Available online at www.sciencedirect.com
SCIENCE DIRECT®
Bioorganic & Medicinal Chemistry 14 (2006) 3923–3928
Bioorganic & Medicinal Chemistry

Correlation weighting of valence shells in QSAR analysis of toxicity

Andrey A. Toropov* and Emilio Benfenati

Laboratory of Environmental Chemistry and Toxicology, Istituto di Ricerche Farmacologiche 'Mario Negri', Via Eritrea 62, 20157 Milan, Italy

Received 8 October 2005; revised 18 January 2006; accepted 20 January 2006
Available online 3 February 2006

Abstract—In the rainbow trout (*Oncorhynchus mykiss*), we studied the acute toxicity LC_{50}-96h of 274 organic pesticides with a wide variety of molecular structures. Optimization of correlation weights of local and global graph invariants (OCWLGI) gave quantitative structure–activity relationships (QSARs) for predicting toxicity. We used a labeled hydrogen-filled graph (LHFG) to elucidate the molecular structure. We also used the extended connectivity of zero (^0EC_k), first (^1EC_k), and second (^2EC_k) order, numbers of path lengths 2 ($P2_k$) and 3 ($P3_k$) starting from a given vertex in the LHFG, and valence shells of second order ($S2_k$). $S2_k$ is the sum of the degree of vertices at distance 2 from a given vertex k. The presence of three-, five-, and six-member cycles and hydrogen bond indices suggested they might be used as global LHFG invariants. We applied this method to a broad set of pesticides, to predict toxicity for the trout. The best model used weighted $S2_k$ and global LHFG invariants. Statistical characteristics of this model are as follows: $n = 233$, $r^2 = 0.7689$, $r^2(\text{pred}) = 0.7688$, $s = 0.75$, $F = 769$ (training set); $n = 41$, $r^2 = 0.6421$, $r^2(\text{pred}) = 0.4241$, $s = 1.14$, $F = 70$ (test set).

1. Introduction

Computer-based models of different kinds of toxicity are necessary for ecology, biology, and medicine, for two main reasons: (1) it is often costly and time-consuming to obtain experimental data, while models offer predicted values easily and quickly; (2) models can provide mechanistic insights useful for developing further knowledge.[1]

The quantitative structure–property/activity relationships (QSPR/QSAR) can provide useful new information and indicate active molecular fragments that influence toxicity. Unfortunately, it is hard to obtain universal QSARs for toxicity endpoints so it may be useful to restrict the problem, though it still remains a challenge. This refers, for instance, to pollutants or pesticides, fertilizers, and to specific questions such as drug design. These approaches are all of interest in chemistry, biology, ecology, and medicine.

By so-called optimization of correlation weights of local and global invariant (OCWLGI) of a molecular graph, we obtained QSPR/QSARs for predicting physico-chemical parameters and biological activity.[2–10] Several papers have reported using QSPR/QSAR modeling with path numbers and valence shells.[11–15] The present study estimated the utility of the OCWLGI, based on path numbers and valence shells, for toxicity prediction. It also set out to improve the OCWLGI,[2–10,15] taking account of the presence of cycles in molecular structure and the ability of substances to generate hydrogen bond interactions. In practice, the OCWLGI scheme can be extended by means of descriptors, which encode the molecular features.[2]

For the rainbow trout (*Oncorhynchus mykiss*) we examined the acute toxicity LC_{50}-96h to test the extended OCWLGI scheme.

2. QSAR method

2.1. Toxicity data

This study checked the rainbow trout acute toxicity experimental LC_{50}-96h, which is the dose that kills 50% of all fish in 96h. As endpoint we used the decimal

Keywords: QSAR; Toxicity; Pesticides; Correlation weights; Path numbers; Valence shells; Cyclicity code; Hydrogen bond index.
* Corresponding author at present address: Sergeli 8-A, Home 4, Room 6, 100085 Tashkent, Uzbekistan; e-mail: aatoropov@yahoo.com

0968-0896/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmc.2006.01.038

logarithm log(1/*C*), where C is the concentration expressed in mmol/L. The toxicity data were kindly provided by Dr. Brian Montague, from the US Environmental Protection Agency. These data are highly reliable because they are obtained according to a standardized protocol accepted by the EPA. For even greater reliability we compared the toxicity data with two other reference databases, and when multiple values were found we selected the values according to a reference protocol; details have been published.[16] In total, 274 compounds were investigated, according to the criteria previously indicated;[16] 41 were randomly selected as an external test set. No information on pesticides from the test set was used in construction of models, described below.

