STUDIA UNIVERSITATIS BABEŞ-BOLYAI, CHEMIA, XLVIII, 2, 2003

MODELING PHYSICO-CHEMICAL PROPERTIES OF POLYCHLORINATED BIPHENYLS ON THE GROUNDS OF HYPERMOLECULE CONCEPT

MIHAELA CAPRIOARA, CSABA L. NAGY AND MIRCEA V. DIUDEA

Faculty of Chemistry and Chemical Engineering,
Babeş-Bolyai University, 400028 Cluj, Romania

ABSTRACT. Biphenyls (PCBs) represent a class of pollutants with great environmental impact. For such chemicals, QSPR studies were performed in order to emphasize the relation between some topochemical and electronic parameters and their molecular properties. The models are described on the ground of hypermolecule concept. Local group mass, autocorrelated by a multivariate regression, are used for molecular description. A map of relevant positions in the hypermolecule, most probably involved in the global (*i.e.*, molecular) property could be drawn. The hypermolecule, in this case, has the meaning of a "mean molecule" in the set. A general procedure for developing and validating models using the above concept is given. Within this frame, a method of data reduction (*i.e.*, selection of relevant descriptors) is exemplified.

INTRODUCTION

An important aspect of modern toxicology research is estimation of properties of environmental pollutants from their molecular structure.[1] The potential toxicity of a compound is normally assessed on the basis of a wide variety of relevant physico-chemical properties. The application of QSPR techniques to the elucidation of the ways in which structure determines chemical and physical properties has already become an essential tool in the area of chemistry.

An enormous attraction of QSPR is that it potentially combines the ability to predict chemical and physical properties of as yet unmeasured or unknown compounds with the ability to understand just how the structure influences a particular chemical and physical property.

Polychlorinated biphenyls (PCBs) are ubiquitous environmental pollutants with different physical and chemical properties, but with quite different biochemical effects.[2] Only few of them exhibit high direct toxicity. Exposure by most of them and by some of their metabolites may result in immunodeficiency, neurotoxic effects including impaired learning, terratogenic effects and endocrine disruptions.[3,4] These compounds originate from various human activities (and also from natural sources) are of special interest because of their global distribution, persistence, high tendency to bioaccumulate and known or suspected toxicity.

The physicochemical properties (*e.g.*, the partition coefficient in octanol-water system, log*P*) and molecular structures of PCBs are strongly related to their toxic effects on biological systems. We tried to identify those structural features that could be relevant in the relationship log*P* - PCBs. In modeling this molecular property, we used some simple molecular descriptors, based on atom group substituent or on the whole molecule.

This paper is focused on development of novel QSPR models using molecular descriptors and available experimental physico-chemical data. General structures of biphenyl herein investigated are presented in Table 1.

MATERIALS AND METHODS

The correlating study was performed on a set of PCBs,[6] being the property log*P*. Table 1 lists the general formula of biphenyls, property and name of compounds.

Among the physico-chemical properties of PCBs, partitioning behavior of these compounds in various media is important in spreading the pollutants in the environment. In general, lipophilicity increases with increasing degree of chlorination. Lipophilicity is known to be involved in their toxicity.

2.1 Chemical Data

The physico-chemical properties were previously modeled by using molecular linear free energy relationship descriptors (LFER). Modeling log*P* for the set of the 16 PCBs led to an *r* value of 0.9868 and predictive correlation of 0.9614.[6-8]

2.2 Calculation of Molecular Descriptors

In order to model the correlation of physico-chemical property with the molecular structure, (QSPR models) the structure must be described in a numerical form. The arrangement of chlorine atoms on the biphenyl structure can be accounted for, by the *hypermolecule HM* concept,[9] viewed as a "mean molecule" of the whole set.[10]

In the construction of the hypermolecule, a *row-vector* P_i of dimension N_{HM} is attached to each molecule M_i :

$$P_i = \{ P_{ij} \; ; \; j = 1,2,\ldots, N_{HM} \} \qquad (1)$$

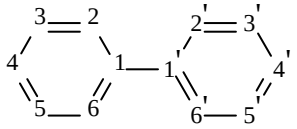
where N_{HM} is the number of vertices in the hypermolecule. The molecules of the set are superimposed according to their maximal common substructures.

In the associated vector, the matching positions take $P_{ij} = 1$ while for the non-matching ones $P_{ij} = 0$. The description of the j^{th} position in *HM* (*e.g.*, the chemical and/or topological nature of the j^{th} atoms) is given by X_{ij}.

