J Mol Model (2008) 14:83–92
DOI 10.1007/s00894-007-0255-y

ORIGINAL PAPER

Molecular electronegativity distance vector model for the Prediction of bioconcentration factors in fish

Shu-Shen Liu · Li-Tang Qin · Hai-Ling Liu · Da-Qiang Yin

Received: 20 August 2007 / Accepted: 8 November 2007 / Published online: 13 December 2007

Abstract Molecular electronegativity distance vector (MEDV) derived directly from the molecular topological structures was used to describe the structures of 122 nonionic organic compounds (NOCs) and a quantitative relationship between the MEDV descriptors and the bioconcentration factors (BCF) of NOCs in fish was developed using the variable selection and modeling based on prediction (VSMP). It was found that some main structural factors influencing the BCFs of NOCs are the substructures expressed by four atomic types of nos. 2, 3, 5, and 13, i.e., atom groups $-CH_2-$ or =CH−, −CH< or =C<, $-NH_2$, and −Cl or −Br where the former two groups exist in the molecular skeleton of NOC and the latter three groups are related closely to the substituting groups on a benzene ring. The best 5-variable model, with the correlation coefficient (r^2) of 0.9500 and the leave-one-out cross-validation correlation coefficient (q^2) of 0.9428, was built by multiple linear regressions, which shows a good estimation ability and stability. A predictive power for the external samples was tested by the model from the training set of 80 NOCs and the predictive correlation coefficient (u^2) for the 42 external samples in the test set was 0.9028.

S.-S. Liu (✉) · H.-L. Liu · D.-Q. Yin
Key Laboratory of Yangtze River Water Environment, Ministry of Education, College of Environmental Science and Engineering, Tongji University, 200092 Shanghai, People's Republic of China
e-mail: ssliuhl@263.net

L.-T. Qin
Department of Material and Chemical Engineering, Guilin University of Technology, 541004 Guilin, People's Republic of China

Keywords Bioconcentration factors (BCF) · Molecular electronegativity distance vector (MEDV) · Nonionic organic compounds (NOCs) · Variable selection and modeling based on prediction (VSMP)

Introduction

A considerable amount of halogenated organic contaminants such as polychlorinated biphenyls, polybrominated biphenyls, chlorinated aliphatic hydrocarbons, polychlorinated benzenes, polybrominated benzenes, polychlorinated anilines, polychlorinated nitrobenzenes, and phenols have been discharged into the environment. Most of these compounds were released from industrial activities, agricultural and residential sources. They are persistent pollutants of the environment thereby producing widespread contamination of water and soil and invariably present in the aquatic environment as highly complex mixtures of isomers and congeners, which complicates environment hazard evaluation and risk assessment. These chemicals show a tendency to accumulate in biota, soils, and sediments and move through food chains and accumulate at sizeable levels in the tissues of animals [1–3]. Human beings have suffered from these chemicals in their daily lives. Consequently, it is necessary to establish scientifically credible risk assessments for the persistent pollutants. In this way, the hazard potential of a chemical can be properly classified and labeled.

The bioconcentration potential of chemicals is normally expressed as the bioconcentration factors (BCF). The BCF constitutes an important parameter to establish the potential hazard of a chemical. The BCF for a particular chemical compound is defined as the ratio of the concentration of a chemical inside an organism (or in the fat, or in a certain tissue of the organism) to the concentration in the

Table 1 The values of five optimal MEDV descriptors and log(BCF) and logP observed and calculated for 122 NOCs

