STUDIA UBB CHEMIA, LX, 4, 2015 (p. 193-200)
(RECOMMENDED CITATION)

QSAR STUDY ON DIOXINS

RALUCA MĂTIEŞ[a], BEATA SZEFLER[b] and MIRCEA V. DIUDEA[a*]

ABSTRACT. This paper presents a QSAR study realized on a set of 40 dioxins, known as pollutants, substances that are toxic for the environment. The study is based on the hypermolecule approach and on the prediction by similarity clustering. The results show a good modeling of logP parameter with the correlation weighted descriptor and some topological indices derived from Cluj matrices and also with the calculated HOMO energy level for the set of studied molecules.

Keywords: *dioxin, hypermolecule, QSAR, Cluj descriptors*

INTRODUCTION

Dioxins are unwanted pollutants in the environment, occurring in industrial processes (incineration pulp and paper bleaching with chlorine), also in the manufacture of pesticides, fungicides or herbicides [1].

The term dioxin refers to dibenzo-p-dioxins (PCDD), polychlorinated dibenzofurans (PCDF) and coplanar polychlorinated biphenyls (PCBs) which show similar biological and toxicological properties. These compounds are contaminants of lipophilic fat and concentrate in biological systems.

PCDDs/ PCDFs and PCBs have toxic effects on the nervous system, immune, endocrine and reproductive systems. International Agency for Research on Cancer has classified 2,3,7,8-TCDD as the most toxic congener of polychlorinated dibenzo-p-dioxins and classified in Group 1 carcinogen to humans [2-3].

Dioxins are odorless and colorless organic compounds, insoluble in water but soluble in fat. These compounds have in their structure carbon, hydrogen, oxygen and chlorine. Dioxins are biodegradable, but are persistent and bio-accumulates in foods [4].

[a] *Department of Chemistry, Faculty of Chemistry and Chemical Engineering, Babeş-Bolyai University, 400028 Cluj, Romania.*
[b] *Department of Physical Chemistry, Collegium Medicum, Nicolaus Copernicus University, Kurpińskiego 5, 85-096, Bydgoszcz, Poland.*
** Corresponding author: diudea@chem.ubbcluj.ro*

In general, exposure to dioxins in humans cause serious health problems such as cancer, chloracne, reproductive and developmental disorders. Human exposure to dioxins is achieved through diet (about 95%) and only a small amount of dioxinis taken by breathing or absorbed through the skin. Dioxins focus on the food chain, accumulate in animal fat, which explains why animals and animal products show a higher content of dioxins than plants or water. [5]

TCDD is a carcinogen, an endocrine disrupter, an agent that induces oxidative stress both in humans and in vertebrates. TCDD exposure causes cardiovascular dysfunction, neuronal degeneration and even craniofacial malformations. [6-8]

Chemical Graph Theory is a branch of mathematics applied to Chemistry. A *graph* $G(V,E)$ is a pair of two sets, *V* (vertices) and *E* (edges), the last one being a binary relation defined on *V* [9]. A molecular graphs, in which vertices are atoms and edges are covalent bonds, can be represented by a number, a sequence number, a matrix or polynomial. A single number representing a graph is also called a topological index and is useful in QSAR/QSPR (Quantitative Structure–Activity/ Property Relationships) studies.

DATA SET

From the PubChem [10] database, we selected 40 molecular structures of dioxins with their log P associated values. LogP is a measure of hydrophobic interactions of ligands with a biological receptor. LogP values can be determined experimentally or computationally. Table1 lists the dioxin names (IUPAC, cf. Figure 1, left, numbering) and their logP values. On the set of all dioxin structures, a hypermolecule (Figure1, right) was built up, as a collection of common skeletal and uncommon substructures [11].

