Molecular Diversity (2006) 10: 159–168
DOI: 10.1007/s11030-005-9007-z

Full-length paper

True prediction of lowest observed adverse effect levels

R. García-Domenech[1,*], J.V. de Julián-Ortiz[2] & E. Besalú[3]
[1]*Unidad de Investigación de Diseño de Fármacos y Conectividad Molecular, Departamento de Química Física, Facultad de Farmacia, Universitat de València, Burjassot, Valencia, Spain;* [2]*Red de Investigación de Enfermedades Tropicales Departamento de Biología Celular y Parasitología, and Departamento de Química Física. Facultad de Farmacia, Universitat de València, Burjassot, Valencia, Spain;* [3]*Institut de Química Computacional. Universitat de Girona, Girona, Spain (*Author for correspondence, E-mail: ramon.garcia@uv.es)*

Received 2 September 2005; Accepted 7 November 2005

Key words: LOAEL, multiple linear regression, linear discriminant analysis, true prediction, ITS method, topological indices

Summary

A database of structurally heterogeneous chemical structures with their experimental values of Lowest Observed Adverse Effect Levels (LOAELs) was modeled using graph theoretical descriptors. Variable selection for multiple linear regression (MLR) and linear discriminant analysis (LDA) was accomplished by the Internal Test Set (ITS) method in order to achieve true predicted LOAEL values. The results obtained can be considered good if we take in count the structural diversity of the training set.

Introduction

The modeling of toxicological properties is an extremely important problem. No empirical toxicological data are available for most chemicals, and the growing new ones must be evaluated or, at least estimated. Thus, reliable methods to predict environmental toxicity are required. Furthermore, the presence of toxic substances in high trophic levels can affect humans. Particular interest in the estimation of chronic lowest observed adverse effect level (LOAEL) has been raised recently due to its environmental implications. LOAEL was defined by the IUPAC as the lowest concentration or amount of a substance, found by experiment or observation, which causes an adverse alteration of morphology, functional capacity, growth, development, or life span of a target organism distinguishable from normal (control) organisms of the same species and strain under defined conditions of exposure [1].

Topological indices (TIs) are non-empirical descriptors calculated from the representation of the molecules as mathematical graphs [2, 3]. These descriptors are able to characterise the most important features of molecular structure: molecular size, binding and branching. The computation of TIs is very swift and they also have the advantage of being true structural invariants. This means that the TIs are independent of the spatial position of the atoms in a particular moment. However, extensions of the TIs that give account of the three-dimensional structure have been also devised [4–6].

TIs have been useful in the prediction of physical [7, 8], chemical and biological properties, even for groups of compounds that show considerable structural diversity. It can be pointed out, among the properties modelled, various therapeutic activities as well as toxicological properties [9], the drug-like character [10, 11], or the molecular similarity/diversity [12].

Today, it is known that topostructural and topochemical information explains the main part of the predicted properties, and that the inclusion of three-dimensional features results in slightly improved predictive models [13].

Logarithms of octanol-water partition coefficients (logP) have been exhaustively used to model toxicological properties. The main reason that moves us to not using it is that logP is a physical empirical descriptor while we attempted to describe toxicological properties only with graph-theoretical structural parameters, which would be capable to give the information contained in logP.

In the QSAR field, mathematical models often are presented as a lineal equation of certain descriptors selected in a particular way with a good adjustment for the experimental data within the series. In this work only linear models will be discussed. These models usually come accompanied by a test of validation of leave-one-out (L-1-O) type in which the value of the property for each molecule is evaluated by an equation obtained with the whole rest of the population, in which the selected variables remain fixed. However, when applying the equations to molecules that do not appear in the series of training, the results of prediction of the property are usually very poor. In part, this is due to the particular procedure which has been followed in order to perform the cross-validation. The methodology explored here is called the Internal Tests Sets (ITS) protocol and, as it will be seen,

constitutes a more severe L-1-O or Leave-many-out procedure. Our experience reveals to us that this method also allows the automatic identification of outliers.

The aim of the present research was to model the true prediction of chronic LOAELs for a heterogeneous group of chemicals by using TIs as molecular descriptors. This study continues from a previous article [14] which has analyzed a compiled database of 234 compounds from different sources [15] in order to assess its homogeneity. This study has concluded that data were not homogeneous, and that only those from the U.S. Environmental Protection Agency (EPA) reports could be well modelled with graph theoretical descriptors by multilinear regression (MLR) and linear discriminant analysis (LDA). In contrast, data estimated from specific procedures from the National Toxicology Program (NTP) database introduced noise and did not render good models either alone, or in combination with the EPA data. In spite of this, this database has been used in a LOAEL prediction study with five submodels for acyclics, alicyclics, single benzenes, multiple benzenes and heteroaromatics; and constitutes part of the training set in which is based the LOAEL module of program TOPKAT [16].

Materials and methods

The EPA database of rat chronic LOAELs consisted of 87 compounds. Some of these chemicals are sodium salts. Since these substances are dissociated in water, only the corresponding anion was considered as toxic agent in the calculations. LOAEL data were originally expressed in units of milligrams of chemical per kilograms of body weight per day (mg/(Kg·day)). These values were converted to their corresponding μmol/Kg in order to compare the respective correlations. Regressions were performed with log(LOAEL). The descriptors used are listed in Table 1. They were calculated with the DESCRI [17] and Molconn-Z [18] programs.

Standard statistical methods have been traditionally used to obtain statistically sound models. The statistical program package BMDP New System 2.0/Dynamic Release 7.0 was used for these calculations in this work [19]. Variable selection was performed by means of the Furnival-Wilson algorithm in MLR and variable sets with the minimal Mallows Cp were selected as best equations, while stepwise procedure was used with LDA [20]. Two tests were performed to test the robustness of the models: randomness and validation in an external set. The first one consists of scrambling the dependent variable value among the molecules. This is particularly interesting in our data set since there are some cases in which different chemical structures show the same experimental LOAEL value. The second is the testing of the prediction power in a set not used in the obtaining of the equation.

