December 2008 Volume 4 Issue (9-11)

Trade Science Inc.

Organic CHEMISTRY

An Indian Journal

Full Paper

OCAIJ, 4(9-11), 2008 [470-474]

Application of molecular topology for the prediction of the toxicity of organic chemicals to *Chlorella vulgaris*

R.García-Domenech*, A.Villanueva, J.Galvez

Unidad de Diseno de Fármacos y Conectividad Molecular. Dpto. Química Física, Facultad de Farmacia, Universitat de Valencia, Avda. V.A. Estellés s/n, 46100-Burjassot, (SPAIN)

Tel. : 34963544291

E-mail : ramon.garcia@uv.es

Received: 14th October, 2008 ; Accepted: 19th October, 2008

ABSTRACT

In this paper a multilinear regression analysis has been carried out in order to look for a mathematical model capable to accurately predict the toxicity of *Chlorella vulgaris* of a set of organic chemicals. The structural description has been achieved through topological indices and the model was validated by a cross-validation test, an external validation test and a randomization test. The results confirm the model's capability to predict the analysed property.

KEYWORDS

Toxicity
Chlorella vulgaris;
Molecular topology;
Multilinear regression analysis.

INTRODUCTION

Toxicity is a property that assesses the degree of toxic or poisonous effects of a chemical compound. The toxicity may relate to the effect produced either on a superior organism, as for instance a human being or over a bacterium or a plant, or a substructure, such as a cell (cytotoxicity).

The emergence of diseases due to poisoning has made us understand the power that can exhibit the toxic compounds. In former times, the main toxic effects studied were those associated with death; however nowadays the toxicity is also assessed for other purposes as for example to combat pests and diseases, for disinfection and even for military endpoints.

Actually, about 28 million chemicals have been synthesized, out of which about 200.000 are sold and used daily. Nearly 3.000 new products are introduced annually by the chemical industry into the marketplace. From these, only 10.000 are based on the ownership of toxicity. Faced with this situation, companies marketing these products, need to conduct pilot studies to verify its use and then throw up for sale proceeds[1,2].

The quantitative structure activity relationships paradigm (QSAR) has been broadly used for chemical hazard assessment[3-5]. One of the most efficient QSAR method is that based on molecular topology[6] (MT) and the multilinear regression analysis.

One of the most significant features is that it has been shown that the toxicity in cellular organisms can be a good reflection of the extent of this property to superior mammalians. In fact, there are a lot of studies based on analysis of toxicity in animals such as fish, using the MT methodology[7,8].

To a lesser extent, experimental studies have been conducted with single-cell organisms such as algae, where the behaviour against the attack of toxic compounds we can yield good predictions. The use of

QSAR for such a goal has been depicted evident in several works[9,10] using as a single body to *Chlorella vulgaris*.

Our scientific aim here is focused at finding mathematical models for predicting the toxicity of organic compounds compared to single-cell organisms. To do that, we have used MT within a framework widely recognized and applied to the prediction of different properties[11-14].

MATERIALS AND METHODS

Analysed compounds

In this study a group of 91 organic compounds with information about toxicological assessment has been selected. Toxicity data[log(1/EC_{50}] (pC) were determined in a biochemical assay utilizing the unicellular algae *C.vulgaris* in the logarithmic phase of its growth cycle. All toxicological analyses were performed in a buffer solution with a pH of 6.9 and a temperature between 25 and 30^0C[10]. Assays were conducted following the protocol described by Worgan et al.[15]. TABLE 1 shows the CAS number, name and experimental toxicity, pC_{exp}, for each compound studied.

Molecular descriptors

A set of well-known topological indexes, TIs, was used in this work. Each compound was characterised by 62 TIs calculated with the aid of the DesMol1 program (available by e-mail request). 32 connectivity Randić-Kier-Hall type indices, $^m\chi_t$, and differences and quotients, ^mD_t and ^mC_t[16], 20 topological charge indices, G_m and J_m,[17] and other 10 discrete invariants[18].

Multilinear regression analysis, MLRA

MLRA was performed with the 2R and 9R modules of the BMDP software, which estimate regression equations for the best subsets of predictor variables and provides detailed residual analysis by using the Furnival-Wilson algorithm[19]. Equations with minimal Mallows C_p parameter were initially chosen[20].