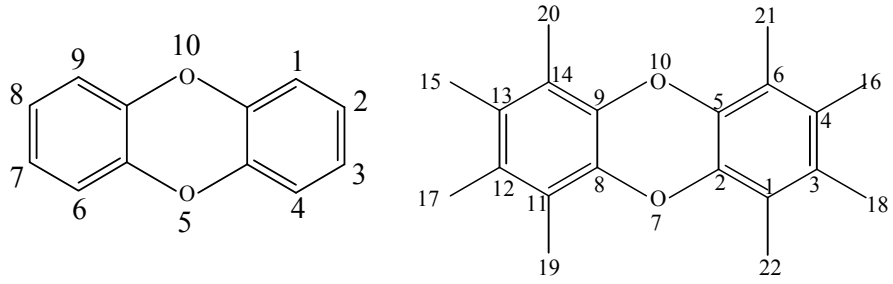

Figure 1. Dioxin IUPAC (left) and the hypermolecule (right) with atom numbering

Table 1. Dioxins –IUPAC name (cf. Figure 1, left) and their log P

	Name	log P
1	2,3,7,8-tetrachlorodibenzo-p-dioxin	6.4
2	Octachlorodibenzo-4-dioxin	8.1
3	2,7-dichlorodibenzo-4-dioxin	4.3
4	1,2,3,6,7,8-hexachlorodibenzodioxin	7.4
5	1,2,3,4,7,8-hexachlorodibenzodioxin	8.4
6	1,2,3,4,6,7,8-heptachlorodibenzodioxin	7.5
7	2-chlorodibenzo-4-dioxin	5
8	2,8-dichlorodibenzo-4-dioxin	4.3
9	1-Chlorodibenzo-p-dioxin	5
10	2,3-dichlorodibenzo-4-dioxin	5.2
11	1,2,3,7,8-pentachlorodibenzo-p-dioxin	6.6
12	1,2,4-trichlorodibenzo-1,4-dioxin	4.9
13	1,2,3,4-tetrachlorodibenzodioxin	7.2
14	1,2,7,8-tetrachlorodibenzo-p-dioxin	6
15	1,3,7,8-tetrachlorodibenzo-4-dioxin	6.3
16	1,6-Dichlorodibenzo-para-dioxin	5.7
17	1,3,6,8-tetrachlorodibenzo-p-dioxin	6.3
18	1,3,7-Trichlorodibenzo-p-dioxin	5.7
19	1,2,3,6,7,9-hexachlorodibenzo-p-dioxin	6.9
20	1,2,3,4,6,7,9-Heptachlorodibenzodioxin	7.5
21	1,2,3,8-tetrachlorodibenzo-p-dioxin	6
22	1,3-Dichlorodibenzo-para-dioxin	5
23	1,2,4,6,7,9-hexachlorodibenzo-p-dioxin	6.8
24	1,2,3,4,7-pentachlorodibenzo-p-dioxin	7.8
25	2,3,7-trichlorodibenzo-p-dioxin	5.8
26	1,2,6,8-tetrachloro dibenzo-p-dioxin	6.4
27	1,4,7,8-tetrachloro dibenzo-p-dioxin	6.4
28	1,4,6,9-tetrachloro dibenzo-p-dioxin	5.6
29	1,2,6,9-tetrachloro dibenzo-p-dioxin	5.6
30	1,2,3,7-tetrachloro dibenzo-p-dioxin	6
31	1,2,4,7,8-pentachlorodibenzo-p-dioxin	6.2
32	1,2,4,8-tetrachloro dibenzo-p-dioxin	6.4
33	1,2,4,7-tetrachloro dibenzo-p-dioxin	6.4
34	1,2,4,6-tetrachloro dibenzo-p-dioxin	5.6
35	1,2,3,9-tetrachloro dibenzo-p-dioxin	6.4
36	1,2,3,6-tetrachloro dibenzo-p-dioxin	6.4
37	1,3,6,9-tetrachloro dibenzo-p-dioxin	6.3
38	1,2,4,9-tetrachloro dibenzo-p-dioxin	5.6
39	1,2,4,6,8,9-hexachloro dibenzo-p-dioxin	6.8
40	1,2,3,4,6,8-hexachloro dibenzo-p-dioxin	7.2

COMPUTATIONAL DETAILS

The structures of dioxins have been optimized, in gas phase, at the Hartree-Fook HF (6-31g(d,p)) level of theory by Gaussian 09 [12]. Topological indices (see Table 2) have been computed by TOPOCLUJ software [13]; HOMO energy (in au) was computed by Gaussian 09.