Nevertheless, the classical leave-one-out (L-1-O) MLR fitting procedure implemented in most programs, as in BMDP, consists of fixing a set of descriptors and *a posteriori* perform the operations of successive remove/replacement of molecules. In this way, the parameters (selected descriptors) entering into the linear equation are maintained fixed during the iterative process of L-1-O. Commonly, some statistical parameters are calculated (r^2, q^2, F, p) in order to be optimized, but this calculation is repeated every time a new set of descriptors is being considered. From the algorithmic point of view, in this classical procedure the external loop is attached to the parameters combination selection and, then, internally, the hide/replace-a-molecule operation (the L-1-O) is done. Usually, the final descriptors entering in the proposed model are the subset which optimizes one of the goal parameters (r^2, q^2, F, p). Recently Livingstone and Salt [21] have justified that this procedure can generate artificial impressive statistical parameters. This is so because the algorithms used try to find optimal goal function values while several sets of descriptors (some times hundreds or millions) are being tested.

In our laboratory we implemented an alternative unsupervised approach, the Internal Tests Sets (ITS) procedure. This protocol demands much more computation time and it consists in performing the hide- molecules operation in the algorithmic external loop and, then, select the descriptors in the internal or core procedure. This constitutes a *true* cross-validation, as the indices entering the equation are re-selected from scratch every time a molecule is hidden from the system. This leads to a different model equation (having different coefficients or even different descriptors) attached to every molecule that is 'left-out'. Every equation is used to make the prediction over the removed structure. It is noticeable that this prediction is unique. So, concerning the predictions, the method is much more risky than the classical approach, but it simulates a real situation: to first build a model and then make predictions with it over a really unknown structure. ITS calculations have been done with a build in-house program, REGRE [22]. This program is able to deal with MLR models or with LDA ones. Its nuclear module iteratively searches for the best set of descriptors entering into the model. The program does not build models in a stepwise manner, but performs an exhaustive search of descriptors combinations. When MLR models are being found, the r^2 parameter is optimized. If the models are LDA functions, the selected model is the one presenting a maximal Mahalanobis distance between two groups (active and less active in our case). This is equivalent to maximizing the Wilks λ parameter.

Additionally, the REGRE core module can be called iteratively removing information from one structure at a time. This implements the L-1-O- ITS protocol described above, as the obtained model is used to make only a single prediction. The module is called as many times as molecules are in the database, thus generating as many models as molecules. Under completion, all the predictions are collected and statistically analyzed. From the numerical point of view, the statistical parameters attached to the predictions are not so spectacular as those coming from simple fittings over entire

Table 1. Descriptors used in this study.

Symbol	Name	Definition	Refs.
MW	Molecular weight		–
R	Ramification	Number of single structural branches	[25]
N_{el}	Number of elements	Number of different elements in the compound	-
nrings	Rings	Number of rings	–
PR_k $k = 0$–3	Pairs of ramifications at distance k	Number of pairs of single branches at distance k in terms of bonds	[25]
${}^k\chi_t$ $k = 0$–4 $t = p, c, pc$	Randić-like indices of order k and type path (p), cluster (c) and path-cluster (pc)	${}^k\chi_t = \sum_{j=1}^{k_{n_t}} \left(\prod_{i \in S_j} \delta_i \right)^{-1/2}$ δ_i, number of bonds, σ or π, of the atom i to non-hydrogen atoms. S_j, jth sub-structure of order k and type t	[26, 27]
${}^k\chi_t^v$ $k = 0$–4 $t = p, c, pc$	Kier-Hall indices of order k and type path (p), cluster (c) and path-cluster (pc)	${}^k\chi_t^v = \sum_{j=1}^{k_{n_t}} \left(\prod_{i \in S_j} \delta_i^v \right)^{-1/2}$ ${\delta_i}^v$, Kier-Hall valence of the atom i. S_j, jth sub-structure of order k and type t.	[28]
G_k $k = 1$–5	Topological charge indices of order k	$G_k = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \lvert \mathbf{M}_{ij} - \mathbf{M}_{ji} \rvert \delta(k, \mathbf{D}_{ij})$ $\mathbf{M} = \mathbf{AQ}$, product of the adjacency and inverse squared distance matrices for the hydrogen-depleted molecular graph. $\mathbf{D}$, distance matrix. δ, Kronecker delta	[29]
${G_k}^v$ $k = 1$–5	Valence topological charge indices of order k	$G_k^v = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \lvert \mathbf{M}_{ij}^v - \mathbf{M}_{ji}^v \rvert \delta(k, \mathbf{D}_{ij})$ $\mathbf{M}^v = \mathbf{A}^v\mathbf{Q}$, product of the electronegativity-modified adjacency and inverse squared distance matrices for the hydrogen-depleted molecular graph. $\mathbf{D}$, distance matrix. δ, Kronecker delta	[29]
J_k and J_k^v $k = 1$–5	Normalized topological charge indices of order k	$J_k = \frac{G_k}{N-1}$ $J_k^v = \frac{G_k^v}{N-1}$	[29]
${}^kD_t$ and ${}^kC_t$ $k = 0$–4 $t = p, c, pc$	Connectivity differences and quotients of order k and type path (p), cluster (c) and path-cluster (pc)	${}^kD_t = {}^k\chi_t - {}^k\chi_t^v$ ${}^kC_t = \frac{{}^k\chi_t}{{}^k\chi_t^v}$	[25]
S_i^T	Atom type electrotopological state indices	Sum of E-state values for all the atoms of a given atom type	[30]
κ_2	Kappa simple index (second-order)	$\kappa_2 = \frac{(A-1)(A-2)^2}{({}^2P_i)^2}$ A = number of atoms; ${}^2P_i$ = number of two-path fragments	[31]
$\kappa_{\alpha 1}$	Kappa alpha index (first-order)	$\kappa_{\alpha 1} = (A+\alpha)(A+\alpha-1)^2({}^1P_i+\alpha)^2$ $\alpha = (\frac{r_x}{r_{Csp3}}) - 1$ r_x = covalent radius of atom x	[31]

molecular tests (and some times selecting a few descriptors from a big pool), but the advantage is that our results constitute a reliable measure of the predictive power of the selected kind of models (MLR or LDA) and descriptors. Additionally, unstabilities among predictions can be associated with the presence of outliers.