The stability of the equation selected was evaluated through a cross-validation by the leave-one-out algorithm. To do this, one compound of the set is extracted, and the model is recalculated using as training set the remaining N-1 compounds. The property is then

TABLE 1: Chemical Abstracts Service (CAS) number, chemical name, experimental and calculated toxicity (log(1/EC_{5J0})(mM) (compounds listed in increasing order of toxicity to *C.vulgaris*)

CAS Nº	Name	*C. vulgaris*		
		PC_{exp}*	PC_{calc}**	Residual
67-56-1	Methanol	-4.06	-3.92	-0.14
64-17-5	Ethanol	-3.32	-3.52	0.20
75-65-0	2-Methyl-propan-2-Ol	-3.16	-2.67	-0.49
78-92-2	Butan-2-Ol	-2.98	-2.49	-0.49
868-77-9[b]	2-Hydroxyethyl methacrylate	-2.82	-2.09	-0.73
818-61-1	2-Hydroxyethyl acrylate	-2.79	-2.39	-0.40
96-33-3	Methyl acrylate	-2.75	-2.75	0.00
71-36-3	Butan-1-Ol	-2.73	-2.71	-0.02
78-93-3	Butanone	-2.51	-2.80	0.29
80-62-6[b]	Methyl methacrylate	-2.24	-2.53	0.29
96-22-0	Pentan-3-One	-2.23	-2.40	0.17
4170-30-3	Crotonaldehyde	-1.98	-2.14	0.16
6728-26-3	Trans-2-hexenal	-1.94	-1.38	-0.56
1576-87-0	Trans-2-pentenal	-1.88	-1.89	0.01
108-95-2[b]	Phenol	-1.46	-1.22	-0.24
96-05-9	Allyl methacrylate	-1.42	-1.95	0.53
62-53-3	Aniline	-1.34	-1.12	-0.22
110-43-0	2-Heptanone	-1.18	-1.53	0.35
100-66-3	Anisole	-1.09	-0.90	-0.19
367-12-4[b]	2-Fluorophenol	-1.08	-1.17	0.09
348-54-9	2-Fluoroaniline	-1.05	-1.07	0.02
108-39-4	3-Cresol	-1.01	-0.76	-0.25
150-76-5	4-Methoxyphenol	-0.97	-0.86	-0.11
95-55-6	2-Hydroxyaniline	-0.91	-0.99	0.08
90-05-1[b]	2-Methoxyphenol	-0.88	-0.84	-0.04
87-62-7	2,6-Dimethylaniline	-0.87	-0.15	-0.72
100-52-7	Benzaldehyde	-0.81	-1.00	0.19
95-48-7	2-Cresol	-0.81	-0.70	-0.11
90-02-8	2-Hydroxybenzaldehyde	-0.8	-0.74	-0.06
98-95-3[b]	Nitrobenzene	-0.78	-0.33	-0.45
950-37-8	Methidathion	-0.73	-0.57	-0.16
106-44-5	4-Cresol	-0.66	-0.75	0.09
95-65-8	3,4-Dimethylphenol	-0.65	-0.31	-0.34
104-87-0	4-Tolualdehyde	-0.65	-0.49	-0.16
94-71-3[b]	2-Ethoxyphenol	-0.62	-0.50	-0.12
24964-64-5	3-Cyanobenzaldehyde	-0.57	-0.73	0.16
99-08-1	3-Nitrotoluene	-0.5	0.00	-0.50
106-48-9	4-Chlorophenol	-0.42	-0.59	0.17
97-02-9	2,4-Dinitroaniline	-0.36	0.24	-0.60
106-41-2[b]	4-Bromophenol	-0.35	-0.14	-0.21
106-40-1	4-Bromoaniline	-0.33	-0.03	-0.30
108-42-9	3-Chloroaniline	-0.31	-0.60	0.29
2495-37-6	Benzyl Methacrylate	-0.21	-0.08	-0.13
618-87-1	3,5-Dinitroaniline	0.03	0.47	-0.44
89-98-5[b]	2-Chlorobenzaldehyde	0.06	-0.29	0.35
540-38-5	4-Iodophenol	0.16	0.21	-0.05
4748-78-1	4-Ethylbenzaldehyde	0.16	-0.08	0.24
58-27-5	2-Methyl-1,4-naphthoquinone	0.16	0.65	-0.49
88-69-7	2-Isopropylphenol	0.17	0.32	-0.15
626-43-7[b]	3,5-Dichloroaniline	0.24	-0.14	0.38
603-71-4	1,3,5-Trimethyl-2-nitrobenzene	0.25	0.62	-0.37
608-31-1	2,6-Dichloroaniline	0.26	-0.03	0.29
88-18-6	2-Tert-Butyl phenol	0.29	0.50	-0.21