Table 2. Topologicaldescriptors and HOMO energy (au)

Mol.	Adjacency	Detour	Distance	IE [CfMax]	IE [CfMin]	IP [CfMax]	IP [CfMin]	HOMO
1	20	1700	570	62	620	420	2800	-0.318
2	24	2700	920	130	950	760	4600	-0.337
3	18	1300	410	36	460	300	2000	-0.308
4	22	2200	730	92	780	590	3600	-0.328
5	22	2200	730	93	770	580	3600	-0.328
6	23	2400	820	110	860	670	4100	-0.332
7	17	1200	340	26	390	250	1600	-0.301
8	18	1300	410	36	460	300	2000	-0.307
9	17	1200	330	26	380	250	1600	-0.302
10	18	1300	410	37	460	300	1900	-0.307
11	21	1900	650	77	690	500	3200	-0.323
12	19	1500	460	49	510	380	2200	-0.314
13	20	1700	540	64	580	440	2600	-0.317
14	20	1700	560	62	610	430	2800	-0.318
15	20	1700	560	62	610	430	2800	-0.320
16	18	1400	400	36	450	320	1900	-0.310
17	20	1700	550	61	600	430	2700	-0.322
18	19	1500	480	48	530	360	2300	-0.315
19	22	2200	720	92	770	590	3600	-0.330
20	23	2400	810	110	850	680	4000	-0.334
21	20	1700	560	62	600	430	2700	-0.318
22	18	1400	400	37	450	310	1900	-0.308
23	22	2200	710	91	760	600	3500	-0.332
24	21	2000	630	77	680	510	3100	-0.323
25	19	1500	490	49	540	360	2400	-0.313
26	20	1700	550	61	600	440	2700	-0.320
27	20	1700	550	62	600	430	2700	-0.321
28	20	1800	540	61	580	450	2600	-0.325
29	20	1700	540	61	590	450	2700	-0.322
30	20	1700	560	62	610	430	2700	-0.318
31	21	2000	640	76	690	500	3200	-0.325
32	20	1700	550	62	600	440	2700	-0.320
33	20	1700	550	62	600	440	2700	-0.320
34	20	1800	540	62	590	450	2600	-0.322
35	20	1700	550	62	590	440	2700	-0.319
36	20	1700	550	62	590	440	2700	-0.320
37	20	1700	540	61	590	440	2700	-0.324
38	20	1800	540	62	590	450	2600	-0.322
39	22	2200	710	91	760	590	3500	-0.332
40	22	2200	720	92	760	590	3500	-0.330

The hypermolecule works like a biological receptor, over which the ligands (i.e. dioxins) are superposed/aligned. According to this superposition, *binary vectors* were constructed, with 1 when for a given position of the hypermolecule exists an atom in the current molecule, and zero, otherwise. In the so built binary vectors, the values 1 are next replaced by the partial charges (given as Supplementary data, at request) of ligand atoms, as computed at the HF level of theory.

RESULTS AND DISCUSSION

Data Reduction

In this step, the descriptors with variance <10% and intercorrelation > 0.80 (two descriptors highly correlated bring quite the same information on the molecule, one of them being sufficient) were discarded.

Evaluation of hypermolecule statistically significant positions was made by removing columns of data that contributes little to the correlation coefficient. From 22 initial positions we selected 16 positions (1, 2, 3, 4, 5, 8, 9, 10, 12, 13, 14, 16, 18, 19, 20 and 21).Correlation weighting [14] was performed on all the statistically significant positions in the hypermolecule: the local descriptors (actually the partial charges, computed at HF level of theory), have been multiplied by the corresponding correlation coefficients thus resulting new weighted vectors CD_{ij}. Next, these new descriptors are summed to give a global descriptor, which is a linear combination of the local correlating descriptors for the significant positions in the hypermolecule [15-17].

Modeling log P

The 40 structures were divided into two sets: the learning set (25 molecules) and the test set (15 molecules: 2; 3; 5; 6; 12; 13; 14; 17; 18; 21; 24; 26; 27; 28 and 37).

The models were performed on the learning set, the best results being listed below and in Table 3. The number of descriptors was limited to three, to fulfill the considerations of Topliss and Costello [18].

