QSPR Study on the Bioconcentration Factors of Nonionic Organic Compounds in Fish by Characteristic Root Index and Semiempirical Molecular Descriptors

Melek Türker Saçan,*,† Safiye Sag Erdem,‡ Gül Altınbas Özpınar,‡ and Isıl Akmehmet Balcıoglu†

Institute of Environmental Sciences, Bogaziçi University, Bebek, Istanbul, Turkey, and Faculty of Arts and Sciences, Chemistry Department, Marmara University, Göztepe, Istanbul, Turkey

Received September 29, 2003

The characteristic root index (CRI) was modeled together with four semiempirical molecular descriptors, namely—energies of the highest occupied and the lowest unoccupied molecular orbital (E_{HOMO} and E_{LUMO}), heat of formation (ΔH_f), and dipole moment (μ)—to predict the fish bioconcentration factor (BCF) of 122 nonionic organic compounds. The best fit equation found by "forward multiple linear regression" showed that the topology based CRI was the most important parameter. The addition of quantum chemical descriptors made only a slight improvement in the predictive capability of the Quantitative Structure−Property Relationship (QSPR) model. The CRI was followed by E_{HOMO}. A two-parameter equation with a correlation coefficient of $r = 0.921$ was obtained for a diverse set of nonionic organic chemicals. Statistical robustness of the developed model was validated by modified jackknife tests where random deletion of a class of compounds and specific deletion of a set of compounds were both performed. The predictive accuracy of the proposed model was compared with the commonly used K_{ow} model and recently published studies in which BCF models were developed. Particular emphasis has been made to clearly define the boundaries for the application of the alternative developed model as well as the quality of estimates.

1. INTRODUCTION

A considerable amount of halogenated organic compounds such as polychlorinated biphenyls, polybrominated biphenyls, chlorinated aliphatic hydrocarbons, polychlorinated benzenes, polybrominated benzenes, polychlorinated anilines, polychlorinated nitro benzenes and phenols, and alkyl benzenes and phenols is found in our environment. Most of them are persistent in the environment and show a tendency to accumulate in biota, soils, and sediments and are also dispersed in the atmosphere.[1] The fate of these chemicals in the environment is controlled by their biological, chemical, and physical properties heavily depending on their structures. One chemical property of interest in modeling the fate and persistence of these chemicals in the environment is bioconcentration factor (BCF). This quantity gives a measure of partitioning of compounds between organisms and their surrounding environment. Barron[2] reported that fishes with an average lipid content of 4.8% are good model animals for bioconcentration relationship studies. Since the experimental determination of BCF values is expensive and time-consuming and experimental data are not available for all chemicals in use, many researchers tend to use estimation methods to supply the missing data. Numerous correlations have been developed relating BCF to the *n*-octanol/water partition coefficient[3] (K_{ow}) as well as other descriptors, for nonionic pollutants.[4−8] Though good results were achieved in some studies, not all BCF always had a good correlation with K_{ow} with any kind of compound, and the application of these methods is limited by the availability of parameter data. Govers[6] et al. modeled BCF and partition coefficients, using molecular weight and molecular connectivity indices. Lu[7,8] et al. used only connectivity indices and both connectivity indices and polarity factors to estimate BCF of nonionic organic compounds. They found that molecular connectivity indices were not good descriptors for polar compounds. When polarity correction factors were introduced into the linear molecular connectivity model, the BCF estimation quality for polar compounds was much increased.

The Characteristic Root Index (CRI) model which comprises all possible orders of path-type valence connectivity indices has been demonstrated correlating many physicochemical properties of organic compounds including solubility,[9] K_{ow},[10] soil-sorption coefficient[11] (K_{oc}), and vapor pressure.[12] Knowing the reasonable literature correlations between the CRI model and the physicochemical properties mentioned above it seems logical to examine the relationship of topology based CRI model with the BCF. Compared to these physicochemical parameters, however, the bioconcentration process becomes far more complicated and a single molecular descriptor may be inefficient in describing the bioconcentration process.

The objective of this study is to describe and discuss a Quantitative Structure−Property Relationship (QSPR) modeling technique for the BCF prediction of 122 halogenated organics, alkyl benzenes, and phenols based on the CRI and utilization of multiple linear regressions in an attempt to increase the predictive power of the CRI model by including semiempirical molecular descriptors.

2. MATERIALS AND METHODS

2.1. Data Set. Log BCF data for 122 nonionic organic compounds were taken from Lu[8] et al. and Devillers[3] et al.

* Corresponding author phone: +90 212 358 15 40/1398; e-mail: msacan@boun.edu.tr.
† Bogaziçi University.
‡ Marmara University.

10.1021/ci0342167 CCC: $27.50
Published on Web 03/10/2004

The studied compounds included halogenated benzenes, halogenated biphenyls, chlorinated aliphatic hydrocarbon, polychlorinated anilines, polychlorinated nitro benzenes and phenols, and alkyl benzenes and phenols. These chemicals are ubiquitous contaminants in the environment due to their wide use in industry and agriculture. The data set contains test species related to several fish such as rainbow trout, guppies, fathead minnows, bluegill sunfish, golden ide, etc.

2.2. Descriptors. The CRI index[9] is a hybrid of path-type valence connectivity index and distance matrix. It is the insertion of valence connectivity into a distance matrix. The graph, *G*, is defined as a square matrix with the entries, w_{ij}, representing the weighted distance traversed in moving from vertex *i*, to vertex *j* in *G*. Thus, the constructed matrix comprises all possible orders of path-type valence connectivity index for a molecule, except zero order.

The starting point of derivation of the CRI is to draw the structural graph, *G*, of a molecule and write the valence values of vertices along the path concerned. Then, by definition, w_{ij} is a weighted path number which appears in the matrix, at the intersection of row/column *i/j* and *j/i* and calculated by

$$w_{ij} = \frac{1}{\sqrt{(m \times n \times p \times ...z)}} \quad (1)$$

where *m, n, p,...z* are the valence values of each vertex along the weighted path. The number of vertex valences that appear in the formula depends on the number of vertices along the path *i* and *j* which joins the vertex *i* and *j*. The entries, w_{ij}, of the matrix are calculated by considering the shortest path to any other vertex. The main diagonal of the matrix contains only zeros, because there is no path bonding to the atom itself. Therefore, entries of this sort (w_{ii}, w_{jj}) are zero. The following procedure after the construction of the matrix is to convert it to a polynomial form using Bocher's[13] formula, which states that the sum of the diagonal elements of a square matrix is equal to the sum of its characteristic values. The constructed matrix was loaded in Excel 2002, and the program written in TURBO PASCAL for personal computer uses this Excel file and converts the stored matrix into the characteristic polynomial. Then the CRI was obtained by summing up the positive characteristic roots of the polynomial calculated with SCIENTIFIC WORKPLACE 3.0.