Results and discussion

Equations with LOAELs in molar units gave always better correlation coefficients that in original form. This is reasonable since molar concentrations are measures of the number of molecules that exert the activity, but it is not possible to establish *a priori* that molar units will be always better predicted than weight ones, since the main part of the molecular descriptors show positive correlation with the molecular mass [23]. Thus, only results with molar units are presented. The MLR correlated property is, then, the decimal logarithm of rat chronic LOAEL expressed in Micromoles of chemical per kilogram of body weight per day μmol/(Kg·day)).

Standard MLR

A six variable (${}^4\chi_{PC}$, G_4^V, J_2, J_5^V, $PR0$, $PR1$) equation has been presented in a previous article [14]. It was obtained with a more reduced pool of variables. The best MLR equation obtained here by standard statistical methods gave the calculated *vs.* experimental LOAEL (μmol/Kg) shown in Figure 1. The MLR equation consisted of a linear combination of 11 variables. The coefficients and statistical parameters are shown in Table 2. The presence of the three electro-topological indices ($S^T(=C<)$, $S^T(\equiv N)$, $S^T(P)$) and the two kappa (κ_2, $\kappa_{\alpha 1}$) diminishes SEE 22% in relation to the six variable equation. The set of descriptors selected showed generally

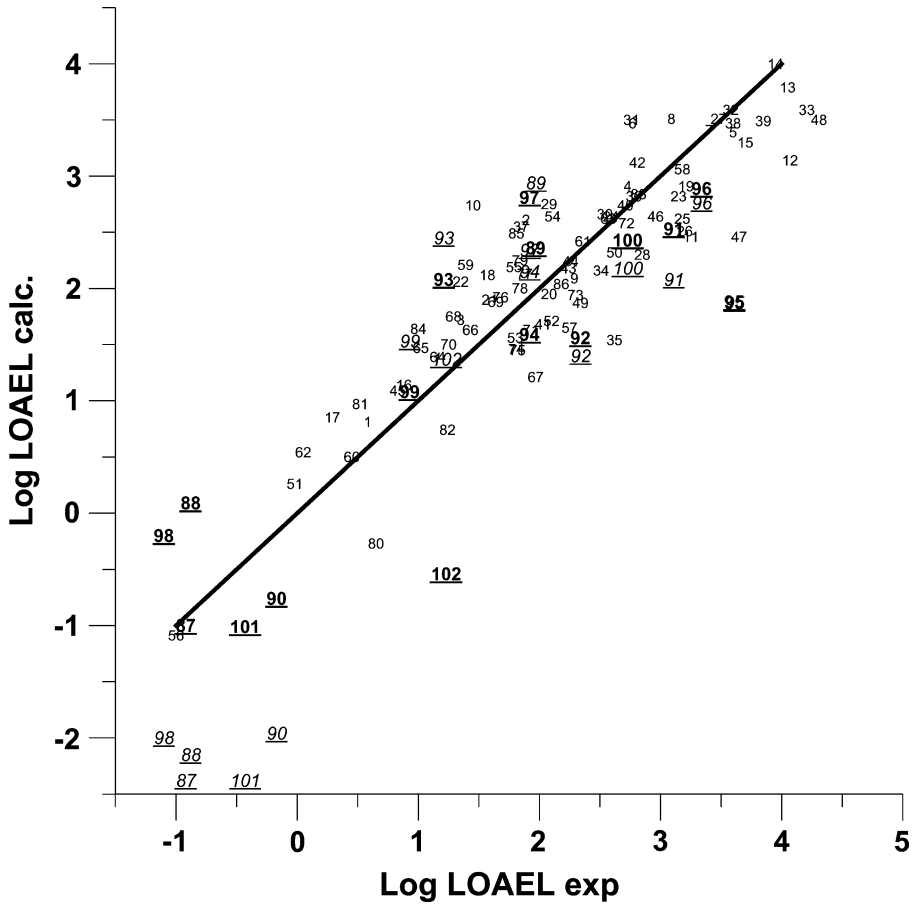

Figure 1. Calculated vs. experimental chronic LOAEL values. Numerical values identify a training molecule. Underlined values stand for test molecules (bold for 11 terms equation and italic for the ITS 2 terms model).

low values of intercorrelation, with an average absolute correlation = 0.196. The strongest intercorrelation indices were $\rho(\kappa_2, \kappa_{\alpha 1}) = -0.873$, $\rho(G_4^V, J_5^V) = -0.718$, ρ ($PR0$, $PR1$) = 0.549 and $\rho(PR0, \kappa_2) = 0.500$.

Table 2. Statistical parameters of the 11-variable equation in the QSAR study for the training group.

Variable	Coefficient	SE	T-Stat	Contribution to r^2
Intercept	3.884	0.218	17.84	–
κ_2	0.295	0.080	3.66	0.0372
$\kappa_{\alpha 1}$	−0.281	0.059	−4.77	0.0631
S^T(=C<)	−0.363	0.108	−3.35	0.0312
S^T(≡N)	−0.112	0.030	−3.67	0.0374
S^T(P)	0.347	0.073	4.77	0.0632
$^4\chi_{PC}$	−0.621	0.127	−4.90	0.0665
G_4^V	0.712	0.151	4.70	0.0613
J_2	−0.997	0.309	−3.22	0.0288
J_5^V	−16.136	3.549	−4.55	0.0574
$PR0$	0.862	0.200	4.32	0.0517
$PR1$	0.318	0.055	5.82	0.0941

$N = 86$; $r^2 = 0.795$; SEE = 0.517; $r_{cv}^2 = 0.719$; $\text{SEE}_{cv} = 0.564$; $F(11, 74) = 26.0$; $p < 0.00001$.

The maximum r^2 obtained in ten runs of the randomness test was 0.14, and the minimum SEE, 0.958, that corresponded to the same equation.