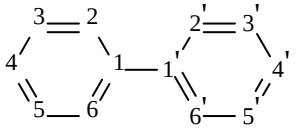
The hypermolecule and the molecules under study can be numerically described by using some molecular *topological descriptors* TD_i:[11,12]

Table 1

General structure of biphenyl. Chemicals and their log*P* [a]

Compound	Structure
Biphenyl	

Code	Compound	log*P*
PCB1	Biphenyl	3.98
PCB2	4-chloro-biphenyl	4.61
PCB3	2,2'-dichloro-biphenyl	4.73
PCB4	4,4'-dichloro-biphenyl	5.58
PCB5	2,4,5-trichloro-biphenyl	5.81

Compound	Structure
Biphenyl	

Code	Compound	log*P*
PCB6	2,2',5-trichloro-biphenyl	5.6
PCB7	2,2',5,5'-tetrachloro-biphenyl	6.09
PCB8	2,3,4,5-tetrachloro-biphenyl	6.41
PCB9	2,2',4,5,5'-pentachloro-biphenyl	6.44
PCB10	2,3,4,5,6-pentachloro-biphenyl	6.52
PCB11	2,2',4,4',5,5'-hexachloro-biphenyl	6.8
PCB12	3,3',4,4',5,5'-hexachloro-biphenyl	7.55
PCB13	2,2',3,3',4,4',6-heptachloro-biphenyl	6.99
PCB14	2,2',3,3',5,5',6,6'-octachloro-biphenyl	7.15
PCB15	2,2',3,3',4,5,5',6,6'-nonachloro-biphenyl	8.16
PCB16	2,2',3,3',4,4',5,5',6,6'-decachloro-biphenyl	8.26

[a]values by Bodor's log*P* dataset

$$TD_i = \sum_j TD_{ij} = \sum_j a_j P_{ij} X_{ij} \quad (2)$$

where P_{ij} and X_{ij} have the meaning above mentioned while a_j is the regression coefficient as given by the multivariate regression log$P = f(X_i)$. The above *TDs* are "ad-hoc" descriptors[13,14] depending on the set of molecules considered and the selected molecular property. Therefore, all the polychlorinated biphenyls can be described using particular local properties that characterize both the substituted/ unsubstituted aromatic positions.

As local property for calculating *TDs*, we used X_{ij} = the group mass related to each position in PCBs.

2.3 Statistics

For developing the QSPR models, the two sets of PCBs were submitted to REGLINEWIN (a home made statistical software package). The correlating equation is of the form:

$$Y_i = b_0 + \sum_{j=1}^{m} b_j \cdot TD_{ij} \quad (3)$$

where Y_i is the dependent variable, TD_{ij} are the predictor variables, $m < n$, n being the number of structures in the set.

A general algorithm for generating a correlating model was carried out:

1) select a training data set of compounds with great structural diversity and generate the *hypermolecule*.

2) perform a Leave-one-out analysis on the input set of local descriptors.

3) calculate the global descriptors by using appropriate weights in the training set.

4) find a regression function by using statistical tools.
5) test the predictive capacity of the model.

RESULTS AND DISCUSSION

The results of correlating analysis will be presented separately for each of the two sets.

3.1 Set of 16 PCBs, Property: logP.

This subset of PCBs consists of 16 polychlorinated compounds and the property is log*P* (Table I). Within this set, we delimited a training set (n = 12) and a test set (n = 4).

Lipophilicity controls the free energies of PCBs binding to AhR.[15] Comparing to other chemicals, PCBs have very high *P* values:[16] log*P*s are in the range [4.5 – 8.5]. Consequently, PCBs tend to adsorb to unpolar surfaces and accumulate in lipophilic matrices. For these compounds, the lipophilicity is related directly to bioconcentration and inversely to aerobic degradation.

Keeping in mind the physical nature of the PCB property we used a mass descriptor *MD* as independent variable. This descriptor accounts for vertices as groups of atoms, *e.g*., CH, C-Cl, *etc*.

3.1.a Hypermolecule model

A *hypermolecule* already generated, we calculated *MD*s for all molecules in the training set. The group mass M_i for each vertex fragment in PCBs is multiplied by the regression coefficients in a multivariate regression including all positions in the hypermolecule (see Table 2).

Table 2

Training set (n = 12). Group mass descriptors M_i and regression coefficients c_i corresponding to aromatic positions.