Countinue in next page

Full Paper

CAS Nº	Name	*C. vulgaris*		
		pC_{exp}**	pC_{calc}**	Residual
95-50-1	1,2-Dichlorobenzene	0.37	-0.19	0.56
99-65-0 [b]	1,3-Dinitrobenzene	0.38	0.26	0.12
51-28-5	2,4-Dinitrophenol	0.4	0.15	0.25
100-25-4	1,4-Dinitrobenzene	0.41	0.27	0.14
99-61-6	3-Nitrobenzaldehyde	0.45	-0.12	0.57
732-11-6	Phosmet	0.47	0.91	-0.44
298-00-0 [b]	Methylparathion	0.6	0.96	-0.36
121-75-5	Malathion	0.64	0.13	0.51
99-30-9	2,6-Dichloro-4-nitroaniline	0.64	0.80	-0.16
86-50-0	Methyl azinphos	0.69	1.55	-0.86
121-14-2	2,4-Dinitrotoluene	0.7	0.44	0.26
2636-26-2 [b]	Cyanophos	0.79	0.77	0.02
3531-19-9	6-Chloro-2,4-dinitroaniline	0.8	0.83	-0.03
99-28-5	2,6-Dibromo-4-nitrophenol	0.81	1.60	-0.79
640-15-3 [a]	Thiometon	0.94	-0.54	1.48
89-61-2	2,5-Dichloronitrobenzene	0.97	0.75	0.22
94-62-2 [b]	Piperine	0.97	1.87	-0.90
939-97-9	4-Tert-butylbenzaldehyde	1	0.63	0.37
634-93-5	2,4,6-Trichloroaniline	1.11	0.46	0.65
83-42-1	2-Chloro-6-nitrotoluene	1.17	0.67	0.50
5388-62-5	4-Chloro-2,6-dinitroaniline	1.19	0.96	0.23
528-29-0 [b]	1,2-Dinitrobenzene	1.23	0.67	0.56
100-00-5 [a]	1-Chloro-4-nitrobenzene	1.25	0.07	1.18
2463-84-5	Dicapthon	1.36	1.42	-0.06
128-37-0	2,6-Di-tert-butyl-4-methyl Phenol	1.45	1.95	-0.50
3481-20-7	2,3,5,6-Tetrachloroaniline	1.48	1.03	0.45
609-89-2 [b]	2,4-Dichloro-6-nitrophenol	1.5	0.82	0.68
83-38-5 [a]	2,6-Dichlorobenzaldehyde	1.5	0.35	1.15
55-38-9	Fenthion	1.56	1.24	0.32
96-76-4	2,4-Di-Tert-butylphenol	1.6	1.33	0.27
87-86-5	Pentachlorophenol	1.69	1.45	0.24
122-14-5 [b]	Fenitrothion	1.71	1.20	0.51
89-69-0	1,2,4-Trichloro-5-nitrobenzene	1.88	1.10	0.78
6284-83-9	1,3,5-Trichloro-2,4-dinitrobenzene	1.89	1.32	0.57
1689-82-3	Phenylazophenol	2.16	1.80	0.36
	4-(Dibutylamino) benzaldehyde	2.18	1.53	0.65
117-18-0 [b]	2,3,5,6-Tetrachloronitrobenzene	2.34	1.77	0.57
608-71-9	Pentabromophenol	3.1	3.73	-0.63

*** Experimental values obtained from Ref. [10]; ** Calculated values from selected topological model; [a] outliers compounds; [b] External test.**

predicted for the removed element. This process is repeated for all the compounds of the set so obtaining a prediction for every one. This procedure also aids in the detection of outlying points[21].

