108***
7	0.76	1.99	0.78	1.87	0.66	*Se nR09 DECC IC2 ATS4m C-003 S-108*
8	0.77	1.92	0.81	1.73	0.66	*Se nR09 DECC IC2 ATS5e C-003 S-108 ALOGPS_logS*
9	0.79	1.85	0.82	1.70	0.97	*nSK nR09 CSI IC2 GGI8 AEigm C-003 S-108 ALOGPS_logS*

Table 2. Brief Description of Molecular Descriptors Involved in Calculated QSAR Models

Constitutionals (0D)[a]
nR09 number of 9-membered rings
nSK number of non-H atoms
Se sum of atomic Sanderson electronegativities (scaled on Carbon atom)
Properties (1D)
ALOGPS_logS calculated aqueous solubility
Atom-Centred Fragments (1D)
S-108 number of R=S bonds
C-003 number of CHR3 groups
Information Indices (2D)
IC1 Information Content index (neighborhood symmetry of 1-order)
IC2 Information Content index (neighborhood symmetry of 2-order)
2D Autocorrelations (2D)
ATS5e Broto-Moreau autocorrelation of lag 5 (log function) weighted by Sanderson electronegativity
ATS5v Broto-Moreau autocorrelation of lag 5 (log function) weighted by van der Waals volume
ATS4m Broto-Moreau autocorrelation of lag 4 (log function) weighted by mass
GGI8 topological charge index of order 8
Topologicals (2D)
DECC Eccentric
CSI eccentric connectivity index
AEigm absolute eigenvalue sum from mass weighted distance matrix

[a]Descriptor's dimensionality.

molecules) in l20%o. Finally, as a further step to assess the robustness of present equation, we apply Y-randomization, demonstrating that the calibration does not result from happenstance and results in a valid structure-carcinogenicity relationship.

Fig. (**1a**, **b**) plot the predicted carcinogenicity as a function of the experimental values for the training and test sets, respectively, showing that there exists a tendency for the points to have a straight line trend. The dispersion plots of residuals (i.e. residuals as function of predicted activities) from Fig. (**2a**, **b**). reveal that the residuals tend to obey a random pattern around the zero line, indicating the absence of non-modeled factors. It is appreciated the presence of 6 outliers molecules, which are: Clopidogrel, Diathymosulfone, Ethylenediamine, Mirex, Octenidine, and Sulbenox. After checking their structures, we assume that this abnormal behavior may be purely attributed to the structural diversity of the present set containing 1464 structures.

The correlation matrix (Table **3**) demonstrates that the six descriptors from Eq. 3 are non-collinear and that each of them includes non-redundant structural information content