In this study, the individual primary semiempirical parameters influencing BCF were inclusive of the energy of the highest occupied molecular orbital (E_{HOMO}), the energy of the lowest unoccupied molecular orbital (E_{LUMO}), dipole moment (μ), and heat of formation (ΔH_f). They were calculated by the quantum chemical PM3 method[14] which utilizes semiempirical parameters. Geometries of 122 nonionic organic compounds were fully optimized. We performed conformational analysis with the PM3 method in the gas phase and with the SM54[15] method in aqueous medium to check the variance in the dipole moment as a function of conformation and test the stability of the model for this variance. SPARTAN PRO[16] software was used for quantum chemical computations.

2.3. Model Development and Robustness Test. Multiple linear regressions were used to fit the bioconcentration data of compounds in the data set. The coefficient of determination for multiple regressions (*r*) and the standard error of the estimate (SE), Fisher statistic (*F*), and *t*-values for individual variables were taken into consideration in testing the quality of the regression. Additionally, the studentized residual was used as an additional statistic in testing the quality of the regression. The predictive performance of the model was checked by the classical internal validation. A modified jackknife test (leave-three-out, leave less than 10% out) was used for regression validation. To verify predictive stability of the model, a strongest cross-validation by the leave-subclass-out procedure (repeated 7 times) was also performed. A multiple linear regression was performed after each deletion to calculate the jackknife coefficient of determination and then compared with the same parameter derived prior to deletion of a chemical group for purpose of robustness evaluation. All of the statistical analyses were performed using STATISTICA 6.0 Software.

3. RESULTS AND DISCUSSION

The forward stepwise multiple linear regression (MLR) was employed in the modeling of experimental log BCF values ranging from 0.16 to 5.92 and resulted in the following QSPR model. The CRI and calculated semiempirical molecular descriptors were selected as independent variables, log BCF as a dependent variable.

$$\log \mathrm{BCF} = 6.001 + 0.933\mathrm{CRI} + 0.660E_{\mathrm{HOMO}} - 0.530E_{\mathrm{LUMO}} - 0.129\mu \quad (2)$$

$$r = 0.926,\ \mathrm{SE} = 0.584,\ F^{4,117} = 177.5,\ n = 122$$

A preliminary study on the model development comprised one additional parameter, namely heat of formation (ΔH_f). However, the stepwise multiple linear regression excluded this parameter from the regression equation. The results of regression analysis indicated that the five observed values (2,3-dichlorobiphenyl, 2,2′,3,4,5′-pentachlorobiphenyl, 2-methyl-4,6-dinitrophenol, 2,3,4,5-tetrachloronitrobenzene, and 2,4,6-tribromophenol) of BCF were discrepant from the others. The existence of outliers would affect the predictive power of the model. Furthermore, the correlation of relevant pairs of explanatory variables is needed to be tested when three or more variables are involved in multicollinearities. The ideal situation in regression is that all of the multiple correlation coefficients are zero, and all of the variance inflation factors (VIFs) are unity. Generally, a value of 1.0 indicates no correlation, a value greater than 2.0 is usually considered problematic. Values over 10.0 indicate an unstable regression that must be reexamined. The correlation coefficient ($r = 0.606$) between E_{LUMO} and μ and VIF (Table 1) were found to be slightly high indicating that they are likely to be collinear to some extent. Therefore, their simultaneous presence in the regression equation violates the basic rule of the MLR method. Consequently, a forward stepwise regression was performed to derive a two-variable or three-variable regression model that will account for the outliers. We did examine both E_{LUMO} and μ as independent variables in regressions of log BCF, but neither variable significantly improved the correlation. Similar results were obtained in the prediction of BCF, K_{ow}, and K_{oc}, that μ and ΔH_f were not important parameters for the BCF prediction,[17] whereas μ and E_{LUMO} were not important parameters for the latter

two properties.[18] When these two semiempirical descriptors were removed, the two-variable regression model resulted in eq 3, and all compounds are predicted within the two standard deviation range.

$$\log \mathrm{BCF} = 4.623 + 1.045\mathrm{CRI} + 0.546E_{\mathrm{HOMO}} \quad (3)$$

$$r = 0.921,\ \mathrm{SE} = 0.599,\ F^{2,119} = 332.5,\ n = 122$$

On the other hand, knowing the importance of bioactive conformations for biological systems and the fact that using the most stable conformation for a flexible molecule may affect the quality of the regression more or less, we examined the conformational flexibility of all the compounds in our data set in more detail as isolated molecules or in aqueous medium. Most of the compounds in the data set have restricted bond rotation or do not show conformational isomerism. Therefore, only six compounds in the data set generate two or three conformations in our calculations which lead to a change in the dipole moment in aqueous medium. These are 1,1,2,2-tetrachloroethane, 1,2-dichloroethane, 2,4-dichlorophenol, 2-chlorophenol, 3-chlorophenol, and 2-methyl-4,6-dinitrophenol with the new dipole moment values 1.52, 2.06, 1.84, 1.92, 1.94, and 6.54, respectively. We performed regression with these new dipole moment values and obtained the following equation.

$$\log \mathrm{BCF} = 4.580 + 1.039\mathrm{CRI} + 0.539E_{\mathrm{HOMO}} - 0.015\mu \quad (4)$$

$$r = 0.921,\ \mathrm{SE} = 0.602,\ F^{3,118} = 220.2,\ n = 122$$

The r values stayed the same, and the coefficients of the variables differed slightly compared to eq 3. The new dipole moment values have no contribution to the regression. Having known that the use of fewer descriptors has important advantages when constructing regression equations, we neglected the dipole moment from the regression.

Two parameters proved to be significant for the obtained QSPR model (eq 3): the CRI and E_{HOMO}. It explained approximately 85% of the variance. Therefore, we examined a two-variable regression model (eq 3) in more detail. The t-test method was used to test the correlation of each independent variable and dependent variable. And, student t-values for partial correlation coefficients in eq 3 were 25.064 and 5.156 for the CRI and E_{HOMO}, respectively. This indicated that the CRI was the most important factor for the prediction of BCF. Considering eq 3, the higher the CRI and E_{HOMO} were, the higher the BCF were. Parameters entered the model reflected the complexity and branching of the molecular structure and reactivity of the tested chemicals. The CRI which comprises all possible orders of path-type molecular connectivity indices encodes global molecular properties such as size, volume, and surface area emphasizing the strong dependence of BCF of nonionic organic compounds on the compound size. Local structural properties and possibly long-range interactions described by path-type molecular indices may also be encoded in the CRI. Since E_{HOMO} descriptor is related to the ability of solutes to participate in electron pair donor−acceptor interactions with biotarget molecules, it indicates that the charge transformation may exist among the tested compound molecule and target-molecule in fish. The observed values, fitted values, and the residuals for 122 chemicals are also listed in Table 2 and are plotted in Figure 1.