No.	Compounds	x_{15}	x_{17}	x_{25}	x_{36}	x_{91}	OBS[1]	CAL[2]	logP	logP[3]
1	hexachloroethane	0.0000	0.0000	0.0000	0.0000	18.7252	2.92	2.27	4.14	3.81
2	pentachloroethane	0.0000	0.0000	0.0000	−1.2597	12.0972	1.83	1.83	3.22	3.15
3	tetrachloromethane	0.0000	0.0000	0.0000	0.0000	9.9197	1.48	1.79	2.83	3.08
4	1,1,1-trichloroethane	0.0000	0.0000	0.0000	0.0000	4.8742	0.95	1.51	2.49	2.66
5	1,1,2,2-tetrachloroethane	0.0000	0.0000	0.0000	−1.4539	6.5459	0.90	1.51	2.39	2.67
6	trichloromethane	0.0000	0.0000	0.0000	−2.9800	4.6437	0.78	1.30	1.97	2.38
7	1,2-dichloroethane	0.0000	0.0000	5.9789	0.0000	0.6530	0.30	1.61	1.48	2.31
8	1,1,2,3,4,4-hexachloro-1,3-butadiene	0.0000	0.0000	0.0000	16.3998	8.5343	3.76	2.81	4.78	4.39
9	tetrachloroethylene	0.0000	0.0000	0.0000	11.7051	4.6375	1.74	2.29	3.40	3.66
10	trichloroethylene	1.9346	0.0000	7.2753	3.7137	2.0292	1.59	2.13	2.42	2.94
11	benzene	0.0000	0.0000	0.0000	0.0000	0.0000	0.64	1.25	2.19	2.25
12	toluene	2.7484	0.0000	0.0000	0.0000	0.0000	1.12	1.43	2.73	2.53
13	ethyl benzene	5.8566	0.0000	0.0000	0.0000	0.0000	1.19	1.64	3.15	2.84
14	*o*-xylene	5.4002	0.0000	0.0000	0.0000	0.0000	1.24	1.61	3.12	2.79
15	*m*-xylene	5.9519	0.0000	0.0000	0.0000	0.0000	1.27	1.64	3.20	2.85
16	*p*-xylene	5.7917	0.0000	0.0000	0.0000	0.0000	1.27	1.63	3.15	2.83
17	isopropylbenzene	4.4844	0.0000	0.0000	0.0000	0.0000	1.55	1.55	3.72	2.70
18	hexachlorobenzene	0.0000	0.0000	0.0000	30.9321	6.3959	4.16	3.67	5.31	5.48
19	2,4,5-trichlorotoluene	12.5470	0.0000	6.9313	8.3689	1.0530	3.87	3.07	4.56	4.33
20	1,2,3,4-tetrachlorobenzene	7.9846	0.0000	6.2094	14.2125	2.5491	3.72	3.21	4.64	4.50
21	1,2,4,5-tetrachlorobenzene	13.0768	0.0000	9.3174	11.4413	2.1171	3.61	3.50	4.82	4.74
22	pentachlorobenzene	7.7604	0.0000	5.2634	20.5753	4.0844	3.45	3.65	5.18	5.16
23	1,2,3,5-tetrachlorobenzene	11.4296	0.0000	8.5306	12.0720	2.2556	3.36	3.40	4.92	4.64
24	1,3,5-trichlorobenzene	11.5933	0.0000	10.0328	5.2621	0.8740	3.26	2.96	4.19	3.95
25	1,2,4-trichlorobenzene	10.2844	0.0000	8.6339	6.6581	1.0193	2.95	2.90	4.02	3.95
26	1,2,3-trichlorobenzene	7.4963	0.0000	6.0866	8.9871	1.4073	2.90	2.75	4.05	3.90
27	1,3-dichlorobenzene	7.2079	0.0000	7.0880	2.8694	0.2814	2.65	2.32	3.60	3.25
28	1,4-dichlorobenzene	7.0293	0.0000	7.4484	2.5689	0.1774	2.47	2.30	3.52	3.20
29	1,2-dichlorobenzene	6.3202	0.0000	5.4427	4.4698	0.5224	2.43	2.29	3.43	3.32
30	chlorobenzene	3.3055	0.0000	3.6916	1.1515	0.0000	1.85	1.74	2.84	2.68