In order to evidence the possible existence of fortuitous regressions, the *randomization test* is adopted in this paper. Thus, the values of the property of each

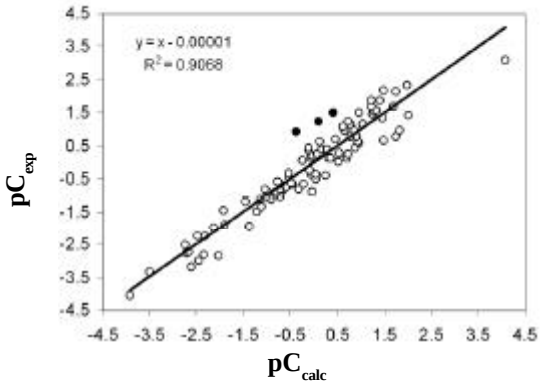

Figure 1: Graphic representation of the pC_{exp} versus pC_{calc} from the topological model selected (outliers compounds: black points)

compound are randomly permuted and linearly correlated with the aforementioned descriptors. This process is repeated as many times as needed. The usual way to represent the results of a randomization test is plotting the correlation coefficients versus predicted ones, r^2 and q^2 respectively.

RESULTS AND DISCUSSION

When working with groups of compounds structurally heterogeneous, as is our case, it is easy to unveil the presence of outliers. For making that, the use of the 2R module from the BMDP software is particularly efficient to figure out, as a first approximation, the most significant variables for predicting the toxicity.

In our case, the outcome resulted in five variables, ${}^0\chi^v$, $G_1{}^v$, $G_5{}^v$, J_4 and V_3, with a variance $r^2 = 0.9068$.

Figure 1 shows the plot of experimental versus calculated pC values for each compound. The compounds labelled with black dots clearly appear as

TABLE 2: Topological model selected with toxicity, pC, through MLRA($pC = -4{,}4937 + 0{,}5679\ {}^0\chi^v - 0{,}1131\ G_1{}^v - 1{,}1609\ G_5{}^v + 10{,}0710\ J_4 + 0{,}1881\ V_3$)

Variables	Standard error	p
Intercept	0.1702	0.0001
${}^0\chi^v$	0.0408	0.0001
$G_1{}^v$	0.0227	0.0001
$G_5{}^v$	0.2353	0.0001
J_4	1.359	0.0001
V_3	0.0241	0.0001

N= 70, SEE = 0.4048, r = 0.9622, p< 0.0001, F= 180, q^2 = 0.9383, r^2 = 0.9282 , Cp Mallow = 6

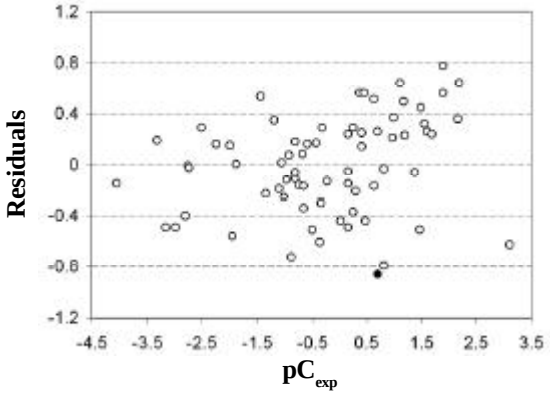

Figure 2: Graphic representation of the residuals versus pC_{exp} obtained with the topological model selected

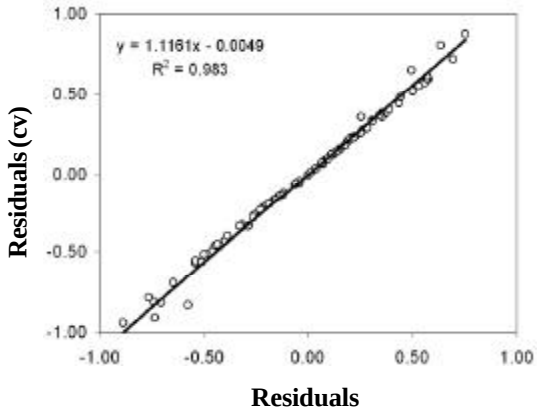

Figure 3: Graphic representation of the residuals obtained in the cross-validation versus the residuals obtained with the topological model selected