Table 1. Correlation Coefficient Matrix for Independent Variables and the Variance Inflation Factors (VIFs) for Eqs 2 and 3

	correlation matrix				VIF	
	CRI	E_{HOMO}	E_{LUMO}	μ	by eq 2	by eq 3
CRI	1.000				2.084	1.001
E_{HOMO}	−0.026	1.000			1.166	1.001
E_{LUMO}	−0.275	0.281	1.000		3.166	
μ	−0.355	−0.138	−0.606	1.000	3.114	

The relative errors in the model for individual chemicals studied were examined using the studentized residuals compared with the calculated log BCF values (Figure 2). As can be seen in Figure 2, the studentized residuals are symmetrically distributed around zero with no specific trend. The four chemicals with the highest residuals are 2,4,6-tribromophenol, 2,3-dichlorobiphenyl, 2,2′,5,5′-tetrabromobiphenyl, and 2,2′,3,4,5′-pentachlorobiphenyl. One possible explanation for this result could be the potential inaccuracies in the measured data itself, since the maximum difference of the multiple measured BCF values for some of the compounds used in this study reported as 2.5 log units.[8] However, more than 91% of plots show good agreement within ±1.0 log unit and the response plot (Figure 1) confirms the model quality, because it shows a good alignment of the nonionic organic compounds along the optimal line.

Considering the relatively small number of collected experimental BCFs in the study, the database was not divided into modeling and testing sets for purposes of model validation. Instead, an alternative approach using a modified jackknife test[19] was applied. We performed a jackknife test with removal of less than 10% randomly selected compounds in each run, and the regression was rerun for all the other observed values. The overall results of the deletion study (leave three/nine-out method) are summarized in Table 3.

The modified jackknife tests validated that the developed CRI-E_{HOMO} based model was statistically robust. The average r values do not have any unduly high variation suggesting that the data set is fairly consistent and the models are not biased by any particular data point. Additionally, the robustness of the model with regard to compound type was tested by deleting each category in sequence one at a time. The jackknife r values calculated on the remaining six categories of chemicals are plotted in Figure 3.

Multiple correlation coefficient decreased to 0.815 when PCBs were deleted implying that PCBs displayed the most significant influence on the model. The decline of the jackknife r seems principally caused by the decrease of data used for the regression, since PCBs was the biggest class including 36 compounds. On the other hand, jackknife r slightly increased to 0.932 and 0.932, when 17 aniline and 20 nitro compounds were deleted from the regression, respectively. This indicates that the determined model was much influenced by the structures of the training compounds and their numbers. The other jackknife r values varied slightly around the original r (0.921) value.

The physical significance of the developed model could be explainable. To be absorbed in a biological system, a chemical must penetrate a sequence of hydrophobic and hydrophilic barriers, and therefore the bioconcentration

Table 2. Bioconcentation Factors and Molecular Structure Descriptors for 122 Nonionic Organic Chemicals