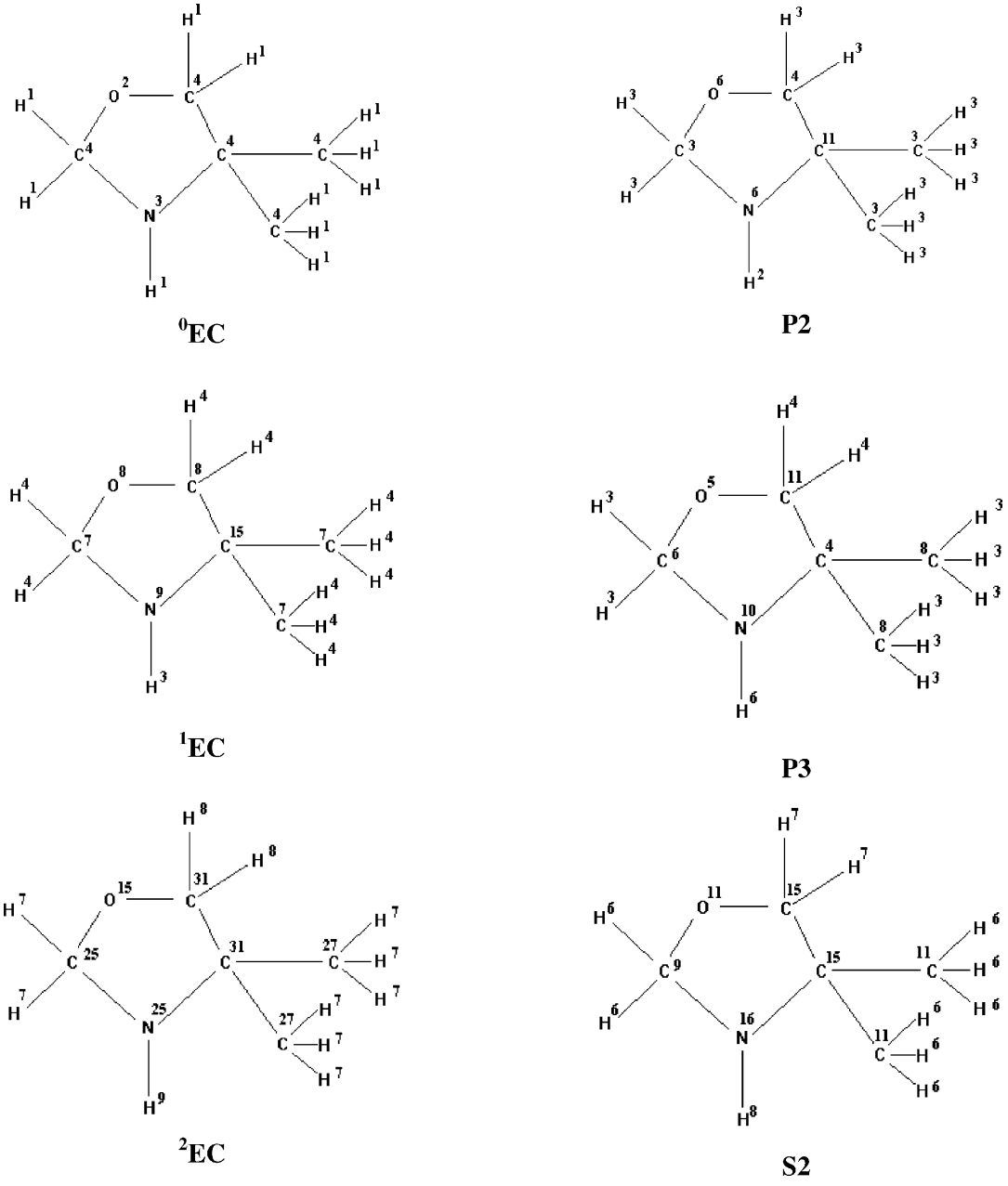
2.2. Optimal descriptors

A labeled hydrogen-filled graph (LHFG) describes the molecular structure used in the present study, which employed two LHFG descriptors. The first were calculated with the equation:

$$ {}^{0}X_{\mathrm{CW}}(\mathrm{LI}) = \sum_{k=1}^{N} \mathrm{CW}(a_k) + \sum_{k=1}^{N} \mathrm{CW}(\mathrm{LI}_k) \qquad (1) $$

where LI is the local invariant, $\mathrm{CW}(a_k)$ and $\mathrm{CW}(\mathrm{LI}_k)$ are correlation weights of the chemical element, a_k, that is the image of the vertex k in the LHFG, and LI_k is one of the following local invariants: extended connectivity of zero ($^{0}\mathrm{EC}_k$), first ($^{1}\mathrm{EC}_k$), and second ($^{2}\mathrm{EC}_k$) order;

Table 1. Examples of calculating $^{0}\mathrm{EC}_k$, $^{1}\mathrm{EC}_k$, $^{2}\mathrm{EC}_k$, $\mathrm{P2}_k$, $\mathrm{P3}_k$, and $\mathrm{S2}_k$ for vertices in LHFG of 4,4-dimethyloxazolidine



$P2_k$ and $P3_k$ are numbers of paths of lengths 2 and 3, and $S2_k$ is the valence shell of second order of vertex k, which is the sum of vertex degrees over all vertices separated by two edges from vertex k. Examples of these local invariant numerical values are shown in Table 1.

The second LHFG descriptors were calculated with an extended version of Eq. 1 taking into account both local and global LHFG invariants:

$$^{0}X_{\mathrm{CW}}(\mathrm{LI},\ \mathrm{GI}) = \mathrm{CW}(\mathrm{CC}) + \mathrm{CW}(\mathrm{HB_o}) + \mathrm{CW}(\mathrm{HB_n}) + \sum_{k=1}^{\mathrm{N}} \mathrm{CW}(a_k) + \sum_{k=1}^{\mathrm{N}} \mathrm{CW}(\mathrm{LI}_k) \quad (2)$$

where CC is a code of cyclicity, calculated as 1000 + 600 C6 + 50 C5 + 3C3. These C6, C5, and C3 are indicators of presence of six-, five-, and three-membered cycles. In other words, the Cx equal to 1, if x-membered cycle is presence, and equal to 0, otherwise. Four-membered cycles were present only in one compound, so we did not use any four-membered cycle in our model; hydrogen bond indices on oxygen (HB_o) and nitrogen (HB_n) were defined as

$$\mathrm{HB_o} = 2000 + 100 \cdot \mathrm{NVO_1} + \mathrm{NVO_2},$$

$$\mathrm{HB_n} = 3000 + 100 \cdot \mathrm{NVN_1} + \mathrm{NVN_3},$$

where NVO_1 and NVO_2 are the numbers of oxygen vertices of valence (degree) 1 and 2, and NVN_1 and NVN_3 are the numbers of nitrogen vertices of valence (degree) 1 and 3.

Numerical values of all CW (I) were obtained by the Monte Carlo optimization procedure,[3] that produces as large as possible a correlation coefficient between toxicity and the $^{0}X_{\mathrm{CW}}(\mathrm{LI})$ or $^{0}X_{\mathrm{CW}}(\mathrm{LI}, \mathrm{GI})$ on compounds of the training set. When CWs are defined, one can calculate $^{0}X_{\mathrm{CW}}(\mathrm{LI})$ or $^{0}X_{\mathrm{CW}}(\mathrm{LI}, \mathrm{GI})$ for all the training set compounds, and the least-squares method gives model toxicity in the forms:

$$\log(1/C) = C_0 + C_1 \cdot {}^{0}X_{\mathrm{CW}}(\mathrm{LI}) \quad (3)$$

$$\log(1/C) = C_0 + C_1 \cdot {}^{0}X_{\mathrm{CW}}(\mathrm{LI},\ \mathrm{GI}) \quad (4)$$

The predictive ability of the model can be validated with compounds of the test set.