Table 3 shows the experimental data, the results of prediction with the 11 variable equation for the training set and the results of cross-validation prediction. This is the prediction for a compound obtained with an equation with the *same variables* (as said above) whose coefficients have been adjusted with all the rest of training compounds (L-1-O). A number of 61 out of 86 compounds (71%) show residuals lower than ±1 SEE. Only two, maleic hydrazide and acrylamide, have residual values greater than ±2 SEE. There are four compounds with residuals greater than ±2 SEE in the crossvalidation results: maleic hydrazide, acrylamide, Bis (2-chloroisopropyl) ether and tralomethrin. They must be considered as outliers.

This model has given good prediction results in an external set of compounds. For this validation, predictions were made for 16 molecules separated prior to *training*. The results are shown in Table 4 (column four). The equation is able to recognize the most toxic compounds, logLOAEL low, except in the case of tetrachlorovinphos. This model is, then, reasonably robust.

Table 3. CAS registry number, name of chemicals and Log chronic LOAEL values obtained in the QSAR study for the training group.

#	CAS registry	Compound	Chronic LOAEL (Log) (μmol/kg) Exp.	Calc	Calc (cv)
1	50-18-0	Cyclophosphamide	0.58	0.81	0.85
2	63-25-2	Carbaryl	1.89	2.61	2.63
3	67-45-8	Furazolidone	1.35	1.72	1.74
4	67-66-3	Chloroform	2.73	2.91	2.94
5	71-55-6	1,1,1-Trichloroethane	3.60	3.39	3.34
6	75-09-2	Dichloromethane	2.77	3.47	3.53
7	75-69-4	Fluorotrichloromethane	3.40	3.41	3.41
8	75-71-8	Dichlorodifluoromethane	3.09	3.51	3.60
9	75-99-0	Dalapon	2.29	2.09	2.05
10	79-06-1	Acrylamide	1.45	2.74	2.83
11	80-62-6	Methyl-methacrylate	3.25	2.46	2.43
12	81-07-2	Saccharin	4.07	3.14	3.04
13	84-66-2	Diethylphthalate	4.05	3.79	3.77
14	84-72-0	Ethylphthalyl-ethyl-glycolate	3.95	4.00	4.02
15	85-44-9	Phthalic-anhydride	3.70	3.30	3.26
16	87-68-3	Hexachlorobutadiene	0.88	1.14	1.20
17	87-84-3	1,2,3,4,5-Pentabromo-6-chlorocyclohexane	0.29	0.85	1.16
18	87-86-5	Pentachlorophenol	1.57	2.12	2.25
19	92-52-4	1,1′-Biphenyl	3.21	2.91	2.90
20	93-65-2	2-(2-Methyl-4-chloro-phenoxy) propionic acid	2.08	1.95	1.94
21	93-76-5	Trichlorophenoxy	1.59	1.90	1.93
22	94-75-7	2,4-Dichlorophenoxy	1.35	2.06	2.11
23	95-53-4	2-Toluidine	3.15	2.82	2.81
24	95-57-8	2-Chlorophenol	2.59	2.64	2.64
25	95-70-5	Toluene-2,5-diamine	3.18	2.62	2.58
26	97-63-2	Ethylmethacrylate	3.20	2.51	2.48
27	100-21-0	4-Phthalic acid	3.48	3.51	3.52
28	101-21-3	Chlorpropham	2.85	2.30	2.26
29	103-69-5	*N*-Ethylaniline	2.08	2.75	2.78
30	106-50-3	4-Phenylenediamine	2.54	2.66	2.67
31	107-07-3	Chloroethanol	2.76	3.50	3.56
32	107-15-3	Ethylenediamine	3.58	3.59	3.60
33	107-21-1	Ethylene-glycol	4.21	3.59	3.54
34	108-31-6	Maleic-anhydride	2.51	2.16	2.11
35	108-60-1	Bis (2-chloroisopropyl) ether	2.62	1.54	1.42
36	108-91-8	Cyclohexylamine	2.78	2.82	2.82
37	109-78-4	Ethylene cyanohydrin	1.85	2.55	2.85
38	110-80-5	2-Ethoxyethanol	3.60	3.47	3.45
39	111-90-0	Diethylene-glycol-monoethyl-ether	3.85	3.49	3.38
40	117-81-7	Di-2-ethylhexyl phtalate	2.71	2.74	2.75
41	120-36-5	Dichloroprop	2.03	1.68	1.66
42	120-61-6	Dimethyl terephthalate	2.81	3.12	3.14
43	120-82-1	1,2,4-Trichlorobenzene	2.24	2.18	2.17
44	120-83-2	2,4-Dichlorophenol	2.26	2.24	2.24
45	121-82-4	RDX Cyclonite	0.83	1.09	1.16
46	122-39-4	*N,N*-Diphenylamine	2.96	2.64	2.62
47	123-33-1	Maleic hydrazide	3.65	2.46	2.35
48	131-11-3	Dimethyl phthalate	4.31	3.50	3.43
49	139-40-2	Propazine	2.34	1.87	1.75
50	148-18-5	Sodium diethyl dithiocarbamate	2.62	2.32	2.29
51	298-00-0	*O,O*-Dimethyl-*O*-(4-nitrophenyl)phosphorothioate	−0.02	0.26	0.32
52	330-55-2	Linuron	2.10	1.71	1.69
53	732-11-6	Phosmet	1.80	1.56	1.51
54	823-40-5	Toluene-2,6-diamine	2.11	2.64	2.67
55	886-50-0	Terbutryn	1.79	2.19	2.29
56	1031-47-6	Triamiphos	−1.00	−1.09	−1.14
57	1071-83-6	Glyphosate	2.25	1.65	1.35
58	1861-32-1	Dacthal	3.18	3.06	3.02
59	1929-77-7	Vernolate	1.39	2.21	2.32
60	2921-88-2	Chlorpyrifos	0.45	0.50	0.51
61	3761-53-3	Sodium 3-OH-4-(2,4-xylylazo)-2,7-Napthalenedisulfonate	2.36	2.42	2.44
62	6923-22-4	Monocrotophos	0.05	0.54	0.66
63	15299-99-7	Napropamide	2.57	2.62	2.63
64	19666-30-9	Oxadiazon	1.16	1.39	1.41
65	21725-46-2	Cyanazine	1.02	1.47	1.68
66	23135-22-0	Oxamyl	1.43	1.63	1.65
67	23564-05-8	Thiophanatemethyl	1.97	1.21	0.99
68	28249-77-6	Thiobencarb	1.29	1.75	1.78
69	34014-18-1	Tebuthiuron	1.64	1.88	1.90
70	40487-42-1	Pendimethalin	1.25	1.50	1.54
71	43121-43-3	Bayleton	1.93	1.63	1.60
72	51218-45-2	Metolachlor	2.72	2.58	2.56
73	51235-04-2	Hexazinone	2.30	1.94	1.81
74	52645-53-1	Permethrin	1.81	1.46	1.41
75	55285-14-8	Carbosulfan	1.82	1.45	1.40
76	55290-64-7	Dimethipin	1.68	1.92	2.05
77	59756-60-4	Fluridone	1.88	2.13	2.17
78	62476-59-9	Sodiumacifluorfen	1.84	2.00	2.02
79	64902-72-3	Chlorsulfuron	1.84	2.25	2.28
80	66841-25-6	Tralomethrin	0.65	−0.27	−0.68
81	68085-85-8	Cyhalothrin	0.52	0.97	1.20
82	68359-37-5	Baythroid	1.24	0.74	0.57
83	74223-64-6	Ally	2.82	2.84	2.84
84	76578-14-8	Assure	1.00	1.64	1.71
85	79277-27-3	Harmony®	1.81	2.49	2.60
86	82558-50-7	Isoxaben	2.18	2.04	2.02