Molecule	M_2	M_3	M_4	M_5	M_6	$M_{2'}$	$M_{3'}$	$M_{4'}$	$M_{5'}$	$M_{6'}$
c_i	0.01	0.02	0.01	0.03	0.00	0.01	0.02	0.01	0.01	-0.01
PCB1	13.01	13.01	13.01	13.01	13.01	13.01	13.01	13.01	13.01	13.01
PCB2	13.01	13.01	47.46	13.01	13.01	13.01	13.01	13.01	13.01	13.01
PCB5	47.46	13.01	47.46	47.46	13.01	13.01	13.01	13.01	13.01	13.01
PCB6	47.46	13.01	13.01	47.46	13.01	47.46	13.01	13.01	13.01	13.01
PCB7	47.46	13.01	13.01	47.46	13.01	47.46	13.01	13.01	47.46	13.01
PCB8	47.46	47.46	47.46	47.46	13.01	13.01	13.01	13.01	13.01	13.01
PCB9	47.46	13.01	47.46	47.46	13.01	47.46	13.01	13.01	47.46	13.01
PCB10	47.46	47.46	47.46	47.46	47.46	13.01	13.01	13.01	13.01	13.01
PCB11	47.46	13.01	47.46	47.46	13.01	47.46	13.01	47.46	47.46	13.01
PCB12	13.01	47.46	47.46	47.46	13.01	13.01	47.46	47.46	47.46	13.01
PCB13	47.46	47.46	47.46	13.01	47.46	47.46	47.46	47.46	13.01	13.01
PCB14	47.46	47.46	13.01	47.46	47.46	47.46	47.46	13.01	47.46	47.46

The leave-one-out LOO procedure is required to select among the 10 positions (k = 10) the corresponding relevant descriptors. It resulted in a map of relevant positions in the hypermolecule, most probably involved in the global (*i.e*., molecular) property (see Table 3). The hypermolecule, in this case, has the meaning of a “mean molecule" in the set.

Table 3

Training set (n = 12). Statistical parameters of the LOO procedure tested on partial descriptors set.

Number of descriptors (k)	Model (multivariate regression)	Irrelevant positions	Correlation coeff. (r) (after descriptor elimination)
k-1	logP = f(M_i), i =1, 2,… k-1	M_6	0.9992
k-2	logP = f(M_i), i =1, 2,… k-2	M_6, $M_{6'}$	0.9979
k-3	logP = f(M_i), i = 1, 2,… k-3	M_2, M_6, $M_{6'}$	0.9966
k-4	logP = f(M_i), i = 1, 2,… k-4	M_2, $M_{5'}$, M_6, $M_{6'}$	0.9945

From Table 3, it appears that a satisfactory statistics is ensured by a map comprising k-4 relevant positions: M_3, M_4, M_5, $M_{2'}$, $M_{3'}$, $M_{4'}$, for which the local parameters global mass descriptor *MD* are given in Table 4.

Table 4

Regression coefficients (c_i) and molecular mass descriptor in the training set.

Number of descriptors (k-4)	M_3	M_4	M_5	$M_{2'}$	$M_{3'}$
c_i	0.019	0.019	0.035	0.019	0.02
Molecules	*MD*	P_{obs}		P_{calc}	Residual
PCB1	1.586	3.98		3.98	0
PCB2	2.226	4.61		4.62	-0.01
PCB5	3.426	5.81		5.82	-0.01
PCB6	3.436	5.6		5.83	-0.23
PCB7	3.436	6.09		5.83	0.26
PCB8	4.071	6.41		6.465	-0.055
PCB9	4.076	6.49		6.47	0.02
PCB10	4.071	6.52		6.465	0.055
PCB11	4.436	6.8		6.83	-0.03
PCB12	5.136	7.55		7.53	0.02
PCB13	4.586	6.99		6.98	0.01
PCB14	4.786	7.15		7.18	-0.03

Model generation. The structure-log*P* model for PCBs fits a linear function, its quality being estimated by the correlation coefficient (r), the standard error of estimate (s), the Fischer ratio (F) and the coefficient of variance (*CV* %), the last one being a ratio of s to the property mean value in the set. The regression analysis was performed using the REGLINWIN package. The following QSPR model was generated:

$$\log P = 2.394 + MD \qquad (4)$$

n = 12; k = 6; r = 0.9946; s = 0.114; F = 902.23; *CV* %=1.852

Random correlation: r_1 = 0.650; r_2 = 0.674; r_3 = 0.535

We tried several partitions of the whole set of 16 molecules into a training set and a test set and we concluded that the final results do not depend significantly on the specific manner of choosing the members of each subset. A random test showed a marked drop in the correlation coefficient, with the meaning that no chance correlation occurred.

