QSAR Study Concerning Toxicity and Threshold Limit Value VL_8 of Chlorine Containing Compounds

ROXANA DONE[1]*, GEORGIANA MÂNDRILA[1], LASZLO TARKO[2]*

[1] University of Bucharest, Department of Physical Chemistry, Faculty of Chemistry, 4-12 Regina Elisabeta Blvd., 030018, Bucharest, Romania

[2] Romanian Academy, Organic Chemistry Center "C.D.Nenitzescu", 202B Spl. Independentei,060023, Bucharest, Romania

The present paper continues precedent works in an attempt to develop methods for estimation of threshold limit value-time weighted average VL_8 for hazardous chemicals in the frame of QSAR studies. The dependent property (toxicity) is defined as $T = \log (k/VL_8)$. A calibration set, with a known value of the dependent property, and a prediction set containing 36 and 17 chlorine containing molecules respectively were used. The analysed molecules were virtually builded using molecular mechanics program PC Model, then optimized with quantum semi-empirical method AM1. For each molecule over 1700 descriptors were calculated using programs MOPAC, PRECLAV and DRAGON. Statistical computations are realized with PRECLAV program. High outlier molecules in calibration set were identified by a specific procedure. The quality of QSAR, in absence of outlier molecule (Epichlorohydrin), is satisfactory ($s = 0.4194$, $r^2 = 0.9060$, $F = 46.6$, $r^2_{cv} = 0.8703$). The toxicity of analysed molecules is directly proportional to weight percent of chlorine and inversely proportional to weight percentage of fluorine. In addition, the toxicity depends on shape and size of molecule. The proposed procedure, based on chemical similarity of prediction set molecules, "statistical outliers" in calibration set and other molecules in calibration set, seems to be useful in identification of the "outlier" molecules in prediction set. In prediction set, some polychlorinated biphenyls, polychlorinated naphthalenes and 2-chloro-phenol (with unknown observed toxicity value) were identified as "high" toxic compounds.

Keywords: QSAR, TLV toxicity, halogenated compounds

American Conference of Governmental Industrial Hygienists (ACGIH), organization of professionals in governmental agencies or educational institutions engaged in occupational safety and health programs, developed and recommended occupational exposure limits for almost 640 chemical substances and physical agents. The threshold limit values (TLV's) are established for the average person equating to a 150 pounds, male, age 25-44. TLV's are general reviewed and updated annually. There are different types of TLV's [1, 2]. The most used TLV is threshold limit value-time weighted average (TLV-TWA) which in Romanian Norms for Work Protection, adopted in 2002 [3] in accordance with U.E. norms is abbreviated as VL_8. This is defined as concentration for a normal 8 - hour work day and a 40 – hour work week to which worker may be repeatedly exposed, day after day, without adverse effects.

The task of evaluating all existing chemicals far exceeds the capacity of toxicology profession world wide [4]. Because chlorinated compounds are frequently implied in different occupational exposures [5] we are interested to develop QSAR studies concerning VL_8 evaluation.

Methods and formulas

The analysed dependent property (toxicity T) is defined as

$$T = \log (k / VL_8) \qquad (1)$$

where VL_8 is expressed in mg/m^3 for particulate and in ppm for gaseous pollutants and constant k = 5000. VL_8 values are extracted from [3].

In present work was used a calibration set with a known value of toxicity T containing 36 and a prediction set of 17 chlorine containing molecules. Both sets are presented in table 1. In this table the observed values of toxicity in prediction set are indicated with "?" sign. The 53 analysed molecules are virtually builded using the PC Model molecular mechanics program [6]. After identification of the minimum energy conformer, a more rigorous geometry optimization by quantum semi-empirical AM1 method [7, 8], included in MOPAC program [9], was made. For each molecule were calculated over 1700 descriptors using MOPAC, PRECLAV [10] and DRAGON [11] programs. Set of descriptors includes parabolic functions of "whole molecule descriptors", calculated by PRECLAV. Identification of the "significant" descriptors uses specific criterion [12]. Also were identified "significant" molecular fragments by a specific procedure [13]. The weight percentage of these fragments, in the molecules in the calibration set presents a good correlation with toxicity values. The statistical computations have been realized with PRECLAV program. In the toxicity definition (1) constant k is the maximum observed value of VL_8 [3] in the used calibration set (k = 5000). Therefore, all the observed values of T, used in calculations, were positive and minimum T value was zero. We obtained tens thousands multilinear QSARs for toxicity.

$$T = c_0 + \sum_{i=1}^{k} c_i \cdot d_i \qquad (2)$$

where:

T is toxicity;

c_0 - intercept;

c_i - weighting factors ;

d_i - (values of) descriptors;

k - number of descriptors.