31	1,2,4,5-tetrabromobenzene	19.5601	0.0000	6.7247	10.9083	0.9739	3.81	3.69	5.13	5.25
32	1,3,5-tribromobenzene	17.9730	0.0000	7.0567	5.1794	0.3903	3.70	3.19	4.51	4.54
33	1,2,4-tribromobenzene	15.4257	0.0000	6.0819	6.3409	0.4605	3.66	3.05	4.66	4.39
34	1,4-dibromobenzene	10.7688	0.0000	5.1014	2.5020	0.0772	2.83	2.41	3.79	3.56
35	1,3-dibromobenzene	11.0082	0.0000	4.8744	2.7936	0.1235	2.80	2.44	3.75	3.61
36	1,2-dibromobenzene	9.0875	0.0000	3.7647	4.1603	0.2343	2.70	2.35	3.64	3.55
37	bromobenzene	4.9619	0.0000	2.4775	1.1080	0.0000	1.70	1.79	2.99	2.85
38	2,2′,3,3′,4,4′,5,6-octachlorobiphenyl	14.7780	0.0000	6.6060	41.3569	7.6484	5.92	5.77	7.35	7.97
39	2,2′,3,4,4′,5-hexachlorobiphenyl	27.6919	0.0000	12.4562	23.2395	3.8228	5.88	5.52	6.82	7.37
40	2,2′,3,3′,4,5,5′,6-octachlorobiphenyl	17.4921	0.0000	8.9922	39.4812	7.3743	5.88	5.94	8.91	8.06
41	2,2′,3,4,4′,5′,6-heptachlorobiphenyl	27.4549	0.0000	14.3090	28.2381	4.9069	5.84	6.00	7.04	7.87
42	2,2′,3,3′,5,5′,6,6′-octachlorobiphenyl	16.4946	0.0000	11.2003	38.3153	6.9956	5.82	5.90	7.73	7.82
43	3,3′,4,4′,5-pentachlorobiphenyl	34.8448	0.0000	12.7916	15.9061	2.4637	5.81	5.45	6.98	7.34
44	2,2′,3,4,5,5′-hexachlorobiphenyl	27.1934	0.0000	12.5442	23.4288	3.8170	5.81	5.51	6.75	7.34
45	2,2′,3,3′,4,4′-hexachlorobiphenyl	25.6581	0.0000	10.7492	24.6196	3.9626	5.77	5.40	6.96	7.30
46	2,2′,3,3′,4,4′,5,5′,6-nonachlorobiphenyl	13.2644	0.0000	5.2518	48.8706	9.5337	5.71	6.21	9.14	8.63
47	2,2′,3′,4,5-pentachlorobiphenyl	29.4849	0.0000	13.2805	16.6475	2.3514	5.43	5.17	6.67	6.86
48	2,2′, 3,3′, 6,6′-hexachlorobiphenyl	18.7476	0.0000	12.6696	24.4568	3.6514	5.43	5.02	7.03	6.57
49	2,2′, 4,5,5′-pentachlorobiphenyl	31.8936	0.0000	15.6768	14.5074	1.9676	5.40	5.29	6.65	6.88
50	2,2′,3,4,5′-pentachlorobiphenyl	27.4011	0.0000	12.7855	17.0358	2.3871	5.38	5.03	6.23	6.68
51	2,2′,3,3′,4,4′,5,5′-octachlorobiphenyl	24.3400	0.0000	9.4127	38.2775	7.1936	5.08	6.33	8.68	8.62
52	2,2′,4,5-tetrachlorobiphenyl	28.2524	0.0000	12.0326	11.9788	1.4627	5.00	4.66	5.69	6.25
53	2,2′,4,4′,6,6′-hexachlorobiphenyl	26.3201	0.0000	16.9001	20.0684	3.1119	4.93	5.42	7.55	6.90
54	2,2′, 3,5′-tetrachlorobiphenyl	26.4792	0.0000	12.2212	11.8082	1.2860	4.84	4.53	5.73	6.04
55	2,2′,4,5′-tetrachlorobiphenyl	29.1498	0.0000	14.3460	9.6444	0.9625	4.84	4.66	5.87	6.10
56	2,2′,4,4′,5,5′-hexachlorobiphenyl	34.9173	0.0000	16.6645	19.4629	3.1069	4.83	5.94	7.75	7.71