Figure 1 shows the plot of log*P vs*. *MD* values in the PCBs training set.

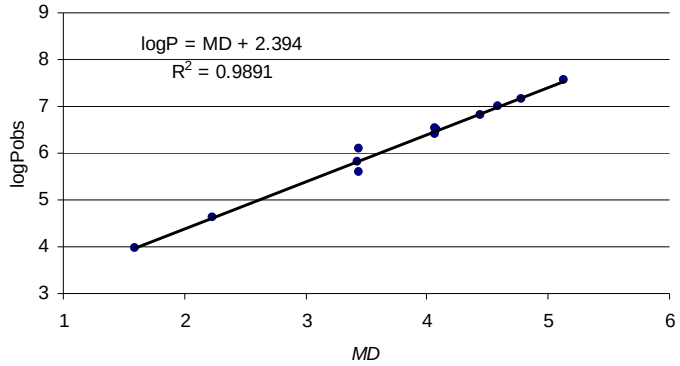


Figure 1. *Plot of logP vs. calculated MD values.*

Model validation. In order to verify the predicting ability of this QSPR model, we first calculated *MDs* for the test molecules (*n* = 4) using their corresponding vertex-group mass and the regression coefficients prior generated for the training set (see Table 5 – we suppose the partition coefficient in the predicting set is unknown).

Table 5

Molecular mass descriptors *MD* in test set (*n* = 4). Estimated log*P* from the linear regression model eq (4)

Molecules	*MD*	logP_{calc} (eq. 4)	log*P*	Residual
PCB3	2.236	4.63	4.73	0.1
PCB4	2.586	4.98	5.58	0.6
PCB15	5.426	7.82	8.16	0.34
PCB16	5.786	8.18	8.26	0.08
$\log P = 0.5318 + 0.96071 \cdot \log P_{calc}$			r = 0.9916	

The model (4) is excellent both in estimation and prediction, as shown in Table 5 and the correlation equation of the calculated (by eq 4) values vs the observed ones.

3.1.b Average Descriptors and the Eigenvalue of S Matrix

The symmetry of the biphenyl skeleton suggests the superimposing of different positions, above considered as separate positions. Thus, we calculated the following descriptors: AM_2 = av(M_2, M_2', M_6, M_6'); AM_3 = av(M_3, M_3', M_5, M_5') and AM_4 = av(M_4, M_4'). Each average value was further divided by 2, thus accounting for 8 possible equivalent structures. Data are given in Table 6.

The second path matrix [17] **S** = **S**(*G*) is defined as the difference between the squared adjacency matrix and the diagonal matrix **DEG** of vertex degrees in *G*:

$$\mathbf{S} = \mathbf{S}(G) = \mathbf{A}^2 - \mathbf{DEG} \tag{5}$$

The characteristic polynomial of **S** is:

$$ChS(x, G) = \det(x \cdot \mathbf{I} - \mathbf{S}) = \sum_{k=0}^{n} b_k(G) \cdot x^{n-k} \tag{6}$$

The matrix **S** counts the all distinct path of length two in *G*. The first eigenvalue of **S** (*i.e.*, the first root of *ChS*(*x*,*G*)) is used in the following for modeling log*P* in the PCBs in Table 1.

In trivariate regression, with the three averaged mass descriptors, one obtains:

$$\log P = 2.918 + 0.022 \cdot AM_2 + 0.062 \cdot AM_3 + 0.038 \cdot AM_4 \tag{7}$$

n = 16; *r* = 0.9661; *s* = 0.353; *CV*% = 5.63; *F* = 56.023

A global descriptor *AMD* is again calculable (cf. eq 2) as a linear combination of the the above three averaged mass descriptors. Data are given in Table 6.