ITS-MLR

REGRE program has been used to obtain linear models under the ITS protocol described above. Table 5 gives the r^2 values between experimental properties and the predicted ones obtained with each number of variables. Note that, for each

Table 4. CAS registry number, name of chemicals and LogLOAEL values obtained in the QSAR study for the external test group (11 variable equation and 2 variable equation).

			logLOAEL (μmol/kg)		
	CAS registry #	Compound	Experimental	Calculated (11 variables)	Calculated (2 variables)
87	57-74-9	Chlordane	−0.92	−1.00	−2.38
88	60-57-1	Dieldrin	−0.88	0.09	−2.15
89	75-35-4	1,1-Dichloroethylene	1.97	2.36	2.94
90	76-44-8	Heptachlor	−0.17	−0.76	−1.96
91	78-59-1	Isophorone	3.11	2.53	2.08
92	80-05-7	Bisphenol A	2.34	1.56	1.40
93	99-55-8	2-Methyl-5-nitroaniline	1.21	2.08	2.45
94	101-61-1	4,4′-Methylenebis-(*N*,*N*-dimethylaniline)	1.92	1.59	2.15
95	103-23-1	Di-(2-ethylhexyl)adipate	3.61	1.88	1.87
96	105-60-2	Caprolactam	3.34	2.89	2.76
97	133-06-2	Captan	1.92	2.81	2.34
98	309-00-2	Aldrin	−1.10	−0.20	−2.00
99	319-84-6	α-Hexachloro-cyclohexane	0.93	1.08	1.53
100	630-20-6	1,1,1,2-Tetrachloroethane	2.73	2.43	2.18
101	959-98-8	α-Endosulfan	−0.43	−1.01	−2.38
102	961-11-5	Tetrachlorovinphos	1.23	−0.54	1.37

Table 5. Maximum r^2 obtained with ITS for each number of variables.

Variables	r^2
1	0.362
2	0.419
3	0.378
4	0.307

number of variables, a total of 87 models were found, and the fitting data of each came from 86 structures. The results can be considered as good since the predictions are unbiased, and the results obtained depend exclusively on the information that the descriptors are able to contain in relation to the predicted property.

The most significant result was obtained with two variables (value of p of the order of 10^{-10}). Models with more than 2 variables are not so useful to generalize. In this case, the same indices are always chosen throughout all the 87 models of two variables. This indicates the robustness of the approach. It must be remembered here that the variables are freely chosen from the pool and, in a set of random values, rarely the best predictions would be made with the same pair twice. The variables chosen were PR1, always with positive sign, and ${}^4\chi_{PC}$, always with negative sign. The true predictions plot with two variables is shown in Figure 2.

Comparing the ITS result with the standard one, it is noteworthy that the two ITS-selected variables are included, with their respective signs, within the 11 variable set. ITS equations are simplified versions of the *standard* equation, but allowing a very different kind of statistical interpretation. Now, we will see how ITS equations predict the studied property in the external test set used to validate the 11 variable standard equation. First, once unambiguously selected the two better variables, $PR1$ and ${}^4\chi_{PC}$, a single equation was obtained with the entire train set. This resulted:

$$\log \text{LOAEL} = (0.201 \pm 0.060)PR1 - (0.719 \pm 0.088){}^4\chi_{PC} + 2.942 \pm 0.136$$
$$N = 86 \quad r^2 = 0.461 \quad \text{SEE} = 0.79$$
$$F(2, 83) = 35.97 \quad p = 1.2 \times 10^{-10}$$

The equation explains almost 50% of the logLOAEL variance, which is a good result if we take account of the number of data. This equation shows great significance, $p = 1.2 \times 10^{-10}$, which is the parameter that determines the chance probability to obtain the same result with random numbers. Thus, the maximum r^2 obtained in ten runs of the randomness test was 0.02, while SEE gave 0.97 for all the runs.