Table 1 (continued)

No.	Compounds	x_{15}	x_{17}	x_{25}	x_{36}	x_{91}	OBS[1]	CAL[2]	logP	logP[3]
57	2,2′,5,5′-tetrachlorobiphenyl	28.7831	0.0000	14.5564	9.7141	0.9102	4.63	4.65	5.79	6.06
58	2,4,4′-trichlorobiphenyl	28.3410	0.0000	11.3837	5.8738	0.4358	4.63	4.16	5.58	5.64
59	2,3′,4′,5-tetrachlorobipenyl	31.3669	0.0000	13.4461	10.0450	1.1156	4.62	4.79	6.07	6.37
60	2,3-dichlorobiphenyl	22.2913	0.0000	5.5403	6.1835	0.5318	4.25	3.47	5.02	5.07
61	2,2′,3,3′-tetrachlorobiphenyl	24.1319	0.0000	9.8794	13.9077	1.6622	4.23	4.41	5.67	6.02
62	2,2′,4,4′-tetrachlorobiphenyl	29.5093	0.0000	14.1180	9.5895	1.0210	4.02	4.67	6.29	6.13
63	2,4,5-trichlorobiphenyl	27.5339	0.0000	9.0978	8.6090	1.0305	4.02	4.20	5.90	5.85
64	2,2′,5-trichlorobiphenyl	25.1005	0.0000	10.7986	7.3485	0.4959	4.01	4.02	5.55	5.45
65	2,5-dichlorobiphenyl	24.1892	0.0000	7.6713	4.2419	0.1807	4.00	3.56	5.16	5.06
66	3,3′,4,4′-tetrachlorobiphenyl	33.4021	0.0000	12.1274	10.5966	1.3714	3.90	4.90	6.63	6.64
67	2,2′,6,6′-tetrachlorobiphenyl	18.7224	0.0000	11.3706	13.5041	1.3708	3.85	4.09	5.94	5.42
68	3,5-dichlorobiphenyl	26.7339	0.0000	7.8360	3.6502	0.2839	3.78	3.71	5.41	5.28
69	2,4′,5-trichlorobiphenyl	28.0792	0.0000	11.6976	5.8369	0.3439	3.75	4.15	5.68	5.61
70	2,4′-dichlorobiphenyl	24.2354	0.0000	7.7069	3.8088	0.0909	3.55	3.53	5.10	5.02
71	4,4′-dichlorobiphenyl	26.3565	0.0000	8.1912	2.7403	0.0541	3.28	3.63	5.58	5.14
72	2,2′-dichlorobiphenyl	21.4449	0.0000	6.9593	5.1402	0.1824	3.26	3.40	4.90	4.87
73	4-chlorobiphenyl	22.4775	0.0000	4.0919	1.3388	0.0000	2.69	3.05	4.63	4.62
74	2,2′,5,5′-tetrabromobiphenyl	37.9237	0.0000	9.6296	8.1995	0.3651	4.80	4.85	7.31	6.80
75	4,4′-dibromobiphenyl	30.6906	0.0000	5.3633	2.4690	0.0221	4.19	3.74	5.72	5.55
76	2,4,6-tribromobiphenyl	30.7339	0.0000	6.4415	7.0691	0.3631	3.93	4.13	6.42	5.98
77	biphenyl	18.6562	0.0000	0.0000	0.0000	0.0000	2.64	2.48	4.09	4.12
78	pentachlorophenol	0.0000	0.0000	0.0000	22.8017	4.1188	2.99	3.00	5.12	4.98
79	2,4,6-trichlorophenol	9.4167	0.0000	6.8095	6.7126	0.8865	2.43	2.74	3.69	4.11
80	2-chlorophenol	3.6166	0.0000	2.5674	1.6053	0.0000	2.33	1.73	2.15	2.85
81	2,4-dichlorophenol	7.2846	0.0000	5.6843	3.3902	0.2844	2.00	2.28	5.53	3.44
82	3-chlorophenol	3.5341	0.0000	3.3684	1.1620	0.0000	1.30	1.74	2.50	2.76
83	phenol	0.0774	0.0000	0.0000	0.0000	0.0000	1.24	1.25	1.46	2.26
84	2,4-dimethylphenol	6.1671	0.0000	0.0000	0.0000	0.0000	2.18	1.66	2.30	2.87
85	4-t-buthylphenol	6.2170	0.0000	0.0000	0.0000	0.0000	2.07	1.66	3.31	2.88
86	*p-sec*-buthylphenol	5.7453	0.0000	0.0000	0.0000	0.0000	1.57	1.63	3.08	2.83
87	2-methylphenol	3.1812	0.0000	0.0000	0.0000	0.0000	1.03	1.46	1.95	2.57
88	4-bromophenol	5.1522	0.0000	2.3722	1.0873	0.0000	1.56	1.79	2.59	2.89
89	2,4,6-tribromophenol	14.0090	0.0000	4.7332	6.2022	0.3865	2.71	2.87	4.13	4.40
90	4-chloroaniline	4.8058	5.2021	3.6256	1.1914	0.0000	0.91	1.14	1.88	1.80
91	3-chloroaniline	4.9495	4.9509	3.4375	1.2690	0.0000	1.06	1.18	1.88	1.87
92	2-chloroaniline	4.6624	3.8115	2.6278	1.8508	0.0000	1.18	1.31	1.90	2.12
93	diphenylamine	9.3251	0.0000	0.0000	0.0000	0.0000	1.48	1.87	3.50	3.19
94	3,4-dichloroaniline	7.9422	4.8133	5.0293	4.6692	0.5283	1.48	1.74	2.78	2.54
95	2,4-dichloroaniline	8.4472	3.7590	5.7907	3.7213	0.2887	1.98	1.88	2.78	2.70
96	2,3,4-trichloroaniline	6.6100	2.9101	4.0081	10.2897	1.4370	2.31	2.28	3.68	3.35
97	2,4,5-trichloroaniline	11.1727	3.2720	6.7820	7.8138	1.0385	2.61	2.50	3.45	3.49
98	3,4,5-trichloroaniline	9.0534	4.3413	5.2294	9.3969	1.4242	2.70	2.26	3.32	3.23
99	2,4,6-trichloroaniline	10.2050	2.0386	6.9211	7.3596	0.9017	2.73	2.57	3.52	3.58
100	2,3,5,6-tetrachloroaniline	7.2714	0.6103	4.7495	14.4116	2.1782	3.03	2.99	4.10	4.30
101	2,3,4,5-tetrachloroaniline	6.4282	2.3292	3.4665	15.8783	2.5961	3.28	2.75	4.27	4.03
102	pentachloroaniline	9.3584	0.0000	6.5819	14.4717	4.4407	3.78	3.44	4.82	4.82
103	aniline	1.3171	5.1461	0.0000	0.0000	0.0000	0.41	0.64	0.90	1.35
104	2-methyl-4,6-dinitrophenol	−0.3591	0.0000	0.0000	0.0000	0.0000	0.16	1.22	2.12	2.22
105	3-nitrophenol	−0.8234	0.0000	0.0000	0.0000	0.0000	1.40	1.19	2.00	2.17
106	2-nitrophenol	−0.5373	0.0000	0.0000	0.0000	0.0000	1.60	1.21	1.79	2.20
107	2,4,5-trichloronitrobenzene	6.0330	0.0000	5.4111	4.4296	0.8899	1.84	2.29	3.48	3.70
108	3-chloronitrobenzene	2.0823	0.0000	2.7215	0.5756	0.0000	1.89	1.57	2.46	2.62
109	2,3,4,5-tetrachloronitrobenzene	3.8569	0.0000	2.6022	10.3247	2.2296	1.89	2.46	4.57	4.25
110	4-chloronitrobenzene	2.4693	0.0000	2.9513	0.7567	0.0000	2.00	1.62	2.39	2.64
111	2,5-dichloronitrobenzene	3.4360	0.0000	4.7136	1.2544	0.1485	2.05	1.82	3.09	3.00
112	2,4-dichloronitrobenzene	4.1529	0.0000	4.6880	1.6142	0.2409	2.07	1.90	3.07	3.08