compounds	logK_{ow}	CRI	-E_{HOMO} (eV)	E_{LUMO} (eV)	μ (debye)	log BCF		
						exp	pred by eq 3	res
1,1,1-trichloroethane	2.49	2.5234	10.75	−0.07	1.380	0.95	1.38	−0.43
1,1,2,2-tetrachloroethane	2.39	2.9801	10.70	−0.10	0.000	0.90	1.89	−0.99
1,1,2,3,4,4-hexachloro-1,3-butadiene	4.78	3.9079	9.37	−0.50	0.241	3.76	3.58	0.18
trichloroethylene	2.42	1.9897	9.38	−0.04	0.491	1.59	1.58	0.01
1,2-dichloroethane	1.48	2.0698	10.68	0.54	0.000	0.30	0.95	−0.65
tetrachloromethane	2.83	2.7088	10.99	−0.63	0.001	1.48	1.45	0.03
trichloromethane	1.97	2.3027	10.84	−0.12	1.016	0.78	1.10	−0.32
hexachloroethane	4.14	3.9907	10.84	−0.56	0.000	2.92	2.87	0.05
pentachloroethane	3.22	3.4849	10.76	−0.37	0.848	1.83	2.38	−0.55
tetrachloroethylene	3.40	2.5788	9.22	−0.32	0.000	1.74	2.28	−0.54
benzene	2.19	1.2238	9.74	0.40	0.000	0.64	0.57	0.07
toluene	2.73	1.6391	9.44	0.38	0.263	1.12	1.17	−0.05
ethyl benzene	3.15	2.0459	9.52	0.37	0.213	1.19	1.56	−0.37
o-xylene	3.12	2.0339	9.30	0.39	0.463	1.24	1.66	−0.42
m-xylene	3.20	2.0279	9.31	0.39	0.267	1.27	1.65	−0.38
p-xylene	3.15	1.9960	9.18	0.36	0.076	1.27	1.69	−0.42
isopropylbenzene	3.72	2.4278	9.53	0.38	0.209	1.55	1.95	−0.40
biphenyl	4.09	2.4084	8.92	−0.36	0.005	2.64	2.26	0.38
1,2,3,4-tetrachlorobenzene	4.64	3.2039	9.28	−0.56	1.008	3.72	2.90	0.82
1,2,3,5-tetrachlorobenzene	4.92	3.2473	9.30	−0.59	0.461	3.36	2.93	0.43
1,2,3-trichlorobenzene	4.05	2.7485	9.38	−0.33	1.350	2.90	2.37	0.53
1,2,4,5-tetrachlorobenzene	4.82	3.2468	9.19	−0.64	0.000	3.61	2.99	0.62
1,2,4-trichlorobenzene	4.02	2.7799	9.24	−0.43	0.666	2.95	2.48	0.47
1,2-dichlorobenzene	3.43	2.2728	9.29	−0.17	1.351	2.43	1.92	0.51
1,3,5-trichlorobenzene	4.19	2.8097	9.59	−0.38	0.001	3.26	2.32	0.94
1,3-dichlorobenzene	3.60	2.2858	9.42	−0.19	0.879	2.65	1.86	0.79
1,4-dichlorobenzene	3.52	2.2750	9.23	−0.24	0.000	2.47	1.95	0.52
hexachlorobenzene	5.31	4.0901	9.31	−0.83	0.000	4.16	3.81	0.35
2,4,5-trichlorotoluene	4.56	3.0889	9.15	−0.40	1.056	3.87	2.85	1.02
chlorobenzene	2.84	1.7607	9.39	0.06	0.953	1.85	1.33	0.52
pentachlorobenzene	5.18	3.6176	9.25	−0.74	0.433	3.45	3.35	0.10
1,2,4,5-tetrabromobenzene	5.13	5.3577	9.96	−0.87	0.000	3.81	4.78	−0.97
1,2,4-tribromobenzene	4.66	4.3372	9.94	−0.66	0.677	3.66	3.72	−0.06
1,3,5-tribromobenzene	4.51	4.2742	10.14	−0.76	0.000	3.70	3.55	0.15
1,3-dibromobenzene	3.75	3.2726	9.95	−0.43	1.064	2.80	2.61	0.19
1,4-dibromobenzene	3.79	3.2573	9.87	−0.32	0.000	2.83	2.63	0.20
bromobenzene	2.99	2.2476	9.81	−0.05	1.184	1.70	1.61	0.09
1,2-dibromobenzene	3.64	3.3061	9.86	−0.45	1.494	2.70	2.69	0.01
2,2′,4,5-tetrachlorobiphenyl	5.69	4.4613	8.96	−0.93	1.497	5.00	4.39	0.61
2,2′,5,5′-tetrachlorobiphenyl	5.79	4.4941	9.07	−0.92	0.003	4.63	4.36	0.27
2,2′,5-trichlorobiphenyl	5.55	3.9522	9.19	−0.28	0.946	4.01	3.73	0.28
2,2′,4,4′-tetrachlorobiphenyl	6.29	4.4487	9.37	−0.41	0.911	4.02	4.15	−0.13
2,2′-dichlorobiphenyl	4.90	3.4059	9.32	−0.07	1.033	3.26	3.09	0.17
2,3′,4′,5-tetrachlorobipenyl	6.07	4.4757	9.16	−0.67	1.237	4.62	4.29	0.33
2,3-dichlorobiphenyl	5.02	3.3954	9.24	−0.35	1.340	4.25	3.12	1.13
2,4,4′-trichlorobiphenyl	5.58	3.8900	9.12	−0.58	0.736	4.63	3.70	0.93
2,4′,5-trichlorobiphenyl	5.68	3.9311	9.13	−0.56	0.708	3.75	3.74	0.01
2,4,5-trichlorobiphenyl	5.90	3.9301	9.09	−0.58	1.043	4.02	3.76	0.26
2,4′-dichlorobiphenyl	5.10	3.4322	9.12	−0.40	1.497	3.55	3.23	0.32
2,5-dichlorobiphenyl	5.16	3.4553	9.12	−0.40	0.321	4.00	3.25	0.75
3,3′,4,4′-tetrachlorobiphenyl	6.63	4.4345	9.10	−0.76	0.533	3.90	4.28	−0.38
3,5-dichlorobiphenyl	5.41	3.4792	9.36	−0.47	1.149	3.78	3.14	0.64
4-chlorobiphenyl	4.63	2.9528	9.06	−0.35	1.054	2.69	2.76	−0.07
4,4′-dichlorobiphenyl	5.58	3.4741	9.03	−0.53	0.000	3.28	3.32	−0.04
2,2′,3,3′-tetrachlorobiphenyl	5.67	4.4149	9.31	−0.29	1.698	4.23	4.15	0.08
2,2′, 4,5,5′-pentachlorobiphenyl	6.65	4.9718	9.24	−0.55	0.702	5.40	4.77	0.63
2,2′,4,4′,5,5′-hexachlorobiphenyl	7.75	5.3952	9.27	−0.64	0.04	4.83	5.19	−0.36
2,2′,4,4′,6,6′-hexachlorobiphenyl	7.55	5.5348	9.57	−0.53	0.000	4.93	5.18	−0.25
2,2′,3,3′,4,4′,5,5′-octachlorobiphenyl	8.68	6.3039	9.35	−0.78	0.545	5.08	6.10	−1.02
2,2′, 3,5′-tetrachlorobiphenyl	5.73	4.4637	9.24	−0.35	1.267	4.84	4.23	0.61
2,2′,4,5′-tetrachlorobiphenyl	5.87	4.3651	9.26	−0.39	0.803	4.84	4.13	0.71
2,2′,6,6′-tetrachlorobiphenyl	5.94	4.4192	9.35	−0.21	0.000	3.85	4.16	−0.31
2,2′,3,4,5′-pentachlorobiphenyl	6.23	4.4391	9.29	−0.49	1.297	5.38	4.21	1.17
2,2′,3′,4,5-pentachlorobiphenyl	6.67	4.8907	9.24	−0.54	1.421	5.43	4.62	0.81
3,3′,4,4′,5-pentachlorobiphenyl	6.98	4.8744	9.16	−0.87	0.695	5.81	4.71	1.10
2,2′,3,3′,4,4′-hexachlorobiphenyl	6.96	5.3125	9.42	−0.57	1.421	5.77	5.02	0.75
2,2′, 3,3′, 6,6′-hexachlorobiphenyl	7.03	5.3998	9.22	0.48	0.727	5.43	5.23	0.20
2,2′,3,4,4′,5-hexachlorobiphenyl	6.82	5.3200	9.29	−0.66	0.934	5.88	5.11	0.77
2,2′,3,4,5,5′-hexachlorobiphenyl	6.75	5.3902	9.28	−0.65	1.054	5.81	5.18	0.63
2,2′,3,4,4′,5′,6-heptachlorobiphenyl	7.04	5.2800	9.29	−0.70	0.466	5.84	5.06	0.78
2,2′,3,3′,4,4′,5,6-octachlorobiphenyl	7.35	6.3413	9.27	−0.78	1.022	5.92	6.18	−0.26
2,2′,3,3′,4,5,5′,6-octachlorobiphenyl	8.91	6.3894	9.27	−0.79	0.463	5.88	6.23	−0.35
2,2′,3,3′,5,5′,6,6′-octachlorobiphenyl	7.73	6.3834	9.20	−0.70	0.000	5.82	6.27	−0.45
2,2′,3,3′,4,4′,5,5′,6-nonachlorobiphenyl	9.14	6.8305	9.30	−0.83	0.473	5.71	6.68	−0.97
2,2′,5,5′-tetrabromobiphenyl	7.31	6.4727	9.86	−0.46	0.146	4.80	6.00	−1.20
2,4,6-tribromobiphenyl	6.42	5.4347	9.80	−0.67	0.536	3.93	4.94	−1.01
4,4′-dibromobiphenyl	5.72	4.3928	9.40	−0.58	0.000	4.19	4.07	0.12
2,4-dichlorophenol	5.53	2.3675	9.09	−0.24	0.521	2.00	2.13	−0.13
pentachlorophenol	5.12	3.2784	9.14	−0.79	1.104	2.99	3.05	−0.06
2,4,6-trichlorophenol	3.69	2.8488	9.12	−0.44	1.002	2.43	2.61	−0.18