This two variable equation gave quadratic deviation = residuals^2 = 20.27, while the 11 variable equation gave 11.62 for the same value. It is not strange since the regression algorithm searches the minimum possible value for this parameter arranging the variables and coefficients with this objective, and the variables involved, at least two of them, really afford structural information to relate the property with the structure. But if we calculate the Pearson correlation coefficient of the observed and predicted values, the information obtained is different. This coefficient, for the 11 variable equation is

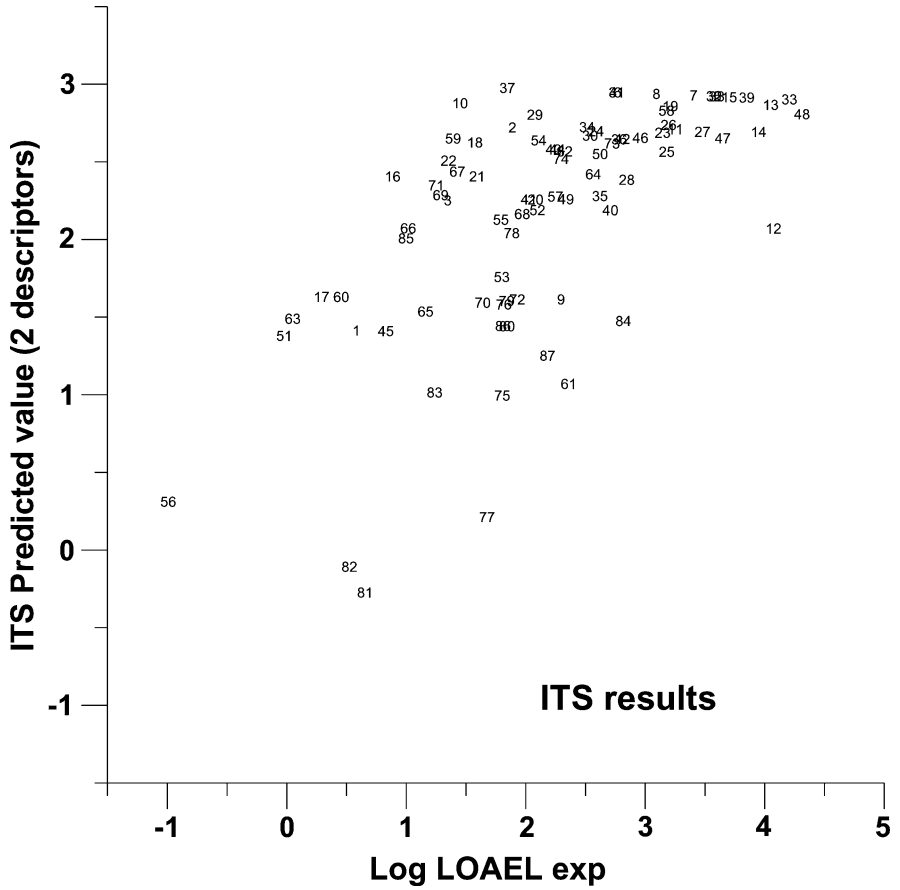

Figure 2. True prediction logLOAEL values for 2 variables obtained following the L-1-O-ITS procedure.

0.844, while for the two variable one is 0.882. In fact tetrachlorovinphos toxicity is better predicted using the last equation. Furthermore, the sign of the predicted values is always successfully predicted by the simplest equation, while the complex one, with nine variables more, fails twice, for dieldrin and for tetrachlorovinphos (see Table 4 and Figure 1). This fact prompted us to try the true prediction ITS method with LDA.

ITS-LDA

Using again the REGRE program, the ITS procedure has been followed to found from 1 up to 3 indices discriminant models. As the direct computation demands to generate an inaccessible number of variable combinations, in order to reach models up to 6 descriptors the parameters entering in the model of 3 indices were kept fixed. In all the cases, the 87 models of 3 descriptors involved the same three indices: the number of distinct elements present in the molecule (N_{el}), and the Randić-Kier-Hall ${}^3\chi_p$ and ${}^4\chi_p$ indices. Table 6 lists the classification ratios achieved. The best model is the one combining 3 descriptors. This model not only presents the best percentage of well done classifications (77%), but also it is a robust one because in all the calculations the three indices mentioned above entered into the discriminant equation. Our experience indicates that, when obtaining predictions following the ITS procedure described here, usually a small number of parameters must enter into the equation. The same occurred in the above linear models. This constitutes a word of caution, because it is very common in the literature to present models of more parameters and be taken as predictive. In fact, in most of these cases overfitting problems are present, but the researcher does not realize it.

The 87 predictions arising from the models of 3 descriptors are depicted in Figure 3. In abscissas the experimental property value is represented. The vertical line denotes the mean experimental value (2.21) and sets the frontier between toxic (left part) and less active (right part) compounds. In ordinates, the discriminant function value is represented. A positive value classifies the compound as being active, and a negative value classifies it as being less toxic. According to this, the depicted boxes delimit the correct classifications, and the points outside them constitute the type I and type II errors. As expected, most of the errors are found near the classification limit (discriminant function equal to zero). The correlation coefficient between the discriminant function value and the property one ($r^2 = 0.414, r = -0.644$) seems to be poor... but it is not taking into account that the

Table 6. Classification table for the set of 87 molecules. The last model concerns the global fitted training model. See text for more details.

Descriptors in model	Classified as actives		Classified as inactives		Overall classifications		Percentage of good classifications among the ones classified as		Overall percentage of well classified
	Success	Wrong	Success	Wrong	Success	Wrong	Actives	Inactives	
1	22	7	37	21	59	28	75.8	63.8	67.8
2	26	10	34	17	60	27	72.2	66.7	69.0
3	33	10	34	10	67	20	76.7	77.3	77.0
4	30	9	35	13	65	22	76.9	72.9	74.7
5	26	16	28	17	54	33	61.9	62.2	62.1
6	29	13	31	14	60	27	69.0	68.9	69.0
3′	34	9	35	9	69	18	79.1	79.5	79.3