3. Results and discussion

Table 2 shows there are some reasonably good models of toxicity based on Eq. 2, namely, $^{0}X_{\mathrm{CW}}(^{1}\mathrm{EC}, \mathrm{GI})$, $^{0}X_{\mathrm{CW}}(\mathrm{P2}, \mathrm{GI})$, and $^{0}X_{\mathrm{CW}}(\mathrm{S2}, \mathrm{GI})$. However, $^{0}X_{\mathrm{CW}}(\mathrm{S2}, \mathrm{GI})$ gave the statistically preferable model. Table 3 lists correlation weights for calculating $^{0}X_{\mathrm{CW}}(\mathrm{S2}, \mathrm{GI})$. Table 4 shows a calculation of this descriptor for LHFG of 4,4-dimethyloxazolidine.

Figures 1 and 2 show the toxicity values on the training and test sets. This model was calculated with the equation:

$$\log(1/C) = -0.594 + 0.130 \cdot {}^{0}X_{\mathrm{CW}}(\mathrm{S2}, \mathrm{GI}) \quad (5)$$

$n = 233$, $r^2 = 0.7689$, $r^2(\mathrm{pred}) = 0.7688$, $s = 0.75$, $F = 769$ (training set); $n = 41$, $r^2 = 0.6421$, $r^2(\mathrm{pred}) = 0.4241$, $s = 1.14$, $F = 70$ (test set).

The model gives a further example of the proposed algorithm to achieve acceptable models for toxicity predictions. The toxicity data set we used was particularly varied, including pesticides of different chemical natures (aromatic or not, chlorinated, triazines, pyrethroids, ureas, etc.) and different modes of toxic action (insecticides, herbicides, etc.). The complexity of the data set required more invariants than in previous models.[2–10] However, the method offers a useful level of performance, of the descriptors used, in terms of simplicity, since they are only 2D and can thus be calculated very fast.

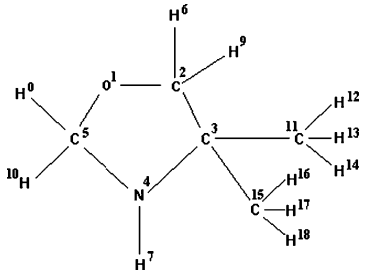
Table 2. Statistical characteristics of the OCWLI and OCWLGI toxicity models, based on Eqs. 1 and 2

Kind of local invariant	Number of optimized parameters	Training set (233)			Test set (41)		
		r^2	s	F	r^2	s	F
OCWLI based on the LHFG descriptor calculated with Eq. 1							
^{0}EC	17	0.4867	1.119	219	0.3293	1.322	19
^{1}EC	28	0.5325	1.068	263	0.4835	1.190	37
^{2}EC	55	0.6208	0.961	378	0.5017	1.233	39
P2	25	0.5072	1.096	238	0.4717	1.194	35
P3	32	0.5612	1.034	295	0.4673	1.237	34
S2	40	0.6522	0.921	433	0.4894	1.317	37
OCWLGI based on LHFG descriptor calculated with Eq. 2							
^{0}EC	58	0.6710	0.896	471	0.5478	1.103	47
^{1}EC	69	0.7068	0.846	557	0.5829	1.071	54
^{2}EC	96	0.7783	0.735	811	0.5706	1.198	52
P2	66	0.6754	0.890	481	0.5686	1.102	51
P3	73	0.7152	0.833	580	0.5640	1.158	50
S2	81	0.7689	0.751	769	0.6421	1.144	70

Table 3. Correlation weights for calculating ${}^{0}X_{CW}$(S2, GI)

a_k	CW	$S2_k$	CW	CC	CW	HB_o	CW	HB_n	CW
H	−0.312	2	7.046	1000	0.847	2000	3.885	3000	0.311
B	15.042	3	−2.182	1050	2.747	2001	7.372	3001	2.766
C	0.876	4	0.486	1600	2.255	2002	7.510	3002	5.040
N	−2.289	5	0.358	1603	10.820	2003	9.431	3003	13.449
O	−1.237	6	0.151	1650	3.659	2004	−1.890	3004	8.790
F	0.160	7	−0.515	1653	20.308	2005	20.311	3006	18.886
P	2.589	8	−0.122			2006	−22.068	3100	1.470
S	5.967	9	0.458			2100	4.167	3101	9.678
Cl	3.593	10	0.454			2101	6.941	3102	4.278
As	9.290	11	1.009			2102	4.876	3200	0.627
Br	5.728	12	−0.559			2103	11.607		
Sn	18.017	13	−0.234			2104	20.457		
J	4.050	14	−0.235			2105	41.382		
		15	−0.161			2200	7.941		
		16	−0.367			2201	5.799		
		17	−3.131			2202	6.131		
		18	0.770			2203	6.977		
		19	−0.492			2300	4.390		
		20	0.139			2301	8.417		
		21	−1.627			2302	6.268		
		22	2.474			2400	13.855		
		23	35.861			2402	5.039		
		24	−26.348			2403	−1.748		
		25	−23.443			2504	53.250		
		26	−8.731			2600	11.381		
		27	32.548						
		28	18.452						