Table 2. (Continued)

compounds	logK_{ow}	CRI	-E_{HOMO} (eV)	E_{LUMO} (eV)	μ (debye)	log BCF exp	log BCF pred by eq 3	log BCF res
2-chlorophenol	2.15	1.8586	9.21	−0.02	0.688	2.33	1.53	0.80
3-chlorophenol	2.50	1.8667	9.23	−0.01	0.445	1.30	1.52	−0.22
2-methylphenol	1.95	1.7542	9.06	0.28	1.357	1.03	1.50	−0.47
phenol	1.46	1.3596	9.17	0.29	1.142	1.24	1.03	0.21
4-t-buthylphenol	3.31	2.7910	9.00	0.35	1.192	2.07	2.62	−0.55
2,4-dimethylphenol	2.30	2.1048	8.86	0.31	1.379	2.18	1.98	0.20
4-bromophenol	2.59	2.3494	9.31	−0.03	1.521	1.56	1.99	−0.43
p-sec-buthylphenol	3.08	2.7840	9.02	0.33	1.209	1.57	2.60	−1.03
2-chloroaniline	1.90	1.9269	8.66	0.13	1.467	1.18	1.90	−0.72
3-chloroaniline	1.88	1.9378	8.76	0.12	1.792	1.06	1.86	−0.80
diphenylamine	3.50	2.5635	8.46	0.07	0.979	1.48	2.68	−1.20
pentachloroaniline	4.82	3.7701	8.80	−0.65	1.808	3.78	3.75	0.03
2,3,4,5-tetrachloroaniline	4.27	3.3252	8.77	−0.47	2.161	3.28	3.31	−0.03
2,3,5,6-tetrachloroaniline	4.10	3.3482	8.91	−0.54	1.441	3.03	3.25	−0.22
2,3,4-trichloroaniline	3.68	2.8670	8.71	−0.26	2.245	2.31	2.86	−0.55
2,4,5-trichloroaniline	3.45	2.8963	8.70	−0.35	1.896	2.61	2.89	−0.28
2,4,6-trichloroaniline	3.52	2.9041	8.72	−0.30	1.458	2.73	2.89	−0.16
3,4,5-trichloroaniline	3.32	2.8774	8.75	−0.26	2.423	2.70	2.85	−0.15
4-chloroaniline	1.88	1.9158	8.58	0.10	1.963	0.91	1.93	−1.02
2,4-dichloroaniline	2.78	2.3276	8.65	−0.12	1.854	1.98	2.33	−0.35
3,4-dichloroaniline	2.78	2.4198	−8.67	−0.11	2.307	1.48	2.41	−0.93
aniline	0.90	1.4048	−8.61	0.42	1.295	0.41	1.38	−0.97
2-nitrophenol	1.79	1.5831	−9.90	−1.22	4.182	1.60	0.86	0.74
2-chloronitrobenzene	2.24	2.0236	−9.99	−1.19	5.272	2.10	1.27	0.83
3-chloronitrobenzene	2.46	1.9745	−10.06	−1.30	4.832	1.89	1.19	0.70
4-chloronitrobenzene	2.39	1.9661	−10.22	−1.36	4.588	2.00	1.09	0.91
2,3-dichloronitrobenzene	3.05	2.4467	−9.82	−1.32	4.860	2.16	1.81	0.35
2,4-dichloronitrobenzene	3.07	2.4820	−10.08	−1.41	4.633	2.07	1.70	0.37
2,5-dichloronitrobenzene	3.09	2.5248	−9.78	−1.36	4.845	2.05	1.92	0.13
3,4-dichloronitrobenzene	3.12	2.4639	−9.82	−1.32	4.860	2.07	1.83	0.24
3,5-dichloronitrobenzene	3.09	2.5280	−10.03	−1.46	4.404	2.23	1.78	0.45
2-methyl-4,6-dinitrophenol	2.12	2.1252	−10.53	−1.82	4.458	0.16	1.08	−0.92
3-nitrophenol	2.00	1.5792	−9.96	−1.18	6.121	1.40	0.83	0.57
pentachloronitrobenzene	4.77	3.8934	−9.74	−1.52	4.218	2.40	3.37	−0.97
2,3,4,5-tetrachloronitrobenzene	4.57	3.4640	−9.77	−1.62	4.248	1.89	2.90	−1.01
2,3,5,6-tetrachloronitrobenzene	3.89	3.4807	−9.70	−1.37	4.443	3.20	2.96	0.24
2,3,4-trichloronitrobenzene	3.68	2.8971	−9.88	−1.50	4.439	2.20	2.25	−0.05
2,4,5-trichloronitrobenzene	3.48	3.0156	−9.79	−1.53	4.444	1.84	2.42	−0.58
4-nitroaniline	1.39	1.6215	−9.42	−1.01	6.636	0.64	1.17	−0.53
3-nitroaniline	1.37	1.6506	−9.29	−1.06	5.695	0.92	1.27	−0.35
2-nitroaniline	1.85	1.6568	−9.16	−1.00	4.88	0.91	1.34	−0.43
2,4,6-tribromophenol	4.13	4.3189	−9.56	−0.80	1.02	2.71	3.91	−1.20

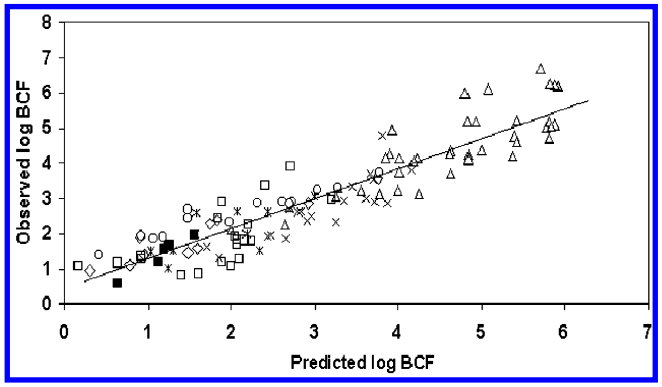

Figure 1. Plot of observed log BCF versus fitted log BCF using eq 3 for seven groups of nonionic compounds. (○ halogenated anilines; □ halogenated nitro aromatics; ■ alkyl benzenes; ◇ chlorinated aliphatic hydrocarbons; * phenols; × halogenated benzenes; △ polyhalogenated biphenyls).