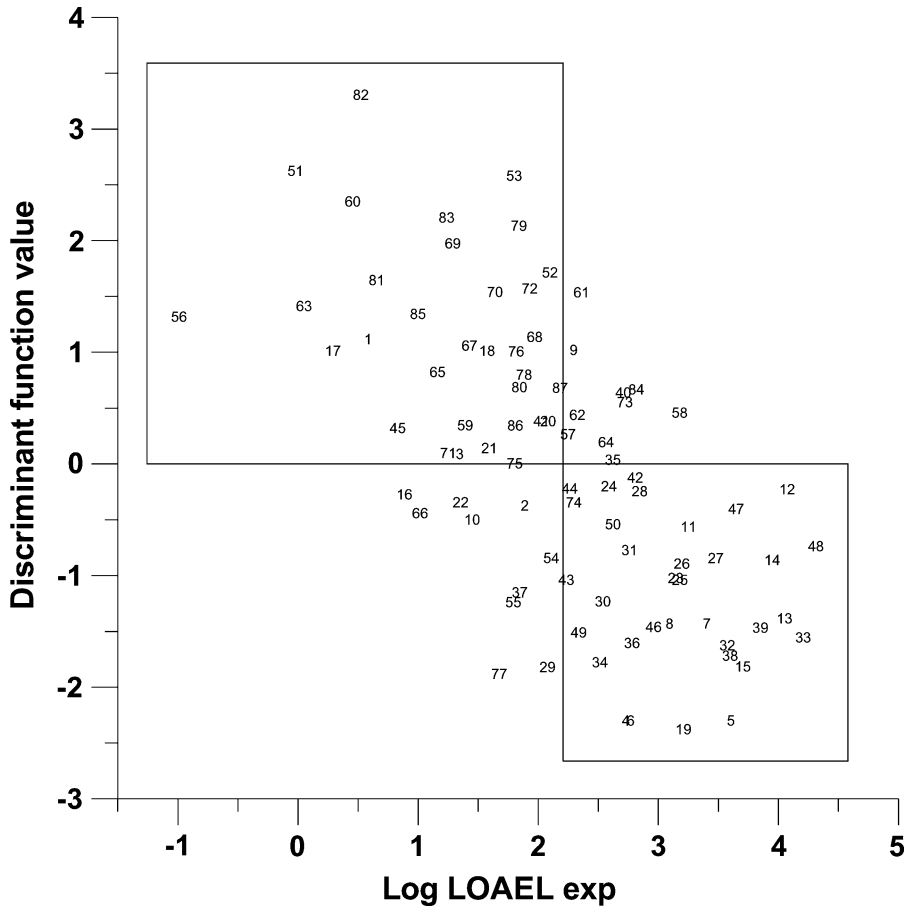

Figure 3. Plot of the true predictions using the 87 models described in the text. In abscissas the experimental property value is represented. In ordinates, the discriminant function value is used. The boxes delimit the correct classifications. See text for more details.

data displayed in Figure 1 comes from individual *true predictions* by molecule. The significance value computed from the F statistic is $p = 0.00004$ [24], and this small ratio gives the probability to obtain such arrangement when performing predictions.

As in the MLR case, the robustness of the 87 equations allowed us searching a unique training model involving all the 87 molecules, this model also selected the three parameters mentioned above:

$$f = 0.81453\, N_{el} + 1.28516^{3}\chi_p - 1.35274^{4}\chi_p - 4.71390$$

$n = 87$ (44 actives and 43 less actives), $\lambda = 0.598$, Mahalanobis distance between groups: 1.61

The classification ratios are listed in the last line of Table 6 (model labeled with 3′ descriptors).

Structure-activity

The training database comprised very structurally diverse compounds. Among others, unstable compounds can be pointed out as cyclonite, phenols, chlorinated phenols, or Michael acceptors. The main part of the training chemicals was monocyclic substances (38). The following group corresponded to acyclic ones (22), and the remainder were tricycles. This fact could explain why, for test compounds aldrin (tetracyclic) and dieldrin (pentacyclic), the predictions were not good. The α-endosulfan is also an outlier because norbornane structures are not represented in training set, making the number of topological paths greater in the molecule. The greatest residual is observed for di-(2-ethylhexyl)adipate, a structurally simple compound with no similar representatives in the training database. Another outlier is tralomethrin, the greatest molecule in the set that shows extreme values for the path and path-cluster connectivity indices and many other descriptors. Outliers were also maleic hydrazide and acrylamide, powerful Michael acceptors. Tetrachlorovinphos, that contains an enolephosphate, a unique structural fragment in the database, cannot be predicted from the two models.

Also, PRi and connectivity indices account mainly for the degree of molecular rigidity and polarisability. These are descriptors that figure in the best correlations.

Conclusions

Internal Test Set method constitutes a conceptually simple algorithm to obtain robust true predictions and to avoid model overfitting. Nevertheless, the equations obtained seem to give poor results, their predictive potential is balanced between the training and the test set. By contrast, complex traditional models often fail to provide better predictions outside the set used to obtain them, and this behavior may be irrespective of the attached statistical significance parameters.

Acknowledgments

The authors acknowledge financial aid to the grant number BQU2003-07420-C05 of the 'Ministerio de Ciencia y Tecnología' within the Spanish Plan Nacional I+D and the 'Red de Investigación de Centros de Enfermedades Tropicales', RICET (C03/04), Fondo de Investigación Sanitaria, Ministry of Health, Spain for financial support. Authors also thank the referees for their useful comments that have improved the present article. Some of them have been included in the revised version. Special thanks to Jennifer Chai-Chang for grammatical advice.

Electronic supplementary material

The following supplementary material is available in electronic format (at: http//dx.doi.org/10.1007/s11030-005-9007-z): Table with values for every computed descriptor for each compound, and corresponding intercorrelation matrix.