Table 1 (continued)

No.	Compounds	x_{15}	x_{17}	x_{25}	x_{36}	x_{91}	OBS[1]	CAL[2]	logP	logP[3]
113	3,4-dichloronitrobenzene	4.5861	0.0000	3.9479	3.1330	0.4653	2.07	2.00	3.12	3.22
114	2-chloronitrobenzene	1.6182	0.0000	2.0146	0.4259	0.0000	2.10	1.49	2.24	2.62
115	2,3-dichloronitrobenzene	3.3834	0.0000	2.9945	2.7091	0.4396	2.16	1.84	3.05	3.16
116	2,3,4-trichloronitrobenzene	4.2154	0.0000	3.1860	6.2791	1.2213	2.20	2.19	3.68	3.71
117	3,5-dichloronitrobenzene	4.3488	0.0000	5.1934	1.5763	0.2448	2.23	1.94	3.09	3.07
118	pentachloronitrobenzene	0.0000	0.0000	0.0000	15.2769	3.5376	2.40	2.46	4.77	4.62
119	2,3,5,6-tetrachloronitrobenzene	4.4146	0.0000	4.1195	7.7348	1.7903	3.20	2.38	3.89	4.14
120	4-nitroaniline	0.6503	4.2026	0.0000	0.0000	0.0000	0.64	0.73	1.39	1.48
121	2-nitroaniline	0.2914	2.9887	0.0000	0.0000	0.0000	0.91	0.87	1.85	1.68
122	3-nitroaniline	0.2877	3.9229	0.0000	0.0000	0.0000	0.92	0.74	1.37	1.50

[1] refers to the log(BCF) observed; [2] to the log(BCF) calculated by the model M2; [3] to the logP calculated by the model M4.

surrounding environment [4–6]. The concentration of the chemical in the organism and the aqueous environment are measured after long-term exposure until steady state is reached. BCF are usually used to estimate the propensity of a chemical to bioaccumulate in various species of fish and other aquatic organisms [6, 7] and it is shown to be highly correlated to the octanol-water partition coefficient (P) for a wide variety of chemicals [8, 9].