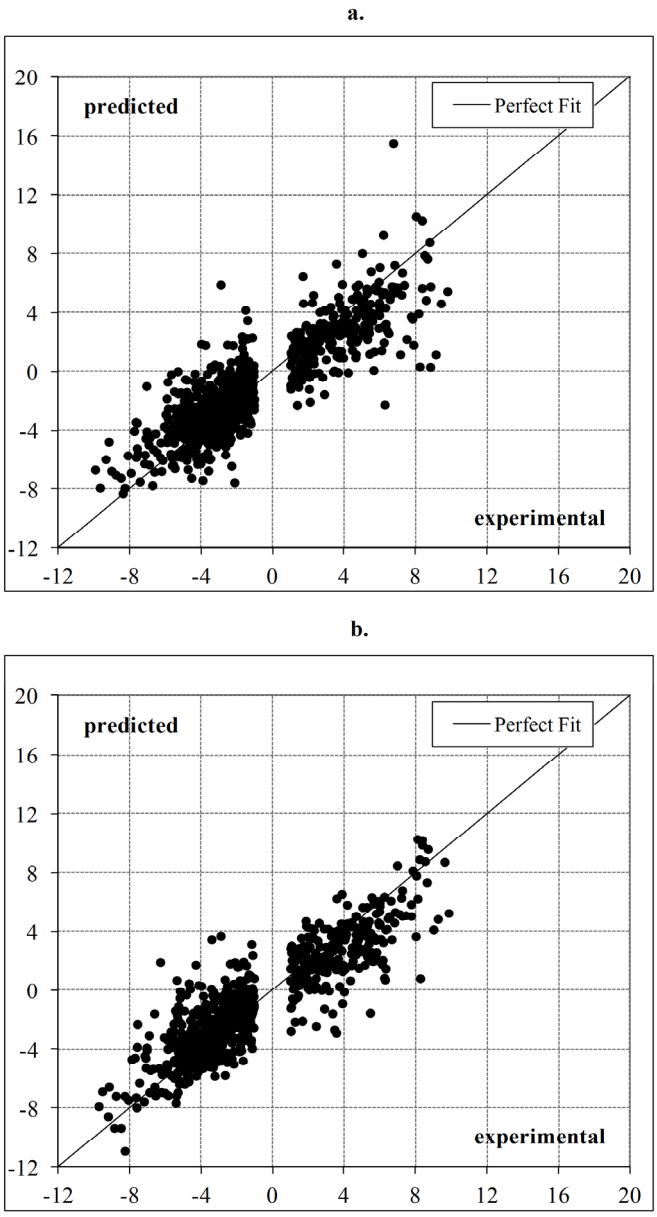

Fig. (1). Predicted and experimental carcinogenicity values for **a**. training set; **b**. test set.

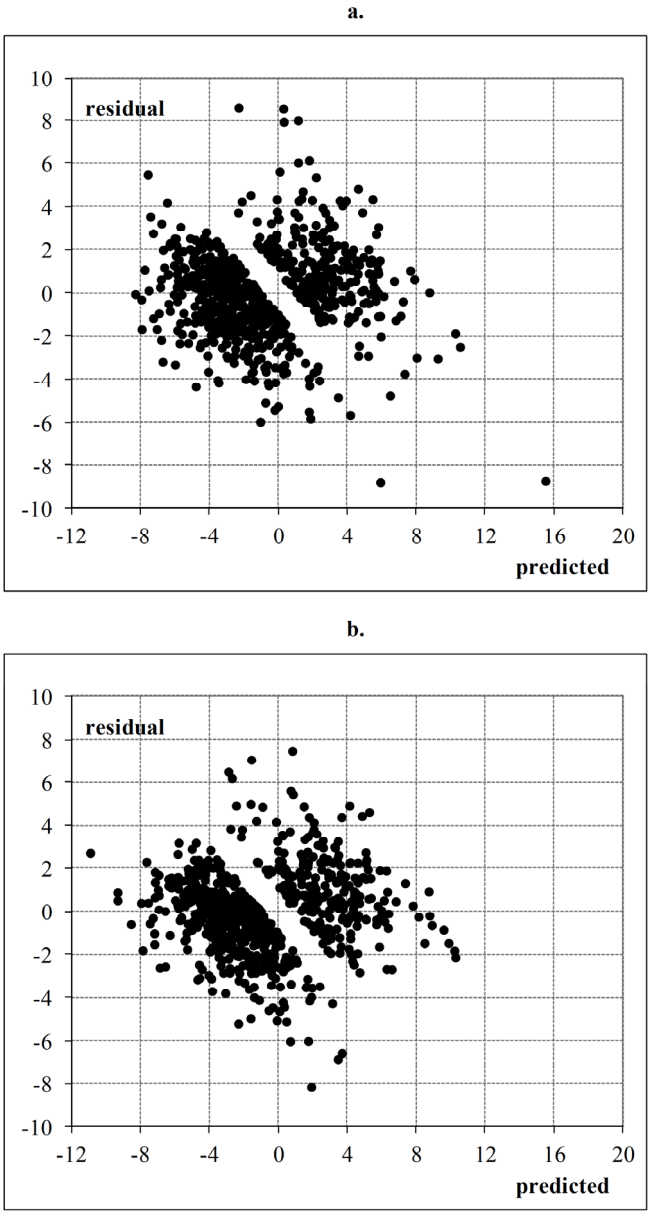

Fig. (2). Dispersion plot of residuals for **a**. training set; **b**. test set.