Table 6

Averaged mass descriptors AM_i and the global descriptor *AMD*

Molecules	log*P*	AM_2	AM_3	AM_4	*AMD*	*EV_S*
c_i		0.022	0.062	0.038		
PCB1	3.98	10.125	13.011	13.011	1.525	2.842
PCB2	4.61	10.253	21.624	13.011	2.064	3
PCB3	4.73	30.238	13.011	13.011	2.718	3.107
PCB4	5.58	13.011	13.011	47.464	2.372	3
PCB5	5.81	10.568	21.624	30.238	3.022	3.282
PCB6	5.6	19.128	13.011	30.238	2.914	3.265
PCB7	6.09	19.236	13.011	47.464	3.561	3.265
PCB8	6.41	19.331	21.624	30.238	3.451	3.338
PCB9	6.49	19.365	21.624	47.464	4.099	3.385
PCB10	6.52	19.358	30.238	30.238	4.295	3.535
PCB11	6.8	19.428	30.238	47.464	3.727	3.385
PCB12	7.55	28.229	30.238	47.464	4.485	3.382
PCB13	6.99	36.702	38.851	13.011	1.973	3.601
PCB14	7.15	36.742	30.238	47.464	2.883	3.591
PCB15	8.16	47.464	47.464	30.238	5.15	3.692
PCB16	8.26	47.464	47.464	47.464	5.797	3.692

The monovariate regression is now:

$$\log P = 2.918 + AMD \tag{8}$$

n = 16; *r* = 0.9661; *s* = 0.3265; *CV*% = 5.186; *F* = 196.15

In view of improving the correlation, we aded a new descriptor, namely the first eigenvalue of the S matrix *EV_S.* The bivariate regression obtained is:

$$\log P = -1.331 + 0.703 \cdot AMD + 1.575 \cdot EV\_S \quad (9)$$

n = 16; r = 0.979; s = 0.268; CV% = 4.251; F = 149.94

The analysis of the residuals, by eq 9, revealed two outliers, as shown in Table 7. The best model is obtained on 14 PCBs:

$$\log P = -1.790 + 0.654 \cdot AMD + 1.741 \cdot E\_S \quad (10)$$

n = 14; r = 0.9908; s = 0.185; CV% = 2.99; F = 296.62

LOO: q = 0.984;Random: r = 0.583

A random mixing of the modeled property showed a significant drop in the correlation coefficient value. Prediction ability of the model (eq 10) is very good (q = 0.984). The explicite variance is higher than 0.98 (see the plot, Figure 2). The data of the model (10) are given in Table 8.

Table 7

Outlier analysis

Number of molecules (n)	Model (multivariate regression)	Outliers	Correlation coefficient (r) (after descriptor elimination)
n	logP = f(M_i), i =1, 2,… n	-	0.979
n - 1	logP = f(M_i), i =1, 2,… n-1	PCB12	0.9858
n - 2	logP = f(M_i), i = 1, 2,… n-2	PCB12, PCB8	0.9908

Table 8

Data of the model (10)

	Molecules	*AMD*	*EV_S*	P_{obs}	P_{calc}	Residual
1	PCB1	1.525	2.842	3.98	4.155	0.175
2	PCB2	2.064	3	4.61	4.783	0.173
3	PCB3	1.973	3.107	4.73	4.909	0.179
4	PCB4	2.883	3	5.58	5.319	-0.261
5	PCB5	2.718	3.282	5.81	5.702	-0.108
6	PCB6	2.372	3.265	5.6	5.446	-0.154
7	PCB7	3.022	3.265	6.09	5.871	-0.219
8	PCB9	3.561	3.385	6.49	6.432	-0.058
9	PCB10	3.451	3.535	6.52	6.621	0.101
10	PCB11	4.099	3.385	6.8	6.784	-0.016
11	PCB13	3.727	3.601	6.99	6.917	-0.073
12	PCB14	4.485	3.591	7.15	7.395	0.245
13	PCB15	5.15	3.692	8.16	8.006	-0.154
14	PCB16	5.797	3.692	8.26	8.43	0.17

The molecular descriptors used in the second part of our study are intercorrelated as shown in Table 9. Their collinearity is pretty good for a multilinear regression.