Table 4. Calculating ${}^{0}X_{CW}$(S2, GI) for LHFG of 4,4-dimethyloxazolidine



Chemical element, a_k	No.	$S2_k$	CW (a_k)	CW ($S2_k$)
O	1	11	−1.237	1.009
C	2	15	0.876	−0.161
C	3	15	0.876	−0.161
N	4	16	−2.289	−0.367
C	5	9	0.876	0.458
H	6	7	−0.312	−0.515
H	7	8	−0.312	−0.122
H	8	6	−0.312	0.151
H	9	7	−0.312	−0.515
H	10	6	−0.312	0.151
C	11	11	0.876	1.009
H	12	6	−0.312	0.151
H	13	6	−0.312	0.151
H	14	6	−0.312	0.151
C	15	11	0.876	1.009
H	16	6	−0.312	0.151
H	17	6	−0.312	0.151
H	18	6	−0.312	0.151

CC = 1050: CW (1050) = 2.74678
HB_o = 2001: CW (2001) = 7.37189
HB_n = 3001: CW (3001) = 2.76625

${}^{0}X_{CW}$(S2, GI) = 13.160.

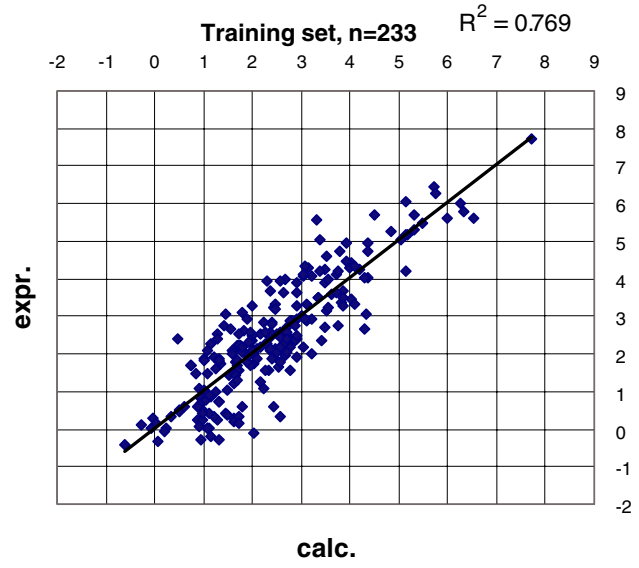


Figure 1. Experimental values and those calculated with Eq. 5 for toxicity on the training set.

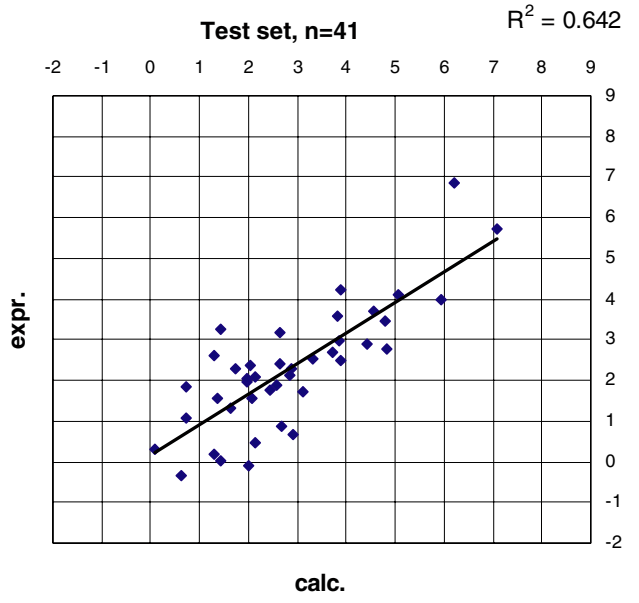


Figure 2. Experimental values and those calculated with Eq. 5 for toxicity on the test set.