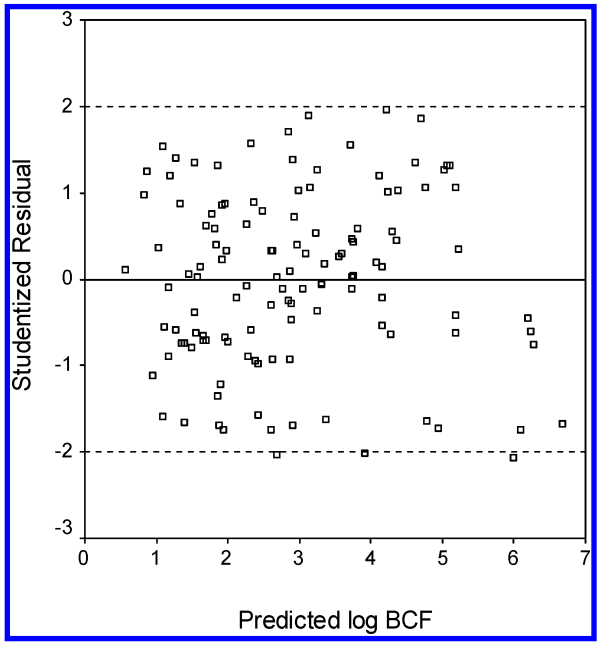

Figure 2. Plot of studentized residual versus calculated log BCF.

process is controlled by polar and nonpolar interactions among water and fish.[8] Most of the estimated log BCF values of aniline compounds were found to be higher than the experimental values. Anilines either may not share a common mode of action which is based on their reactivity and/or polar interactions are utilized to provide an attractive force between anilines and water, rather than fish tissue. In addition, a significant body of evidence now supports the role of metabolism in producing BCF values that are lower than expected.[20] Examples include chlorinated anilines[21] and certain organophosphates.[22] Thus, these compounds cannot reach their maximum in bioaccumulation. On the other hand, the predicted BCF values of halogenated nitro benzenes were

Table 3. Summary of Results of Random Deletion Test

no. of cases deleted (<10% of *N*)	no. of regression runs	log BCF	
		av *r*	av SE
3	40	0.921	0.599
9	13	0.921	0.597
av		0.921	0.598

found to be lower than the experimental ones. It is likely that part of the estimation error might be caused by the observed data themselves. The data collected for these chemicals come from a single source,[23] making it difficult to avoid any potential systematic error.

The selection of test species also induces a source of error in the estimation of BCFs. In aquatic toxicology, it is generally accepted that lipit content of an animal is an important determinant of bioconcentration.[2] The BCF data was not corrected for fish lipid content because these data are not available in most publications. Normalizing data to lipid content can eliminate some interspecies variation in BCF values. Error in the prediction model may also be generated from other sources. Uncertainty associated with a given measured BCF may arise from the exposure concentration, test condition, duration of the experiment, and the determination of the concentration in water and fish. It is also likely that the two-descriptors selected for the modeling may still not be sufficient for a full explanation of all of the structural features of these chemicals. Nevertheless, these

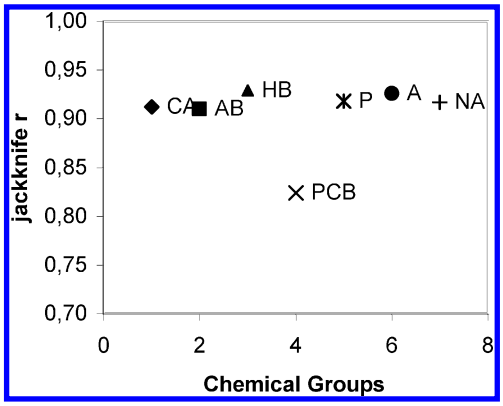

Figure 3. Distribution of the jackknife *r* resulted from deletions of categorized chemicals. (◆ chlorinated aliphatic hydrocarbons (CA), ■ alkyl benzenes (AB), ▲ halogenated benzenes (HB), × polychlorinated biphenyls (PCBs), ∗ phenols (P), ● anilines (A), + nitro aromatics (NA)).

are the common problem of all the other estimation methods based on topological and other descriptors which uses experimental values of BCF.

It is of interest to compare the results of the current study (the CRI-E_{HOMO} based model) with those of recently published studies in which BCF models were developed (Table 4). The CRI-E_{HOMO} based model is superior to the models reported in Table 4 in terms of number of parameters. Use of fewer descriptors has important advantages when con-

Table 4. QSPR Bioconcentration Factor Prediction Models

model no.	model type	no. of parameters	*n*	r^2	SE	*F*	investigators
1	MLR	5	9/44[b]	0.792	0.432	24	Fatemi et al. (2003)[25]
2	MLR	4	27	0.931	0.137	103	Wei et al. (2001)[17]
3	NLR[a]	13	239	0.810	0.615	NR[c]	Lu et al.(2000)[8]
4	MLR	5	80	0.907	0.364	NR	Lu et al.(1999)[7]
5	MLR	2	122	0.851	0.599	338	current study

[a] NLR = nonlinear regression. [b] Notation indicates: $n_{\text{prediction set}}/n_{\text{training set}}$. [c] NR = not reported.

Table 5. Predicted log BCF Values of Chlorophenols, PCBs, and Alkylphenols with Eq 3

compound	CRI	$-E_{HOMO}$ (eV)	log BCF	compound	CRI	$-E_{HOMO}$ (eV)	log BCF
			Chlorophenols				
4-chlorophenol	1.845	9.009	1.61	2,3,6-	2.822	9.286	2.490
2,3-	2.338	9.323	1.97	2,4,5-	2.836	8.985	2.660
2,6-	2.349	9.193	2.04	3,4,5-	2.824	9.113	2.590
3,4-	2.356	9.055	2.12	2,3,4,5-	3.266	9.050	3.082
3,5-	2.391	9.422	1.97	2,3,4,6-	3.278	9.151	3.042
2,3,4-	2.795	9.069	2.57	2,3,5,6-	3.273	9.193	3.016
2,3,5-	2.839	9.114	2.60				
			PCBs				
2,6-dichloro	3.437	9.340	3.112	2,3,4,5-	4.347	9.215	4.13
3,3′-	3.447	9.283	3.152	2,4,4′,5-	4.414	9.105	4.26
3,4-	3.428	9.081	3.236	2,3,5,6-	4.187	9.118	4.01
2,2′,3-trichloro	3.947	9.269	3.684	2,2′,3,4,4′-	4.905	9.391	4.63
2,2′,4-	3.934	9.338	3.635	2,2′,3,5′,6-	4.956	9.231	4.76
2,3,3′-	3.924	9.320	3.634	2,2′,4,4′,5-	4.931	9.248	4.73
2,3,4′-	3.964	9.204	3.735	2,2′,4,6,6′-	4.936	9.423	4.64
2,4,6-	3.943	9.453	3.580	2,3,3′,4,4′-	4.898	9.183	4.73
3,4,4′-	3.886	9.070	3.72	2,3,3′,4′,6-	4.898	9.275	4.68
2,2′,3,4-	4.423	9.344	4.15	2,3′,4,4′,5-	4.916	9.139	4.77
2,2′,3,5-	4.464	9.200	4.26	2,3,4,5,6-	4.817	9.178	4.64
2,2′,5,6-	4.477	9.179	4.28	2,2′,4,5,5′,6-	5.449	9.280	5.26
2,3,4,4′-	4.319	9.391	4.01	2,2′,3,3′,4,5′,6,6′-	6.387	9.210	6.27
			Alkyl Phenols				
2,3-dimethyl	2.102	8.990	1.89	2,6-	2.114	8.961	1.92
2,5-	2.130	8.972	1.93	3,4-	2.111	8.878	1.95