($R^2_{ij\max} = 0.66$). Among such descriptors, the Constitutionals (0D) are: *Se*, sum of atomic Sanderson electronegativities (scaled on Carbon atom) and *nR09*, number of 9-membered rings; the Atom-Centred Fragments (1D) are *C-003*, number of CHR3 groups and *S-108*, number of R=S bonds; the Information Index is: *IC2*, Information Content index (neighborhood symmetry of 2-order); and the Topological (2D) is: *DECC*, Eccentric [10]. The ranking of contributions of these descriptors reveals that *Se* and *IC2* contribute most to the predicted carcinogenicity values:

$$Se(0.65) > IC2(0.62) > DECC(0.50) > nR09(0.37) > C-003(0.30) > S-108(0.26) \quad (4)$$

The relative magnitude of the coefficients b_j^s (shown in parentheses) suggests that the numerical variables complement each other for predicting the activity. In addition, as the chosen set of descriptors take positive numerical values, it is concluded that the sign of the regression coefficients in Eq. 3 determine the value of the predicted carcinogenicity. Therefore, lower values for *Se* and *IC*2 and higher values for *nR*09, *DECC*, *C*−003 and *S*−108 would lead to higher predicted DF_{carc} values.

According to the applicability domain analysis as defined by Eq. 1, the results obtained indicate that 698 out of 732 compounds included in the test set belong to the AD of Eq. 3. The following 34 molecules have leverage values exceeding the warning leverage: Lorajmine, Stigmasterol, Vinconate, Metralindole, Dihydrocodeine, Pyrazophos, Tolciclate, Tilorone, Phosalone, Thiamylal, Dihydrotachysterol, Nitroscanate, Amoscanate, Diazinon, Cholesterol, Ergosterol, Rosaramicin, Ticarbodine, Nitrodan, Etrimfos, Dicapthon, Triazophos, Sulfathiourea, Fenthion,

Table 3. Correlation Matrix for Eq. 3

	Se	*nR09*	*DECC*	*IC2*	*C-003*	*S-108*
Se	1.00	0.03	0.66	0.40	0.17	0.01
nR09		1.00	0.00	0.04	0.14	0.00
DECC			1.00	0.32	0.04	0.00
IC2				1.00	0.05	0.00
C-003					1.00	0.00
S-108						1.00

Methidathion, Aldrin, Endrin, Paromomycin, Oxendolone, Mibolerone, Iloprost, Dromostanolone, Mesterolone, and Cortivazol. The predicted carcinogenicity for such compounds may not be considered as reliable, although this necessarily does not mean that the residuals for these test set compounds should be high. The leverage value for each compound is provided in Table **S2**.

4. CONCLUSIONS

The statistical results achieved for the heterogeneous molecular set composed of 1464 organic compounds are satisfactory and in line with previous reported studies from the literature. A main characteristic of the linear models established here is that they are based on geometry-independent molecular descriptors. In addition, we do not discard any outlier compound from the training set, including all of them during the analysis. We succeeded in establishing a quite simple QSAR model that includes only six molecular descriptors, by using an elaborated variable subset selection technique such as the Replacement Method and the Enhanced Replacement Method.

As future tasks for our drug research and development program, we pretend to continue exploring different ADME/Tox properties of drugs, including new attempts for QSAR improvements in this Galvez carcinogenicity data set, using different classes of structural descriptors in combination with linear/nonlinear methodologies.

CONFLICT OF INTEREST

The authors do not have competing financial interests to declare.

ACKNOWLEDGEMENTS

We thank the financial support provided by the National Research Council of Argentina (CONICET) PIP112201001 00151 project, and to Ministerio de Ciencia, Tecnología e Innovación Productiva for the electronic library facilities.