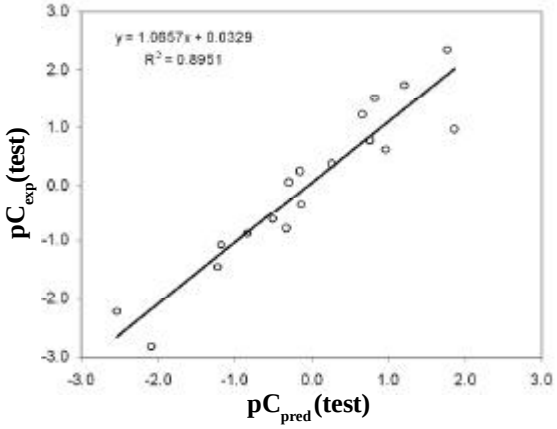

Figure 4: Graphic representation of the pC_{exp} versus pC_{pred} for the external test

outliers. Namely thiometon, 1-chloro-4-nitrobenzene and 2,6-diclorobenzaldehide display values of standard error of estimation above ± 2SEE. These three

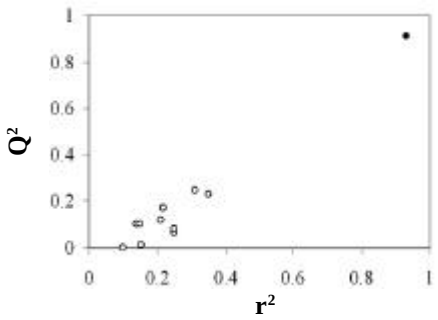

Figure 5: Graphic representation of the prediction coefficient ,Q^2, versus correlation coefficient, r^2, obtained by randomization study

molecules were removed from the training group and the 88 remaining compounds were hence left for the predictive analysis.

The topological indices selected show a relatively poor intercorrelation (r <0.700), although they all are statistically significant for predicting the property studied. The index $^0\chi^v$, would take into account of topological and structural aspects related to molecular volume[22], whereas G_1^v, G_5^v and J_4 would profile the intramolecular charge transfer responsible for the value of the toxicity of each compound and V_3 the presence of multiple bonds in the molecule.

Selection of the best prediction regression function

After removing the three causing a loss of predictive capability, we applied the module 9R of the BMDP software, using a training set of 70 compounds and leaving out at random the 18 remaining compounds as a test group.

TABLE 2 shows the selected function together with the associated statistical information. As can be seen, all the indexes therein show a statistical significance above 99.9% (p <0.001) (see TABLE 2 column 3). Furthermore, the selected function explains over 92% of the variance (r^2=0.9282).

Figure 2 shows the graphic representation of the residual versus pC_{exp} obtained with the topological model selected. The toxicity of every compound was predicted in a satisfactory manner (see TABLE 1) except for azinphos methyl, with an residual slightly higher than ± 2SEE.

Randomization and predictive ability tests

The cross-validation analysis of the training group,

shows similar result to the obtained in the analysis of multilinear regression ($q^2 \approx 0900$) what demonstrates the stability of the selected function, see figure 3.

To complete the validation of the predictive model, we extracted from the matrix of data (N = 88), 18 compounds, actually 20% of the overall group. TABLE 1 (black mark) and figure 4 show the values obtained for the external test group. It has obtained a prediction coefficient $q^2 = 0.8951$ bringing us confirmation of the good performance of the selected function.

Finally, a study of randomness was carried out, in order to evidence the possible existence of fortuitous regressions. The values of the pC_{exp} of each compound were randomly permuted and linearly correlated with the aforementioned descriptors. The process was repeated ten times. Figure 5 shows the graphic representation of the prediction coefficient, q^2, versus correlation coefficient, r^2, obtained in this study. In all cases, the values of q^2 were below 0.5 (the black point in figure 5 belongs to selected model), therefore, the selected prediction equation is not fortuitous.

CONCLUSIONS

Toxicity is a biological property whose assessment is more and more important nowadays. In our work, we have tried to obtain a mathematical function capable to predict the toxicity by means of topological indices. Initially, we performed a literature search of possible databases to provide a reliable, commercial and transferred to the experimental world. Following the topological indices more suitable for predicting the toxicity and the possible emergence of outliers, were selected using standard and well proven algorithms. The selection of five topological indices and the emergence of three outliers was the first outcome of the study.