Because the experimental determination of BCF is time-consuming, difficult and expensive [10, 11], it is important to develop the quantitative structure-activity relationship (QSAR) models to provide reliable predictions for a large number of chemical compounds. It is also necessary to determine the physico-chemical properties of these organic chemicals and their environmental partitioning. Numerous attempts have been made for modeling the accumulation of organic chemicals and estimation of BCF values, such as characteristic root index and semi-empirical molecular descriptors [12], Padmakar-Ivan (PI) index [13], genetic algorithm and artificial neural network [14], molecular connectivity indices and polarity correction factors [15], and fragment constant method [16, 17]. The other approach based on the experimental relationship between BCF and the physico-chemical parameters such as the octanol/water partitioning coefficient (P), water solubility or the soil adsorption coefficient was also used to estimate a chemical's BCF [8, 9, 18].

In the present paper, a QSAR model for the prediction of BCF of 122 nonionic organic compounds (NOCs) was developed by using the molecular electronegativity distance vector (MEDV) [19] obtained directly from two dimensional topological molecular structures and the electrotopological state index [20, 21]. The MEDV descriptors had been applied previously to the quantitative structure-activity/property relationship (QSAR/QSPR) studies on many complicated molecular systems, such as cyclooxygenase-2 (COX-2) inhibitors [22, 23], polychlorinated naphthalenes [24], polybrominated diphenyl ethers (PBDEs) [25, 26], and polychlorinated biphenyls [27, 28]. The VSMP [29, 30] was employed to select an optimal subset from the original descriptor set and then the subset was used to create a relationship model between the MEDV and BCF of NOCs.

Materials and methods

Data set

The values of the log(BCF) and logP for 122 nonionic organic compounds (NOCs) were directly taken from the literature [12]. In our paper, 122 NOCs were arranged according to the classes of compounds such as 10 chlorinated aliphatic hydrocarbons (from no. 1 to 10), seven substituted benzenes (11 to 17), 13 polychlorinated benzenes (18 to 30), seven polybrominated benzenes (31 to 37), 36 polychlorinated biphenyls (38 to 73), three polybrominated biphenyls (74 to 76), biphenyl (77), five polychlorinated phenols (78 to 82), five substituted phenols (83 to 87), two polybrominated phenols (88 to 89), 14 polychlorinated anilines (90 to 103), and 19 substituted nitrobenzene (104 to 122). The names of the NOCs as well as their log(BCF) and logP values observed are listed in Table 1. The distribution of log(BCF) and logP for 122 NOCs are shown in Fig. 1 where 40 NOCs display the logBCF values between 0.0 and 2.0 (low concentration), 52 ones between 2.0 and 4.0 (moderate concentration), and 30 between 4.0 and 6.0 (good concentration). The logP has a similar distribution. The NOC with the minimum log(BCF) value of 0.16 is 2-methyl-4,6-dinitrophenol (no.104) and one with the maximum log(BCF) value of 5.92 is 2,2′,3,3′,4,4′,5,6-octachlorobiphenyl (no.38).

MEDV method

Two concepts, atomic types and atomic attributes, were introduced into the MEDV procedure to describe each non-