PATIENT'S CONSENT

Declared none.

SUPPLEMENTARY MATERIAL

Supplementary material is available on the publisher's web site along with the published article.

REFERENCES

[1] Hansch C, Leo A. Exploring QSAR. Fundamentals and applications in chemistry and biology. Washington, D. C. In American Chemical Society, 1995.

[2] Benfenati E. Quantitative structure-activity relationships (QSAR) for pesticide regulatory purposes. Amsterdam: Elsevier Science, 2007.

[3] Kubinyi H. QSAR: Hansch analysis and related approaches. New York: Wiley-Interscience, 2008.

[4] Puzyn T, Leszczynski J, Cronin MT. Recent advances in QSAR studies: methods and applications. New York: Springer, 2009.

[5] Esposito E, Hopfinger AJ. Multi-dimensional QSAR: methods and applications for drug discovery and polymer science. New York: CRC Press, 2012.

[6] Davis AM, Riley R. Predictive ADMET studies, the challenges and the opportunities. Curr Opin Chem 2004; 8: 378-86.

[7] Schuster D, Laggner C, Langer T. Why drugs fail?-a study on side effects in new chemical entities. Curr Pharm Des 2005; 11: 3545-59.

[8] Singh N, Guha R, Giulianotti MA, Pinila C, Houghten RA, Medina-Franco JL. Chemoinformatic analysis of combinatorial libraries, drugs, natural products, and molecular libraries small molecule repository. J Chem Inf Model 2009; 49: 1010-24.

[9] González-Díaz H, Duardo-Sánchez A, Ubeira FM, *et al*. Review of MARCH-INSIDE & complex networks prediction of drugs: ADMET, anti-parasite activity, metabolizing enzymes and cardiotoxicity proteome biomarkers. Curr Drug Metab 2010; 11: 379-406.

[10] Todeschini R, Consonni, V. Molecular descriptors for chemoinformatics. Weinheim: Wiley-VCH, 2009.

[11] Dehmer M, Varmuza K, Bonchev D. Statistical modelling of molecular descriptors in QSAR/QSPR (Quantitative and Network Biology). Berlin: Vch verlagsgesellschaft Mbh, 2012.

[12] Konovalov DA, Llewellyn LE, Heyden YV, Coomans D. Robust cross-validation of linear regression QSAR models. J Chem Inf Model 2008; 48: 2081-94.

[13] Schüürmann G, Ebert RU, Chen J, Wang B, Kühne R. External validation and prediction employing the predictive squared correlation coefficient test set activity mean *vs* training set activity mean. J Chem Inf Model 2008; 48: 2140-5.

[14] Chirico N, Gramatica P. Real external predictivity of QSAR models: how to evaluate it? Comparison of different validation criteria and proposal of using the concordance correlation coefficient. J Chem Inf Model 2011; 51: 2320-35.

[15] Roy K, Mitra I. On various metrics used for validation of predictive QSAR models with applications in virtual screening and focused library design. Comb Chem High Throughput Screen 2011; 14: 450-74.

[16] Morales AH, Perez MAC, Combes RD, Gonzalez MP. Quantitative structure-activity relationship for the computational prediction of nitrocompounds carcinogenicity. Toxicology 2006; 220: 51-62.

[17] Benigni R, Giuliani A. Chapter 19. QSAR approaches in mutagenicity and carcinogenicity estimation. In: van de Waterbeemd H, Testa B, Folkers G, editors. Computer-assisted lead finding and optimization: current tools for Medicinal Chemistry. Zürich: Verlag Helvetica Chimica Acta, 2007.

[18] González-Díaz H, Vilar S, Santana L, Uriarte E. Medicinal chemistry and bioinformatics-current trends in drugs discovery with networks topological indices. Curr Top Med Chem 2007; 7: 1015-29.

[19] González-Díaz H, González-Díaz Y, Santana L, Ubeira FM, Uriarte E. Proteomics, networks and connectivity indices. Proteomics 2008; 8: 750-78.