Altogether, a mathematical model with five variables enable to estimate the toxicity with a $r^2 > 0.92$ in a group including 70 training compounds. The validation of the model was conducted with the help of a cross-validation leave-one-out and an external test

ACKNOWLEDGMENTS

We thank the Fondo de Investigación Sanitaria, Ministerio de Sanidad, Spain (project SAF2005–PI052128) for support of this work.

REFERENCES

[1] M.Cronin, J.Jaworska, J.Walker, M.Comber, C. Watts, A.Woth Environ. Health Perspect., **111**, 1391 **(2003)**.
[2] J.Walker, J.Mol.Struc.Theochem, **622**, 167 **(2003)**.
[3] R.Garcia-Domenech, P.Alarcón-Elbal, G.Bolas, R. Bueno-Mari, F.A.Chorda-Olmos, S.A.Delacour, M.C.Mourino, A.Vidal, J.Galvez, SAR and QSAR in Environ.Res., **18**, 745 **(2007)**.
[4] J.M.Luco, J.Gálvez, R.Garcia-Domenech, J.V.de Julian-Ortiz, Mol Divers., **8(4)**, 331 **(2004)**.
[5] R.Garcia-Domenech, J.V.de Julián-Ortiz, M.J. Duart, J.M.Garcia-Torrecillas, G.M.Anton-Fos, I. Rios-Santamarina, C.de Gregorio-Alapont, J. Galvez, SAR and QSAR in Environ.Res., **12(1-2)**, 237 **(2001)**.
[6] A.A.Lagunin, A.V.Zakharov, D.A.Filimonov, V.V. Poroikov, SAR and QSAR Environ.Res., **18(3-4)**, 285 **(2006)**.
[7] C.L.Russom, S.P.Bradbury, S.J.Broderius, D.E. Hammermeister, R.A.Drummond; Environ.Toxicol., **16**, 948 **(1997)**.
[8] M.T.D.Cronin, J.C.Dearden, A.J.Dobbs; Sci.Total Environ., **109-110**, 431 **(1991)**.
[9] A.D.P.Worgan, J.C.Dearden, R.Edwards, T.I. Netzeva, M.T.D.Cronin; QSAR Comb.Sci., **22**, 204 **(2003)**.
[10] M.T.D.Cronin, T.I.Netzeva, J.C.Dearden, R. Edwards, A.D.P.Worgan; Chem.Res.Toxicol., **17**, 545 **(2004)**.
[11] R.Garcia-Domenech, J.Gálvez, J.V.de Julian-Ortiz, L.Pogliani; Chemical Reviews, **108(3)**, 1127 **(2008)**.
[12] R.Garcia-Domenech, J.V.Julian-Ortiz, E.Besalu; Molecular Diversity, **10(2)**, 159 **(2006)**.
[13] J.V.de Julián-Ortiz, R.Garcia-Domenech, J.Gálvez, L.Pogliani; SAR and QSAR in Environ.Res., **16(3)**, 263 **(2005)**.
[14] M.J.Duart, G.M.Antón-Fos, J.Galvez, R.García-Domenech; Chemistry an Indian Journal, **1(1)**, 67 **(2003)**.
[15] T.I.Netzeva, A.D.Worgan, J.Dearden, C.Edwards, M.T.D.Cronin; J.Chem.Inf.Comput.Sci., **44(1)**, 258 **(2004)**.
[16] L.B.Kier, L.H.Hall; J.Pharm.Sci., **72**, 1170 **(1983)**.
[17] J.Gálvez, R.García, M.T.Salabert, R.Soler; J.Chem.Inf.Comput.Sci., **34(3)**, 520 **(1994)**.
[18] J.V.De Julian-Ortiz, R.García-Domenech, J.Galvez, R.Soler, F.J.Garcia-March, G.M.Anton-Fos; J.Chromatogr.A, **719**, 37 **(1996)**.
[19] G.M.Furnival, R.Wilson; Technometrics, **16**, 499 **(1974)**.
[20] C.L.Mallows; Technometrics, **15**, 661 **(1973)**.
[21] E.Besalu; J.Math.Chem., **29**, 191 **(2001)**.
[22] J.Gálvez; J.Chem.Inf.Comput.Sci., **43**, 1231 **(2003)**.