This new version of optimization of correlation weights of solely local invariants of the LHFG gives reasonably good models of toxicity in the setting examined. The best model was obtained by correlation weighting of the cyclicity code, hydrogen indices on oxygen and nitrogen, together with valence shells of second order.

According to Ref. 17, correlation coefficient for models of toxicity on the trout ranges from 0.26 to 0.99. In Ref. 18, excellent statistical characteristics on all substances under consideration ($R^2 = 0.945$) are accompanied by six outliers on an external test set of 33 substances. However, these authors took a disputable approach to final validation, namely removing 20 compounds: it might have been more logical to use these 20 compounds as an external test set. It is more typical to construct models of toxicity, in general, and toxicity toward trout, in particular, for separate congeneric classes of chemicals.[19] However, searching for as universal as possible a model, should also be an important task of biochemistry.

Comparison with other reports on QSAR modeling for the rainbow trout has shown that the statistical characteristics of Eq. 5 are reasonably good.

4. Conclusions

1. We describe the optimization of correlation weights of local and global graph invariants on a training set, for predicting the toxicity values on pesticides of an external test set.
2. Taking into account information on the quality of cycles in the molecular structure one can improve the statistical characteristics of the model. In other words, toxicity for the rainbow trout is sensitive to cycles in the molecular structure of pesticides.

Acknowledgment

This work was funded by EU contract QLK5-CT-2002-00691, Demetra.

References and notes

1. Benfenati, E.; Piclin, N.; Roncaglioni, A.; Varì, M. R. *SAR QSAR Environ. Res.* **2000**, *12*, 593.
2. Toropov, A. A.; Toropova, A. P. *Russ. J. Coord. Chem.* **2002**, *28*, 877.
3. Toropov, A. A.; Schultz, T. W. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 560.
4. Toropov, A. A.; Benfenati, E. *J. Mol. Struct. (THEOCHEM)* **2004**, *676*, 165.
5. Toropov, A. A.; Benfenati, E. *J. Mol. Struct. (THEOCHEM)* **2004**, *679*, 225.
6. Toropov, A. A.; Toropova, A. P. *J. Mol. Struct. (THEOCHEM)* **2002**, *578*, 129.
7. Toropov, A. A.; Toropova, A. P. *J. Mol. Struct. (THEOCHEM)* **2003**, *637*, 1.
8. Toropov, A. A.; Toropova, A. P. *J. Mol. Struct. (THEOCHEM)* **2002**, *581*, 11.
9. Toropov, A. A.; Toropova, A. P. *J. Mol. Struct. (THEOCHEM)* **2001**, *538*, 287.
10. Toropov, A. A.; Toropova, A. P.; Nesterova, A. I.; Nabiev, O. M. *Chem. Phys. Lett.* **2004**, *384*, 357.
11. Randic, M.; Basak, S. C. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 261.
12. Randic, M. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 627.
13. Lukovits, I.; Nikolic, S.; Trinajstic, N. *Chem. Phys. Lett.* **2002**, *354*, 417.
14. Amic, D.; Lucic, B.; Nikolic, S.; Trinajstic, N. *Croat. Chem. Acta* **2001**, *74*, 237.
15. Toropov, A. A.; Nesterov, I. V.; Nabiev, O. M. *J. Mol. Struct. (THEOCHEM)* **2003**, *637*, 37.

16. Roncaglioni, A.; Benfenati, E.; Boriani, E.; Clook, M. *J. Environ. Sci. Health., Part B* **2004**, *B39*, 641.
17. Delistraty, D.; Taylor, B.; Anderson, R. *Ecotoxicol. Environ. Saf.* **1998**, *39*, 195.
18. Tao, S.; Xi, X.; Xu, F.; Dawson, R. *Water Res.* **2002**, *36*, 2926.
19. Roy, D. R.; Parthasarathi, R.; Maiti, B.; Subramanianb, V.; Chattaraj, P. K. *Bioorg. Med. Chem.* **2005**, *13*, 3405.