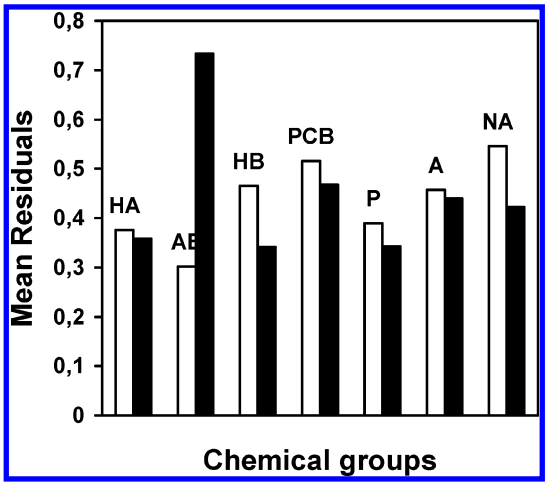

Figure 4. Mean residuals of various chemical classes. (□ the CRI-E_{HOMO} based model; ■ the model 3).

structing regression equations.[24] Of the results reported in Table 4, model numbers 1, 2, and 4 have lower errors than the CRI-E_{HOMO} based model; however, it must be noted that the total number of compounds in their data set was 34%, 58%, and 78% less than the data set used by our group, respectively. The model 3 utilized by Lu et al.[8] has been analyzed in more details since their data set was structurally diverse. Jackknife *r* values displayed that PCBs have the most significant influence both on the current model and the model 3. However, chlorinated dibenzo-dioxins and chlorinated dibenzofurans which are not included in the current model exhibited relatively more effects on the latter model.

Furthermore, to compare the predictive performance of the developed model with the model 3, RMS (i.e., root-mean-square error) values (eq 5) of both models were calculated.[3]

$$\mathrm{RMS} = ((\sum \log \mathrm{BCF}_{obs} - \log \mathrm{BCF}_{calc})^2)/\text{number of observation})^{1/2} \quad (5)$$

The RMS values are 0.592 and 0.596 for the CRI-E_{HOMO} based model and the model 3, respectively. Although the RMS results confirm that these two models yield the equivalent results, it is interesting to note that model 3 with five connectivity indices and eight correction factors (polarity corrections for polar functional groups) was not able to discriminate between two pairs, 2,4,4′- and 2,4′,5- trichlorobiphenyl and 2,2′,3,4,4′,5- and 2,2′,3,4,5,5′-hexachlorobiphenyl, of PCB isomers, and one pair, 2,4- and 3,4-dichloroaniline, and one triplet, 2,4,5- and 2,4,6- and 3,4,5-trichloroaniline, of chlorinated aniline isomers, and one pair, 2,4- and 2,5-dichloronitrobenzene, of chlorinated nitrobenzene isomers. This feature reveals that the CRI-E_{HOMO} based model is more sensitive to the substitution pattern (Table 2) than the model 3, since it can differentiate between the BCF value of compounds within isomeric groups. Additionally, the analysis of absolute mean residuals of the two models has been performed on a chemical class by class basis (Figure 4). The mean residuals of alkyl benzenes were 0.30 and 0.73 for the current model and the model 3, respectively. Although there were seven chemicals in the sample for this class, their structure was fairly described by the model 3. The mean residuals of other classes are only slightly different from each other (Figure 4). The overall averages of the groups as a whole were 0.484 and 0.477 log unit for the CRI-E_{HOMO} based model and the model 3, respectively.

Moreover, the predictive accuracy for the CRI-E_{HOMO} based model was compared with the results of commonly used K_{ow} models. Seven linear and nonlinear BCF models based on log K_{ow} have been compared in order to estimate their accuracy, predictive power, and domain of application.[3] They showed that for chemicals with log $K_{ow} < 6$, the different models yielded equivalent results. At the opposite, for highly hydrophobic chemicals (log $K_{ow} > 6$), the bilinear model[26] was superior to the other studied models. Experimentally determined K_{ow} values were collected for the data set from the literature[3,8,10,27,28] (Table 2). The log K_{ow} values for all the studied compounds ranged from 0.90 to 9.14, and 17% of the data set have log $K_{ow} > 6$. Therefore, the predictive accuracy of the CRI-E_{HOMO} based model was compared with the results of the bilinear K_{ow}-based model.[26] The RMS values are 0.59 and 0.66 (for $n = 342$)[3] for the CRI-E_{HOMO} based model and the bilinear K_{ow}-based model, respectively. This result confirms that the current model is not inferior to that achieved using K_{ow}. Overprediction of BCF is usually pronounced for highly hydrophobic chemical.[20] However, the BCFs of compounds having log $K_{ow} > 6$ are not systematically over- or underestimated by the developed model. It is superior to the K_{ow}-based model in many aspects, including model structure and parameter availability. However, data on traditional descriptors such as K_{ow} have to be experimentally determined or estimated/calculated by other methods. In general, the reliability of calculation methods decreases as the complexity of the compound under study increases.

The level of accuracy of the developed model is quite good considering the many sources of error that may impact the model. Because of its high statistical significance, the validated model eq 3 has been used to predict the log BCFs of those compounds where there are no experimental measurements (Table 4). Two compounds—2,4,5-trichlorophenol and 2,3,5,6- tetrachlorophenol—out of 43 have reported BCF values[29] ranging from 2.40 to 3.28 and from 2.17 to 3.16, respectively. However, it should be noted that the predicted BCF values of these compounds (2.68 and 3.02, respectively) are within the range of reported values. Finally, this result also supports the predictive power of the model.

4. CONCLUSION

The information presented in this study shows that a fairly good relationship existed between the CRI-E_{HOMO} and the log BCF of 122 nonionic compounds with a log BCF values below 6. However, the linearity observed in this model must be checked beyond the BCF values of 6. It was also shown that by using a small set of descriptors, an alternative reliable prediction model was developed for BCF of compounds containing varied groups such as $-CH_3$, $-NO_2$, $-Cl$, $-Br$, $-OH$, and $-NH_2$. All the methods used to check the robustness of the model indicated that our model is stable and has a high prediction power considering the variations in the experimental data. This QSPR equation can be used to predict the BCF values for compounds that have similar structural characteristics with the modeled compounds. The descriptors used in this MLR are attractive because they can be calculated easily and rapidly and are error free. The

developed model, as BCF models published in the literature, does not take into account the transformation of the chemicals in the fish.