Quality of equations was computed using a specific formula [15]. It is needed to mention that program PRECLAV verified if the calibration set is a representative sample in the calibration set + prediction set bulk. In the

* *email: rdone2006@yahoo.com and ltarko@cco.ro*

QSAR practice one consider *a priori* this condition is respected. According to PRECLAV, this validation influences the selection of significants descriptors. The selection is very much influenced by the presence in calibration and prediction set of the *outlier* molecules. The QSAR equation obtained in presence of prediction set is different from that obtained in its absence. Identification of the high outliers (from statistical and structural point of view) in the prediction set is a problem yet unsolved in the QSAR practice.

Usually, the prediction set is constituted so, that is considered without statistical or chemical outliers. Here, the structural diversity of prediction set molecules is visible much more larger than structural diversity of calibration set. In our case, majority of calibration set molecules contain only hydrogen, carbon and chlorine atoms, but in the prediction set situation is very different. The presence of outliers in prediction set is probable. We considered that outliers molecules included in prediction set are:

- similar, from chemical structure point of view, with very few other molecules of calibration/prediction set;
- similar from structural point of view with molecules identified as statistical outliers in calibration set.

The high statistical outlier molecules in the calibration set were identified by a specific procedure [15]. Evaluation of chemical similarity was realized using a previously described procedure [13]. A certain group of molecules with high chemical similarity (cluster) can include high statistical outlier molecules of calibration set, other molecules of calibration set and molecules of the prediction set. If, within analysed cluster, sum percentages of statistical outliers molecules of calibration set + prediction set molecules is over 60%, then the molecules of prediction set are considered outliers of prediction set. This empirical procedure of outliers' identification in prediction set considers the above conditions a) and b). Utility of predictors (i.e. the descriptors present in QSAR equation having the highest predictive power) was calculated using a specific formula [14].

Results and discussions

The quality of equation obtained in presence of high outlier(s) is satisfactory:

$$s = 0.5340,\ r^2 = 0.8484,\ F = 28.0,\ r^2_{cv} = 0.8024$$

The molecule identified as outlier in the calibration set is Epichlorohydrin (**28** in table 1). Twenty molecular fragments were identified, but, in absence of outlier

Table 1
OBSERVED AND COMPUTED TOXICITY VALUES OF CHLORINATED COMPOUNDS

No.	Name	CAS	VL_8	T_{obs}	T_{calc}
1	Ethyl chloride	75-00-3	1000	0.699	1.259
2	*Iso*-propyl chloride	75-29-6	400	1.097	1.527
3	Dichloro-methane	75-09-2	174	1.458	1.261
4	1,1-Dichloro-ethane	75-34-3	405	1.092	1.348
5	1,2-Dichloro-ethane	107-06-2	30	2.222	1.640
6	1,2-Dichloro-propane	78-87-5	100	1.699	2.175
7	Chloroform	67-66-3	10	2.699	2.022
8	1,1,1,2-tetrachloro-ethane	630-20-6	20	2.398	1.764
9	Pentachloro-ethane	76-01-7	40	2.097	2.240
10	Hexachloro-ethane	67-72-1	5	3.000	2.922
11	Lindane	608-73-1	0.3	4.222	4.188
12	Chloro-bromo-methane	74-97-5	700	0.854	1.075
13	Vinyl chloride	75-01-4	7.77	2.809	2.429
14	Allyl chloride	107-05-1	3	3.222	3.334
15	Metallyl chloride	563-47-3	80	4.135	3.987
16	1,2-dichloro-ethene	540-59-0	200	1.398	2.105
17	Tetrachloro-ethylene	127-18-4	50	2.000	2.592
18	Chloro-benzene	108-90-7	47	2.027	1.546
19	2-chloro-toluene	95-49-8	150	1.523	1.746
20	1,2-dichloro-benzene	95-50-1	122	1.613	1.607
21	Trichloro-toluene	98-07-7	2	3.398	2.968
22	1,2,4-trichloro-benzene	120-82-1	15.1	2.520	2.207
23	Hexachloro-benzene	118-74-1	0.5	4.000	4.180
24	Thionyl chloride	7719-09-7	15	2.523	2.200
25	2-Chloro-aniline	95-51-2	0.22	4.357	4.352
26	2,2 –Dichloro-N,N-di-2-propenyl-acetamide	37764-25-3	7	2.854	2.890
27	Bis-(2-Chloro-ethyl) ether	111-44-4	40	2.097	2.534
28	Epichlorohydrin	106-89-8	1	3.699	outlier
29	Aldrin	309-00-2	0.2	4.398	4.590
30	Chloro-difluoro-methane	75-45-6	3600	0.143	0.263