(i) Monovariate regression

$$\log P = 21.806 + 0.918 \times DS$$

(ii) Bivariate regression

$$\log P = 19.822 + 0.823 \times DS + 0.001 \times Distance$$

(iii) Three-variate regression

$$\log P = 17.628 + 0.842 \times DS + 0.220 \times Adj. - 0.001 \times Detour$$

Table 3. QSAR models and their statistics

	Descriptors	R^2	Adjust. R^2	St. Error	F
1	DS	0.908	0.904	0.253	226.639
2	Distance	0.832	0.825	0.341	114.197
3	IP min	0.828	0.820	0.345	110.431
4	IE min	0.826	0.818	0.347	108.872
5	IE max	0.824	0.816	0.349	107.438
6	Adjacency	0.817	0.809	0.356	102.717
7	DS, Adjacency	0.909	0.901	0.257	109.741
8	DS, Distance	0.909	0.901	0.257	109.879
9	DS, IE max	0.909	0.900	0.257	109.279
10	DS, IE min	0.909	0.901	0.257	109.639
11	DS, HOMO	0.908	0.900	0.257	109.069
12	Distance, IE min	0.845	0.831	0.335	60.045
13	DS, Adjacency, Detour	0.912	0.899	0.259	72.272
14	DS, Detour, IE max	0.911	0.898	0.260	71.706
15	Distance, IE max, IE min	0.847	0.825	0.340	38.809
16	Adjacency, Detour, Distance	0.839	0.816	0.350	36.399
17	Detour, Distance, IE max	0.838	0.815	0.350	36.225
18	Detour, IE max, IP max	0.8352	0.8117	0.3534	35.4878

External Validation

For the 25moleculesin the learning set, the best model was recorded for the trivariate model (DS, Adjacency and Detour), as shown in Table 3, (eq. 13). This model was used to predict log P of the molecules in the test set (15 molecules). Data for this external validation are listed in Table 4 while the plot of calculated logP vs. database-values is shown in Figure 2.

Table 4. LogP calc. cf (13) Table 3 on the molecules of the test set

Molecule	log P	log $P_{calc.}$
2	8.1	7.4
3	4.3	5.1
5	8.4	7.7
6	7.5	7.8
12	4.9	5.7
13	7.2	6.6
14	6	5.7
17	6.3	6.7
18	5.7	6.2
21	6	6.7
24	7.8	7.2
26	6.4	5.9
27	6.4	6.0
28	5.6	6.0
37	6.3	5.8

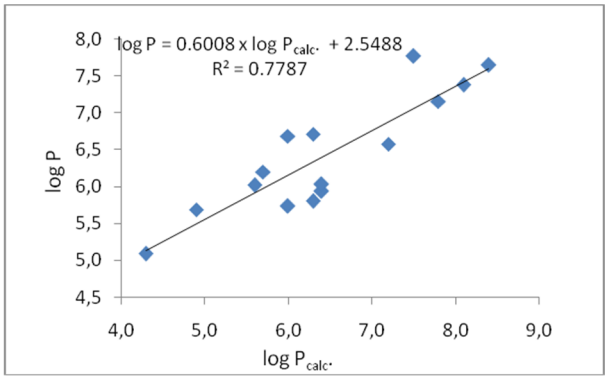

Figure 2. The plot log P vs. log $P_{calc.}$ for the test set (external validation)

Prediction by Clusters of Similarity

For the molecules in the test set, prediction can be done by means of similarity clusters: each of the 15 molecules in the test set is the leader of its own cluster, selected by 2D similarity among the 25 structures of the learning set (each cluster comprising about 15-20 molecules). The values log P were predicted by 15 new equations (the leader being left out) with the same descriptors as in eq. 13, Table 3. Data are listed in Table 5 and the monovariate correlation:

$$\log P = 0.763 \times \log P_{calc.} + 1.392 \text{ ; } n = 15; R^2 = 0.924; s = 0.331; F = 157.899$$

is plotted in Figure 3.

Table 5. Log $P_{calc.}$ on molecules leading to clusters of similarity

Molecule	log P	log $P_{calc.}$
2	8.1	7.8
3	4.3	4.6
5	8.4	7.6
6	7.5	7.7
12	4.9	5.2
13	7.2	6.7
14	6	5.8
17	6.3	6.3
18	5.7	5.9
21	6	6.3
24	7.8	7.1
26	6.4	5.9
27	6.4	6.2
28	5.6	5.7
37	6.3	6

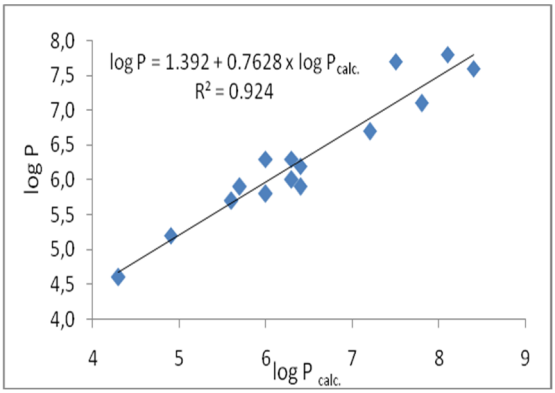

Figure 3. The plot log P vs. log $P_{calc.}$ by similarity clusters

CONCLUSION

A regression analysis was performed for 40 molecules of dioxins class. The study was focused on the correlation weighting of predictor variables describing the hypermolecule built up on the data set and prediction by clusters of similarity. The results showed a good modeling of log P by Cluj topological indices, HOMO energy level and the descriptor summing the contributions of the statistically significant positions in the hypermolecule. Validation of the model was done by an external set as well as by means of similarity clusters. Similarity calculation (in 2D) was done using the program TOPOCLUJ.