[20] González-Díaz H, Prado-Prado F, Ubeira FM. Predicting antimicrobial drugs and targets with the MARCH-INSIDE approach. Curr Top Med Chem 2008; 8: 1676-90.

[21] Garcia-Domenech R, Galvez J, de Julian-Ortiz JV, Pogliani L. Some new trends in chemical graph theory. Chem Rev 2008; 108: 1127-69.

[22] Helguera AM, Cordeiro MNDS, Pérez MAC, Combes RD, González MP. Quantitative structure carcinogenicity relationship for detecting structural alerts in nitroso-compounds. Species: rat; sex: male; route of administration: water. Toxicol Appl Pharmacol 2008; 231: 197-207.

[23] Massarelli I, Imbriani M, Coi A, Saraceno M, Carli N, Bianucci AM. Development of QSAR models for predicting

hepatocarcinogenic toxicity of chemicals. Eur J Med Chem 2009; 44: 3658-64.

[24] Benfenati E, Benigni R, DeMarini D, *et al*. Predictive models for carcinogenicity: frameworks, state-of-the-art, and perspectives. J Environ Sci Health 2009; 27: 57-90.

[25] González-Díaz H. Editorial [Hot topic: QSAR and complex networks in pharmaceutical design, microbiology, parasitology, toxicology, cancer and neurosciences]. Curr Pharm Des 2010; 16: 2598-600.

[26] Valerio Jr. LG. Computational science in drug metabolism and toxicology. Expert Opin Drug Metabol Toxicol 2010; 6: 781-4.

[27] Toropov AA, Toropova AP, Benfenati E, Gini G, Leszczynska D, Leszczynski J. SMILES-based QSAR approaches for carcinogenicity and anticancer activity: comparison of correlation weights for identical SMILES attributes. Anti-Cancer Agents Med Chem 2011; 11: 974-82.

[28] Putz MV, Ionaşcu C, Putz A-M, Ostafe V. Alert-QSAR. Implications for electrophilic theory of chemical carcinogenesis. Int J Mol Sci 2011; 12: 5098-134.

[29] Putz MV. Residual-QSAR. Implications for genotoxic carcinogenesis. Chem Cent J 2011; 5: 1-11.

[30] Kar S, Roy K. First report on development of quantitative interspecies structure-carcinogenicity relationship models and exploring discriminatory features for rodent carcinogenicity of diverse organic chemicals using OECD guidelines. Chemosphere 2012; 87: 339-55.

[31] Kar S, Deeb O, Roy K. Development of classification and regression based QSAR models to predict rodent carcinogenic potency using oral slope factor. Ecotoxicol Environ Saf 2012; 82: 85-95.

[32] Tarko L, Putz MV. On quantitative structure-toxicity relationships (QSTR) using high chemical diversity molecules group. J Theor Comput Chem 2012; 11: 265-72.

[33] Galvez J. Personal webpage. http://www.uv.es/-galvez/tablevi.pdf

[34] Hemmateenejad B, Safarpour MA, Miri R, Nesari N. Toward an optimal procedure for PC-ANN model building: prediction of the carcinogenic activity of a large set of drugs. J Chem Inf Model 2005; 45: 190-9.

[35] Deeb O, Hemmateenejad B, Jaber A, Garduno-Juarez R, Miri R. Effect of the electronic and physicochemical parameters on the carcinogenesis activity of some sulfa drugs using QSAR analysis based on genetic-MLR and genetic-PLS. Chemosphere 2007; 67: 2122-30.

[36] Kar S, Roy K. Development and validation of a robust QSAR model for prediction of carcinogenicity of drugs. Indian J Biochem Biophys 2011; 48: 111-22.