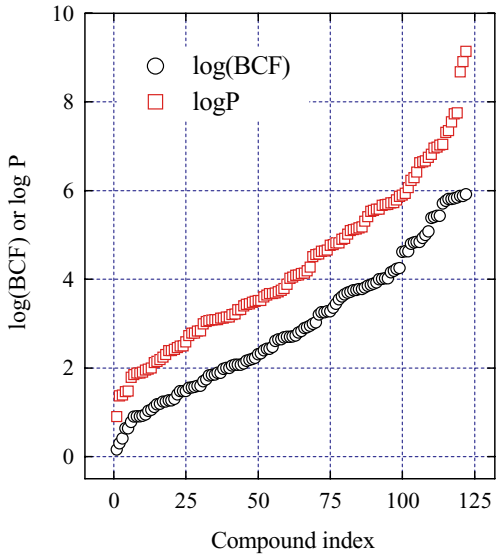

Fig. 1 Distribution of log(BCF) and logP of 122 NOCs

hydrogen atom of the examined molecule. There are 13 atomic types and 43 atomic attributes for most organic compound molecules [19, 31]. The atomic attributes, substructures (groups), and corresponding atomic types of 122 NOCs were listed in Table 2. According to the literatures [19, 31], the original MEDV-13 descriptor, x_z (z=1, 2, 3, ..., 91), is calculated by the following.

Firstly, the intrinsic state (I) of an atom is calculated:

$$I = \sqrt{\frac{\nu}{4}\frac{(2/n)^2\delta^\nu + 1}{\delta}} \tag{1}$$

where the symbol $\boldsymbol{v}$ is the number of valence electrons; $\boldsymbol{n}$ is the principal quantum number for the valence shell of that atom; and δ^ν and δ are the molecular connectivity delta values which are given as follows:

$$(\delta = \sigma - h,\ \delta^\nu = \sigma + \pi - h) \tag{2}$$

where σ and δ are respectively the number of electrons in σ and π orbitals and $\boldsymbol{h}$ is the number of hydrogen atoms bonded to the atom.

Secondly, the relative electronegativity (q) of a non-hydrogen atom is calculated using the atomic type, atomic attributes (Table 2), and intrinsic state (I) of the atom:

$$q_i = I_i + \sum_{j\neq i}^{all\ j}\frac{(I_i - I_j)}{d_{ij}^2} \tag{3}$$

where d_{ij} is the shortest graph distance between two atoms, atom i and j.

Then, the MEDV-13 descriptor, $\boldsymbol{x}_z$, is calculated as follows:

$$x_z = m_{kl} = \sum_{i\in k, j\in l}\frac{q_i q_j}{d_{ij}^2}(k,\ l = 1, 2, \ \ldots\ , 13; l \geq k;\ z = 1, 2, \ \ldots\ , 91) \tag{4}$$

where k and l are the atomic types of the ith and jth non-hydrogen atoms (Table 2) in which i and j are the serial number of the non-hydrogen atoms in a molecule, respectively. $\boldsymbol{z}$ is the serial number of the MEDV descriptors. In general, there are 91 MEDV descriptors for a given molecule. However, for the 122 NOCs under study, there are only 8 atomic types (atomic types 1,2,3,4,5,7,9, and 13, Table 2) which result in 34 nonzero MEDV descriptors according to Eq. (4). The other 57 out of 91 MEDV descriptors with zero values were deleted from the model to be developed because of no contribution to the model. Furthermore, among the 34 nonzero MEDV descriptors, 18 ($\boldsymbol{x}_1$, $\boldsymbol{x}_2$, $\boldsymbol{x}_3$, $\boldsymbol{x}_4$, $\boldsymbol{x}_7$, $\boldsymbol{x}_9$, $\boldsymbol{x}_{13}$, $\boldsymbol{x}_{16}$, $\boldsymbol{x}_{18}$, $\boldsymbol{x}_{27}$, $\boldsymbol{x}_{29}$, $\boldsymbol{x}_{37}$, $\boldsymbol{x}_{42}$, $\boldsymbol{x}_{46}$, $\boldsymbol{x}_{49}$, $\boldsymbol{x}_{51}$, $\boldsymbol{x}_{55}$, and $\boldsymbol{x}_{64}$) have too few number of nonzero samples (≤12) to have significant meaning to the model and should be also deleted before the model development. Thus, there are only 16 MEDV descriptors ($\boldsymbol{x}_{14}$, $\boldsymbol{x}_{15}$, $\boldsymbol{x}_{17}$, $\boldsymbol{x}_{19}$, $\boldsymbol{x}_{21}$, $\boldsymbol{x}_{25}$, $\boldsymbol{x}_{26}$, $\boldsymbol{x}_{28}$, $\boldsymbol{x}_{30}$, $\boldsymbol{x}_{32}$, $\boldsymbol{x}_{36}$, $\boldsymbol{x}_{66}$, $\boldsymbol{x}_{70}$, $\boldsymbol{x}_{77}$, $\boldsymbol{x}_{81}$, and $\boldsymbol{x}_{91}$) entering the next QSAR analysis.