31	Dichloro-difluoro-methane	75-71-8	2000	0.398	-0.018
32	1,2-Dichloro-tetrafluoro-ethane	75-43-4	42	2.076	1.058
33	1,2-Dichloro-1,1,2,2-tetrafluoro-ethane	76-14-2	3000	0.222	0.224
34	Endrin	72-20-8	0.03	5.222	5.106
35	Trichloro-fluoro-methane	75-69-4	4000	0.097	0.578
36	1,1,2-trichloro-1,2,2-trifluoro-ethane	76-13-1	5000	0.000	0.670
**	*******************************	********	*****	******	******
37	Hexachloro-butadiene	87-68-3	?	?	2.890
38	Ethyl-chloro-thioformate	2941-64-2	?	?	1.023
39	2,4,6-trichloro- s-Triazin	108-77-0	?	?	1.195
40	p-nitro-chloro-benzene	100-00-5	?	?	outlier
41	4-Chloro-1,3-dinitro-benzene	97-00-7	?	?	outlier
42	bis-Chloro-methyl ether	542-88-1	?	?	outlier
43	Chloro-acetic acid	79-11-8	?	?	outlier
44	Chloro-acetaldehyde	107-20-0	?	?	0.702
45	2-Chloro-acetophenone	532-27-4	?	?	outlier
46	1,1, -dichloro-pinacolone	22591-21-5	?	?	2.692
47	2,2',4' Trichloro biphenyl	37680-66-3	?	?	3.261
48	PCB 99	38380-01-7	?	?	3.652
49	Ethyl chloro-formate	541-41-3	?	?	outlier
50	α-chloro-*iso*-butiraldehyde	917-93-1	?	?	1.358
51	1,5-dichloro-naphthalene	1825-30-5	?	?	3.176
52	1,3,5,7-tetrachloro-naphthalene	53555-64-9	?	?	3.548
53	2-Chloro-phenol	95-57-8	?	?	3.168

Table 2
THE "SIGNIFICANT" MOLECULAR FRAGMENTS

Fragment	Pearson correlation	Fisher function F
F	- 0.5482	146.1
Cl-C=C-Cl	0.4771	100.2

molecule, only two are "significant" molecular fragments, presented in table 2.

It is obviously, from the table 2, that toxicity for the molecules in the calibration set is directly proportional to weight percentage of fragment Cl-C=C-Cl, which is present only within the **29** (Aldrin) and **34** (Endrin) molecules. The toxicity is inversely proportional to weight percentage of fluorine. Other molecular fragments are "non-significant".

The maximum quality QSAR equation, obtained in absence of outlier molecule is:

$$T=0.3652 + 0.8580 \cdot D_1 + 0.7372 \cdot D_2 - 1.5754 \cdot D_3 + 0.8030 \cdot D_4 + 0.6486 \cdot D_5 + 1.0360 \cdot D_6 \quad (3)$$

where:

D_1 - is the Burden descriptor weighted by atomic electronegativities [16, 17];

D_2 – is the parabolic function the "variation coefficient of the sum of absolute values of net charges";

D_3 - is the number of multiple topologic paths of second order;

D_4 - is maximum value of nucleophilic reactivity index for carbon atoms;

D_5 - is the number of chlorine atoms;

D_6 – the capacity to form hydrogen bonds [18] weighted by molecular mass;

Quality of equation (3):

$s = 0.4194$, $r^2 = 0.9060$, $F = 46.6$, $r^2_{cv} = 0.8703$.

Minimum correlation of predictors with toxicity: for D_6 ($r^2 = 0.2023$)

Maximum intercorrelation of predictors: for pair D_1/D_3 ($r^2 = 0.6028$)

Utility of the predictors in description of toxicity is very high for D_4 and D_5, high for D_3 and moderate for D_1, D_2 and D_6.

The algebraic sign of coefficients indicates a direct proportional or inverse proportional influence of predictors to the toxicity. Physical meaning of DRAGON descriptors D_1 and D_3 is difficult to interpret. The toxicity values for the molecules in calibration set, computed in absence of statistical outlier molecule and in absence of prediction set molecules, are presented in the sixth column of the table 1. The molecules **40-43**, **45** and **49** are identified as outliers in the prediction set. After elimination of the outlier molecules from both calibration and prediction set, the used software verify the "representative sample" character of the new calibration set (36 molecules) within the new database (36 + 11 molecules). From point of view of now identified "significant" descriptors, the condition of representative sample is respected. However, this character is not very high. In fact the predictive power of the new obtained QSAR equation (in presence of prediction set) is low enough ($s = 0.7170$, $r^2 = 0.7252$, $F = 28.1$ $r^2_{cv} = 0.6640$). The toxicity values for the molecules in prediction set, calculated in absence of all outliers and in the presence of the prediction set, are presented in the last column of table 1, rows 37-53.