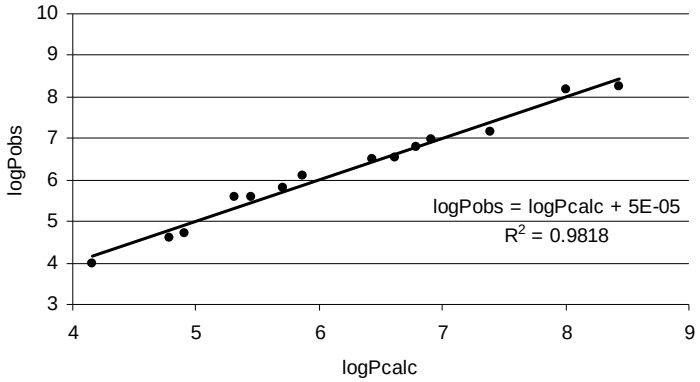


Figure 2. *Plot of logP vs. logP calculated by eq. 10*

Table 9

Intercorrelation of the averaged mass and *EV_S* descriptors

	AM_2	AM_3	AM_4	*EV_S*
AM_2	1	0.794	0.153	0.790
AM_3		1	0.157	0.849
AM_4			1	0.360
EV_S				1

CONCLUSIONS

In the present paper, we have used (local) molecular descriptors for encoding a *hypermolecule*, as a "mean molecule" in the set, in case of a physico-chemical property.

The local descriptors are fitted by means of multivariate regression and represent "ad-hoc" (or "autocorrelated") descriptors, which change with the set and selected property.

A general procedure for developing and validating models using the concept of the hypermolecule was given. Within this frame, two ways of data reduction (*i.e.*, selection of relevant descriptors) were exemplified.

REFERENCES

1. H. Kubinyi, *Quant. Struct.-Act. Relat*. **1994**, *13*, 285-294.
2. M. V. Diudea, *J. Chem. Inf. Comput. Sci*. **1997**, *37*, 300-305.
3. J. G. Topliss and R. P. Edwards, *J. Med. Chem*. **1979**, *22*, 1238-1244.
4. S. C. Basak, G. D. Grunwald, B. D. Gute, K. Balasubramanian, and D. Opitz, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 885-890.

5. G. G. Oakley, U. Devanaboyina, L. W. Robertson, and R. C. Gupta, *Chem. Res. Toxicol.* **1996**, *9*, 1285-1292.
6. M. D. Erickson, *Analytical Chemistry of PCB*s, Butterworth Publishers, Boston, **1986**.
7. E. M. Silberhorn, H. P. Glauert, and L. W. Robertson, *CRC Crit. Rev. Toxicol.* **1990**, *20*, 439-496.
8. D. Mackay, W. Y. Shiu, and K. C. Ma, *Illustrated Handbook of Physical-Chemical Properties and Environmental Fate for Organic Chemicals*. Volume I+II. Lewis Publishers Inc., Boca Raton, FL, USA, **1992**.
9. J. A. Platts, M. H. Abraham, D. Butina, and A. Hersey, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 71-80.
10. P. Buchwald, and N. Bodor, *Curr. Med. Chem.* **1998**, *5*, 353-380.
11. R. P. Schvarzenbach, P. M. Gschwend, and D. M. Imboden, *Environmental Organic Chemistry*, John Wiley&Sons, New York, **1992**.
12. O. M. Minailiuc, and M. V. Diudea, TI-MTD Model. Applications in Molecular Design, in: *QSPR/QSAR Studies by Molecular Descriptors*, Nova Science Publishers, Inc., **2001**, pp. 363-388.
13. R. P. Sheridan, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1456-1469.
14. I. Motoc, S. Holban, R. Vancea, and Z. Simon, *Studia Biophys.* (*Berlin*) ***1977***, *66*, 75-78.
15. A. T. Balaban, A. Chiriac, I. Motoc, and Z. Simon, *Steric Fit in QSAR* (*Lectures Notes in Chemistry*, Vol. 15), Springer, Berlin, **1980**, Chap. 6.
16. A. A. Toropov, and A. P. Toropova, *Internet Electronic Journal of Molecular Design*, **2002**, *1*, 108-114.
17. D. J. G. Marino, P. J. Peruzzo, E. A. Castro and A. A. Toropov, *Internet Electronic Journal of Molecular Design*, ***2002***, *1*, 115-133.
18. S. A. Kafafi, H. Y. Afeefy, A. H. Ali, H. K. Said, and A. G. Kafafi, *Environ. Health Perspect*. ***1993***, *101*, 422-428.
19. H. Fiedler, H. Hoff, J. Tolls, C. Mertens, A. Gruber, and O. Hutzinger, *Environmental Fate of Organochlorines in the Aquatic Environment. Organohalogen Compd. 15*, ECO-INFORMA Press, Bayreuth, 199 Pages, ISBN 3928379119, **1994**.
20. P. E. John and M. V. Diudea, *MATCH Commun. Math. Comput. Chem*., **2003** (accepted).