References

1. McNaught, A.D. and Wilkinson, A., *IUPAC Compendium of Chemical Terminology*, 2nd Edition Blackwell Science, 1997.
2. Pogliani, L., *From molecular connectivity indices to semiempirical connectivity terms: recent trends in graph theoretical descriptors*, Chem Rev., 100 (2000) 3827–3858.
3. Gozalbes, R., Doucet, J.P. and Derouin, F., *Application of topological descriptors in QSAR and drug design: history and new trends,* Curr. Drug Targets Infect. Disord., 2 (2002) 93–102.
4. Torrens, F., *A new topological index to elucidate apolar hydrocarbons,* J. Comput.-Aid. Mol. Des., 15 (2001) 709–719.
5. Besalú, E., Gironés, X., Amat, L. and Carbó-Dorca, R., *Molecular quantum similarity and the fundamentals of QSAR,* Acc. Chem. Res., 35 (2002) 289–295.
6. Golbraikh, A., Bonchev, D. and Tropsha, A., *Novel ZE-isomerism descriptors derived from molecular topology and their application to QSAR analysis*, J. Chem. Inf. Comput. Sci., 42 (2002) 769–787.
7. Tomovic, Z. and Gutman, I., *Modeling boiling points of cycloalkanes by means of iterated line graph sequences*, J. Chem. Inf. Comput. Sci., 41 (2001) 1041–1045.
8. Torrens, F., *Table of periodic properties of fullerenes based on structural parameters*, J. Chem. Inf. Comput. Sci., 44 (2004) 60–67.
9. Estrada, E., Patlewicz, G., Chamberlain, M., Basketter, D. and Larbey, S., *Computer-aided knowledge generation for understanding skin sensitization mechanisms: The TOPS-MODE approach*, Chem. Res. Toxicol., 16 (2003) 1226–1235.
10. Gálvez, J., Julián-Ortiz, J.V. de and García-Domenech, R., *General topological patterns of known drugs*, J. Mol. Graph. Model., 20 (2001) 84–94.
11. Murcia-Soler, M., Pérez-Giménez, F., García-March, F.J., Salabert-Salvador, M.T., Díaz-Villanueva, W. and Castro-Bleda, M.J., *Drugs and nondrugs: An effective discrimination with topological methods and artificial neural networks*, J. Chem. Inf. Comput. Sci., 43 (2003) 1688–1702.
12. Ivanciuc, O. and Klein, D.J., *Computing wiener-type indices for virtual combinatorial libraries generated from heteroatom-containing building blocks,* J. Chem. Inf. Comput. Sci., 42 (2002) 8–22.
13. Basak, S.C., Mills, D.R., Balaban, A.T. and Gute, B.D., *Prediction of mutagenicity of aromatic and heteroaromatic amines from structure: A hierarchical QSAR approach,* J. Chem. Inf. Comput. Sci., 41 (2001) 671–678.
14. Julián-Ortiz, J.V. de, García-Domenech, R., Gálvez, J. and Pogliani, L., *Predictability and prediction of lowest observed adverse effect levels in a structurally heterogeneous set of chemicals,* SAR QSAR Environ. Res., 16 (2005) 263–272
15. Mumtaz, M.M., Knauf, L.A., Reisman, D.J., Peirano, W.B., DeRosa, C.T., Gombar, V.K., Enslein, K., Carter, J.R., Blake, B.W., Huque, K.I. and Ramanujam, V.M.S., *Assessment of effect levels of chemicals from quantitative structure-activity relationship (QSAR) models. I. Chronic lowest-observed-adverse-effect level (LOAEL),* Toxicol Lett., 79 (1995) 131–143.
16. Venkatapathy, R., Moudgal, C.J. and Bruce, R.M., *Assessment of the oral rat chronic lowest observed adverse effect level model in TOPKAT, a QSAR software package for toxicity prediction*, J. Chem. Inf. Comput. Sci., 44 (2004) 1623–1629.
17. García-Domenech, R., DESCRI version 2003. Department of Physical Chemistry. University of Valencia, Spain. It is a home-made PC program for the calculation of 62 molecular structural invariants which accepts lists of MDL MOL files as inputs, freely available to academia upon request to the author (Ramon.Garcia@uv.es).

18. Hall, L.H., Molconn-Z (version 3.0) EastemNazaree College, Quincy, Massachusetts, USA.
19. BMDP New System 2.0., Statistical Solutions Ltd., Saugus, MA, USA, 2001.
20. Furnival, G.M. and Wilson, R.W., *Regresions by leaps and bounds,* Technometrics, 16 (1974) 499–511.
21. Livingstone, D.J. and Salt, D.W., *Judging the significance of multiple linear regression models*, J. Med. Chem., 48 (2005) 661–663.
22. Besalú, E., Regre v 1.57. Institute of Computational Chemistry. University of Girona. Spain, 2005.
23. García-García, A., Gálvez, J., Julián-Ortiz, J.V. de, García-Domenech, R., Muñoz, C., Guna, R. and Borrás, R., *Search of chemical scaffolds for novel antituberculosis agents*, J. Biomol. Screen., 10 (2005) 206–214.
24. Besalú, E. and Julián-Ortiz, J.V. de, *Equivalence of the Pecka-Ponec correlation probability and the statistical F significance for MLR models,* J. Math. Chem., 36 (2004) 361–363.
25. Gálvez, J., García-Domenech, R., Julián-Ortiz, J.V. de and Soler. R., *Topological approach to drug design*, J. Chem. Inf. Comp. Sci., 35 (1995) 272–284. Errata: J. Chem. Inf. Comp. Sci. 35 (1995) 938.
26. Kier, L.B., Murray, W.J., Randič, M. and Hall, L.H., *Molecular connectivity V: connectivity series concept applied to density,* J. Pharm. Sci., 65 (1976) 1226–1230.
27. Julián-Ortiz, J.V. de, Gálvez, J., Muñoz-Collado, C., García–Domenech, R. and Gimeno–Cardona, C., *Virtual combinatorial syntheses and computational screening of new potential anti-herpes compounds*, J. Med. Chem., 42 (1999) 3308–3314.
28. Kier, L.B. and Hall, L.H., *General definition of valence delta values for molecular connectivity*, J. Pharm. Sci., 72 (1983) 1170–1173.
29. Gálvez, J., García-Domenech, R., Salabert, M.T. and Soler, R., *Charge indexes. New topological descriptors*, J. Chem. Inf. Comp. Sci., 34 (1994) 520–525.
30. Hall, L.H. and Kier, L.B., *Electrotopological state indices for atom types: a novel combination of electronic, topological, and valence state information*, J. Chem. Inf. Comput. Sci., 35 (1995) 1039–1045.
31. Kier, L.B., *Indexes of molecular shape from chemical graphs*, Medicinal Research Reviews, 7 (1987) 417–440.