ACKNOWLEDGEMENS

R.M. acknowledges to Project POSDRU/159/1.5/S/132400, “Tineri cercetători de succes – dezvoltare profesională în context interdisciplinary şi internaţional”.

REFERENCES

[1] Schecter, L. Birnbaum, J.J. Ryan, J.D. Constable, *Environmental Research*, **2006**, *419*, 428.
[2] B. Olanca, G.C. Cakirogullari, Y. Ucar, D. Kirisik, D. Kilic, *Chemosphere*, **2014**, *94*, 13.
[3] J.M. Llobet, J.L. Domingo, A. Bocio, C. Casas, A. Teixidó, L. Müller, *Chemosphere*, **2003**, *50*, 1193.
[4] P.L. Galbenu, *Lucrări Ştiinţifice Medicină Veterinară*, Timişoara, **2009**, XLII (2).
[5] M. De Vries, R.P. Kwakkel and A. Kijlstra, *NJAS*, **2006**, *54-2*, 207.
[6] Q. Liu, M.L. Rise, J.M. Spitsbergen, T.S. Hori, M. Mierity, S. Geis, J.E. McGraw, G. Goety, J. Larson, R.J. Huty, M.J. Carvan III, *Aquatic Toxicology*, **2013**, *140-141*, 356.
[7] G. Xu, Q. Zhou, C. Wan, Y. Wang, J. Liu, Y. Li, X. Nie, C. Cheng, G. Chen, *NeuroToxicology*, **2013**, *37*, 63.
[8] Y. Li, G. Chen, J. Zhao, X. Nie, C. Wan, J. Liu, Z. Duan, G. Xu, *Toxicology*, **2013**, *312*, 132.
[9] M.V. Diudea, I. Gutman, L. Jäntschi, *Molecular Topology***,** NOVA, New York, 2002.
[10] PubChemdatabase; accessed 23.03. 2015.
[11] A.T. Balaban, A. Chiriac, I. Motoc, and Z. Simon, Steric Fit in QSAR (*Lectures Notes in Chemistry*, Vol. 15), Springer, Berlin, **1980**.
[12] **Gaussian 09**, Gaussian Inc Wallingford CT, Revision A.1, M.J. Frisch, G.W. Trucks, H.B. Schlegel, G.E. Scuseria, M.A. Robb, J.R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G.A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H.P. Hratchian, A.F. Izmaylov, J. Bloino, G. Zheng, J.L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J.A. Montgomery, J.E. Peralta, F. Ogliaro, M Bearpark, J.J. Heyd, E. Brothers, K.N. Kudin, V.N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J.C. Burant, S.S. Iyengar, J. Tomasi, M. Cossi, N. Rega, N.J. Millam, M. Klene, J.E. Knox, J.B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R.E. Stratmann, O. Yazyev, A.J. Austin, R. Cammi, C. Pomelli, J.W. Ochterski, R.L. Martin, K. Morokuma, V.G. Zakrzewski, G.A. Voth, P. Salvador, J.J. Dannenberg, S. Dapprich, A.D. Daniels, Ö. Farkas, J.B. Foresman, J.V. Ortiz, J. Cioslowski, D.J. Fox. **2009**.
[13] O. Ursu, M.V. Diudea, “TOPOCLUJ software program”, Babes-Bolyai University, Cluj, **2005**.
[14] A.A. Toropov, A.P. Toropova, *J. Mol. Struct*. (*Theochem*), **2001**, *538*, 287.
[15] C.D. Moldovan, A. Costescu, G. Katona, M.V. Diudea, *MATCH Commun. Math. Comput. Chem*., **2008**, *60*, 977.
[16] T.E. Harsa, A.M. Harsa, B. Szefler, *Cent. Eur. J. Chem*., **2014**, *12*, 365.
[17] A.M. Harsa, T.E. Harsa, S. Bolboaca, M.V. Diudea, *Curr. Comput.-Aided Drug Design*, **2014**, *10*, 115.
[18] J.G. Topliss, R.J. Costello, *J. Med. Chem*., **1972**, *15*, 1066.