The values of toxicity for the outlier molecules in the prediction set should be estimated using other calibration set, containing molecules with higher chemical similarity with those.

Table 3
PREDICT VL_8 VALUES

No.	log VL_8	VL_8	No.	log VL_8	VL_8
37	0.80897	6.441	46	1.00697	101.6
38	2.67597	474.2	47	0.43797	2.741
39	2.50397	319.1	48	0.04697	1.140
40	outlier	outlier	49	outlier	outlier
41	outlier	outlier	50	2.34097	219.3
42	outlier	outlier	51	0.52297	3.334
43	outlier	outlier	52	0.15097	1.416
44	2.99697	993.0	53	0.53097	3.396
45	outlier	outlier			

According to certain statistical criterion [19], the estimated toxicity for molecules **47, 48, 51, 52** and **53** can be considered "high" (in comparison with other molecules of the prediction set). The estimated toxicity for molecules **38, 39, 47** and **50** can be considered "low" (also by comparison with the other molecules of the prediction set).

Finally, using toxicity computed values for the molecules of the prediction set and equation (1), we can write:

$$\log VL_8 = 3.69897 - T \quad (4)$$

According to equation (4) the VL_8 values are given in table 3.

Conclusions

The quality of QSAR, in absence of outlier molecule (Epichlorohydrine), is satisfactory ($s = 0.4194$, $r^2 = 0.9060$, $F = 46.6$, $r^2_{cv} = 0.8703$).

The toxicity of analysed molecules is directly proportional to weight percentage of chlorine and inversely proportional to weight percentage of fluorine. In addition, the toxicity depends on shape and size of molecule.

The proposed procedure, based on chemical similarity of prediction set molecules, "statistical outliers" in calibration set and other molecules in calibration set, seems to be useful in identification of the "outlier" molecules in prediction set.

In prediction set, some polychlorinated biphenyls, polychlorinated naphthalenes and 2-chloro-phenol (with unknown observed toxicity value) were identified as "high" toxic compounds.

References

1. STĂNCULESCU, I., DONE, R., MANDRAVEL, C., Proc. 32nd A.R.A. Congress, Boston, USA, July 22-24, 2008, p.615
2. MÂNDRILĂ, G., DONE, R., CHIOSA, V., Anal. Univ. Buc., **XVII**, 1, 2008, p. 57
3. SILION, I., CORDUNEANU, C., Bazele medicinii muncii, Ed. PIM, Iaşi, 2003, p.16, 326
4. SMITH, J.S., MENDELOFF, J. M., Risk Analysis, **19**, 1999, p.1223
5. CHIOSA, V., STĂNCULESCU, I., MANDRAVEL, C., Evaluarea toxicităţii poluanţilor atmosferici, Ed. Universităţii Bucureşti, Bucureşti, 2007, p.140
6. PC Model program is available from Gilbert, K.; Gajewski, J.J., Serena Software, Box 3076, Bloomington, IN, USA; http://www.serenasoft.com/index.html
7. DEWAR, M. J. S. et al., J. Am. Chem. Soc., **107**, 1985, p. 3902.
8. DEWAR M. J. S., ZOEBISCH, E. G., Theochem, **180**, 1988, p. 1
9. MOPAC program is available from Stewart, J.J.P., 15210 Paddington Circle, Colorado Springs, CO 80921; E-mail: MrMOPAC@OpenMOPAC.net; http://www.openmopac.net/
10. PRECLAV program is available from Tarko, L., Center of Organic Chemistry (CCO) – Bucharest, Romanian Academy, E-mail: ltarko@cco.ro
11. DRAGON program is available from Todeschini, R., Talete srl., via V. Pisani, 13-20124, Milano, Italy, website http://www.talete.mi.it
12. TARKO, L., Rev. Chim. (Bucureşti), **48**, no. 6, 1997, p. 676
13. TARKO, L., Rev. Chim. (Bucureşti), **55**, no. 7, 2004, p. 539
14. TARKO, L., Rev. Chim. (Bucureşti), **56**, no. 6, 2005, p. 639
15. TARKO, L., Rev. Chim. (Bucureşti), 2009, in press
16. PEARLMAN, R. S., SMITH, K. M., Novel Software Tools for Chemical Diversity in 3D QSAR in Drug Design - Vol. 2, Ed. Kubinyi – Folkers - Martin, Kluwer/ESCOM, Dordrecht (The Netherlands), 1998, p. 339
17. BURDEN, F. R., J. Chem. Inf. Comput. Sci., **29**, 1989, p. 225
18. TARKO, L., Rev. Roum. Chim., **51,** 2006, p. 463
19. TARKO, L., Rev. Chim. (Bucureşti), **57**, no. 10, 2006, p. 1014

Manuscript received: 13.03.2009