ACKNOWLEDGMENT

The financial support of Bogaziçi University Research Fund (Project Number 02Y104) is appreciated.

REFERENCES AND NOTES

(1) Wania, F.; Mackay, D. Tracking the distribution of persistent organic pollutants. *Environ. Sci. Technol.* **1996**, *30*, 390A-396A.
(2) Barron, M. G. Bioconcentration. *Environ. Sci. Technol*. **1990**, *24*, 1612−1618.
(3) Devillers, J.; Bintein, S.; Domine, D. Comparison of BCF Models Based on log P. *Chemosphere* **1996,** *33*, 1047−1065.
(4) Sabljic, A.; Protic, M. Molecular Connectivity: A Novel Method for Prediction of Hazardous Chemicals. *Chem.-Biol. Interact*. **1982**, *42*, 301−310.
(5) Koch, R. Molecular Connectivity Index for Assessing Ecotoxicological Behaviour of Organic Compounds. *Toxicol. Environ. Chem.* **1983**, *6*, 87−96.
(6) Govers, H.; Ruepert, C.; Aiking, H. Quantitative Structure−Activity Relationships for Polycyclic Aromatic Hydrocarbons: Correlation between Molecular Connectivity, Physicochemical Properties, Bioconcentration and Toxicity in *Daphnia pulex*. *Chemosphere* **1984**, *13*, 227−236.
(7) Lu, X.; Tao, S.; Cao, J.; Dawson, R. W. Prediction of Fish Bioconcentration Factors of Nonpolar Organic Pollutants Based on Molecular Connectivity Indices. *Chemosphere* **1999**, *39*, 987−999.
(8) Lu, X.; Tao, S.; Hu, H.; Dawson, R. W. Estimation of Bioconcentration Factors of Nonionic Organic Compounds in Fish by Molecular Connectivity Indices and Polarity Correction Factors. *Chemosphere* **2000**, *41*, 1675−1688.
(9) Saçan, M. T.; İnel, Y. Prediction of Aqueous Solubility of PCBs Related to Molecular Structure. *Turk. J. Chem.* **1993**, *17*, 188−195.
(10) Saçan, M. T.; İnel, Y. Application of the Characteristic Root Index Model to the Estimation of Octanol/Water Partition Coefficients: Polychlorinated Biphenyls. *Chemosphere* **1995**, *30*, 39−50.
(11) Saçan, M. T.; Balcıoğlu, I. A. Prediction of the Soil Sorption Coefficient of Organic Pollutants by the Characteristic Root Index Model. *Chemosphere* **1996**, *32*, 1993−2001.
(12) Saçan, M. T.; Balcıoğlu, I. A. Estimation of Liquid Vapor Pressures for Low Volatility Environmental Chemicals. *Chemosphere* **1998**, *36*, 451−460.
(13) Istefanopulos, Y. *State Variables and Linear Control Systems*; BoğaziçiUniversity Publications: 1987; pp 61−63.
(14) Stewart, J. J. P. Optimization of Parameters for Semiempirical Methods. *J. Comput. Chem.* **1989**, *10*, 209−220.
(15) Chambers, C. C.; Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem.* **1996**, *100*, 16385.
(16) PC Spartan Pro, Wavefunction, Inc. 18401 Von Karman, Suite 370, Irvine, CA 92612, U.S.A.
(17) Wei, D.; Zhang, A.; Wu, C.; Han, S.; Wang, L. Progressive Study and Robustness Test of QSAR Model Based on Quantum Chemical Parameters for Predicting BCF of Selected Polychlorinated Organic Compounds (PCOCs). *Chemosphere* **2001**, *44*, 1421−1428.
(18) Dai, J.; Sun, C.; Han, S.; Wang, L. QSAR for Polychlorinated Organic Componds (PCOCs). I. Prediction of Partition Properties for PCOCs Using Quantum Chemical Parameters. *Bull. Environ. Contam. Toxicol.* **1999**, *62*, 530−538.
(19) Dietrich, D. W.; Dreyer, N. D.; Hansch, C. J. Confidence Interval Estimators for Parameters Associated with Quantitative Structure−Activity Relationships. *J. Med. Chem.* **1980**, *23*, 1201−1205.
(20) Meylan, W. M.; Howard, H. P.; Boethling, R. S.; Aronson, D.; Printup, H.; Gouchie, S. Improved Method for Estimating Bioconcentration/Bioaccumulation Factor from Octanol/Water Partition Coefficient. *Environ. Toxicol. Chem.* **1999**, *18*, 664−672.
(21) De Wolf, W.; Seinen, W.; Opperhuizen, A.; Hermens, J. L. M. The Influence of Biotransformation on the Relationship between Bioconcentration Factors and Octanol−Water Partition Coefficients. *Environ. Sci. Technol.* **1992**, *26*, 1197−1201.
(22) De Bruijn, J.; Seinen, W.; Hermens, J. Biotransformation of Organic Compounds by Rainbow Trout (*Oncorhynchus mykiss*) Liver in Relation to Bioconcentration. *Environ. Toxicol. Chem.* **1993**, *12*, 1041−1050.
(23) Niimi, A. J.; Lee, H. B.; Kissoon, G. P. Octanol/Water Partition Coefficients and Bioconcentration Factors of Chloronitrobenzenes in Rainbow Trout (*Salmo gairdneri*). *Environ. Toxicol. Chem.* **1989**, *8*, 817−823.
(24) Randic, M. The Connectivity Index 25 years after. *J. Mol. Graphics Modell.* **2001**, *20*, 19−35.
(25) Fatemi, M. H.; Jalali-Heravi, M.; Konuze, E. Prediction of Bioconcentration Factor using Genetic Algorithm and Artifical Neural Network. *Anal. Chim. Acta* **2003**, *486*, 101−108.
(26) Bintein, S.; Devillers, J.; Karcher, W. Nonlinear Dependence of Fish Bioconcentration on *n*-octanol/water Partition Coefficients. *SAR QSAR Environ. Res.* **1993**, *1*, 29−39.
(27) EPA (U.S. Environmental Protection Agency) and SRC (Environmental Science Center), 2000, EPIWINVersion3.11.
(28) Basak, S. C.; Gute, B. D.; Mills, D.; Hawkins, D. M. Quantitative Molecular Similarity Methods in the Property/Toxicity Estimation of Chemicals: A comparison of Arbitrary versus Tailored Similarity Spaces. *J. Mol. Struct. (THEOCHEM)* **2003**, *622*, 127−145.
(29) Shiu, W.-Y.; Ma, K.-C.; Varhanikova, D.; Mackay, D. Chlorophenols and alkylphenols: A Review and Correlation of Environmentally Relevant Properties and Fate in an Evaluative Environment *Chemosphere* **1994**, *29*, 1155−1224.

CI0342167