[37] The PubChem Project. http://pubchem.ncbi.nlm.nih.gov

[38] HyperChem 7 (Hypercube, Inc.). http://www.hyper.com

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

[40] Recon Version 5.5. Rensselaer Polytechnic Institute, Troy, New York, USA http://www.drugmining.com

[41] Lavine BK, Davidson CE, Breneman C, Katt W. Electronic van der waals surface property descriptors and genetic algorithms for developing structure-activity correlations in olfactory databases. J Chem Inf Comput Sci 2003; 43: 1890-905.

[42] Worachartcheewan A, Nantasenamat C, Naenna T, Isarankura-Na-Ayudhya C, Prachayasittikul V. Modeling the activity of furin inhibitors using artificial neural network. Eur J Med Chem 2009; 44: 1664-73.

[43] Bader RFW. Atoms in Molecules-A Quantum Theory. Oxford: Clarendon Press, 1990.

[44] Breneman CM, Weber LW. Transferable atom equivalents. Assembling accurate electrostatic potential fields for large molecules from ab initio and PROAIMS results on model systems. In: Jeffrey GA, Piniella JF, editors. The Application of Charge Density Research to Chemistry and Drug Design. New York: Plenum, 1991.

[45] Duchowicz PR, Castro EA, Fernández FM. Alternative algorithm for the search of an optimal set of descriptors in QSAR-QSPR studies. MATCH Commun Math Comput Chem 2006; 55: 179-92.

[46] Duchowicz PR, Castro EA, Fernández FM, González MP. A new search algorithm of QSPR/QSAR theories: normal boiling points of some organic molecules. Chem Phys Lett 2005; 412: 376-80.

[47] Duchowicz PR, Talevi A, Bruno-Blanch LE, Castro EA. New QSPR study for the prediction of aqueous solubility of drug-like compounds. Bioorg Med Chem 2008; 16: 7944-55.

[48] Goodarzi M, Duchowicz PR, Wu CH, Fernández FM, Castro EA. New hybrid genetic based support vector regression as QSAR approach for analyzing flavonoids-GABA(A) complexes. J Chem Inf Model 2009; 49: 1475-85.

[49] Pomilio AB, Giraudo MA, Duchowicz PR, Castro EA. QSPR analyses for aminograms in food: citrus juices and concentrates. Food Chem 2010; 123: 917-27.

[50] Talevi A, Goodarzi M, Ortiz EV, *et al*. Prediction of drug intestinal absorption by new linear and non-linear QSPR. Eur J Med Chem 2011; 46: 218-28.

[51] Pasquale G, Romanelli GP, Autino JC, García J, Ortiz EV, Duchowicz PR. Quantitative structure-activity relationships on chalcone derivatives: mosquito larvicidal studies. J Agric Food Chem 2012; 60: 692-7.

[52] Mercader AG, Duchowicz PR, Fernández FM, Castro EA. Replacement method and enhanced replacement method versus the genetic algorithm approach for the selection of molecular descriptors in QSPR/QSAR theories. J Chem Inf Model 2010; 50: 1542-8.

[53] Mercader AG, Duchowicz PR, Fernández FM, Castro EA. Advances in the replacement and enhanced replacement method in QSAR and QSPR theories. J Chem Inf Model 2011; 51: 1575-81.

[54] Matlab 7.0. Masachussetts, USA: The MathWorks, Inc., http://www.mathworks.com

[55] Golbraikh A, Tropsha A. Beware of q2! J Mol Graphics Modell 2002; 20: 269-76.

[56] Wold S, Eriksson L. Statistical validation of QSAR results. In: van de Waterbeemd H, editor. Chemometrics methods in molecular design. Weinheim: VCH, 1995. p. 309-18.

[57] Gramatica P. Principles of QSAR models validation: internal and external. QSAR Comb Sci 2007; 26: 694-701.

[58] Eriksson L, Jaworska J, Worth AP, Cronin MT, McDowell RM, Gramatica P. Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. Environ Health Perspect 2003;111: 361-75.

[59] Draper NR, Smith H. Applied regression analysis. New York: John Wiley&Sons, 1981.

Received: August 8, 2012 Revised: September 24, 2012 Accepted: September 26, 2012