Variable selection and modeling

Not all MEDV descriptors affect the BCF value. Therefore, it is necessary to select the optimal variables from the

Table 2 The atomic attributes, substructures (group), and atomic types existed in 122 NOCs

No.	atomic attribute	substructure (group)*	atomic type	No.	atomic attribute	substructure (group)*	atomic type
1	1	−CH_3	1	9	16	cCcc	3
2	2	−CH_2−	2	10	17	−OH	9
3	3	−CH<	3	11	19	=O	9
4	4	>C<	4	12	21	−NH_2	5
5	6	=CH−	2	13	23	>N−	7
6	7	=C<	3	14	30	≥N=	7
7	14	cCHc	2	15	37	−Cl	13
8	15	−cCc	3	16	38	−Br	13

*: − refers to a single bond, = to a double bond, > or < to two single bonds, c to a conjugated bond, and ≥ to a single bond plus a double bond.

Table 3 Some statistics in the optimal QSAR models having various *m* values base on the training set of 80 compounds

m	r^2	*RMSE*	q^2	*RMSV*	Optimal MEDV descriptors						
1	0.6555	0.89	0.6397	0.91	x_{25}						
2	0.8970	0.49	0.8862	0.51	x_{15}	x_{36}					
3	0.9122	0.45	0.9014	0.48	x_{15}	x_{17}	x_{36}				
4	0.9225	0.42	0.9121	0.45	x_{15}	x_{17}	x_{25}	x_{36}			
5	0.9357	0.39	0.9211	0.43	x_{15}	x_{17}	x_{25}	x_{36}	x_{91}		
6	0.9357	0.39	0.9209	0.43	x_{15}	x_{17}	x_{25}	x_{28}	x_{36}	x_{91}	
7	0.9362	0.38	0.9190	0.43	x_{15}	x_{17}	x_{21}	x_{25}	x_{28}	x_{36}	x_{91}

original set of 16 nonzero MEDV descriptors. The VSMP program [29, 30] developed in our laboratory was used to select a set of optimal MEDV descriptors. The VSMP is a modified all-subset regression technique based on prediction rather than estimation. Some attempts have been made for selection of descriptors such as stepwise regression, genetic algorithm and artificial neural network. Most of these methods aimed at the estimated statistics of a model and usually only account for the estimation abilities for internal samples, while the VSMP technique examined not only the estimated statistics but also the leave-one-out (LOO) cross-validated ones. The main steps of VSMP can be seen in literature [29].

Results and discussion

MEDV model for the prediction of the log(BCF)

To validate and develop a credible QSAR model, it is not enough to only model the whole data set. So, it is necessary to split the whole dataset into a training set and a test set. The log(BCF) values of 122 NOCs were sorted ascending and then 80 NOCs were equidistantly picked up as a training set and the remaining 40 NOCs made up a test set. The log(BCF) of the NOCs in the training set were used as a dependent variable and the nonzero MEDV descriptors as predictive variables to construct a QSAR model. The optimal descriptors were selected by VSMP whose results are shown in Table 3. In Table 3, the *r* and *q* refer to the correlation coefficient in the modeling and LOO validation step, while the *RMSE* and *RMSV* to the root mean square error calibrated and validated, respectively.

From Table 3, the best MEDV model for the training set is a 5-variable model (m=5) and the five variables are respectively the MEDV descriptors of nos. $\boldsymbol{x}_{15}$, $\boldsymbol{x}_{17}$, $\boldsymbol{x}_{25}$, $\boldsymbol{x}_{36}$, and $\boldsymbol{x}_{91}$. The descriptors reflected the interactions between the pairs of atomic types 2 and 3 ($\boldsymbol{x}_{15}$), 2 and 5 ($\boldsymbol{x}_{17}$), 2 and 13 ($\boldsymbol{x}_{25}$), 3 and 13 ($\boldsymbol{x}_{36}$), and 13 and 13 ($\boldsymbol{x}_{91}$). Fig. 2 illustrated the relations between four atomic types (i=2, 3, 5, and 13) and five MEDV descriptors ($\boldsymbol{x}_j$, j=15, 17, 25, 36, and 91) existing in the molecule of 3,4-dichloroaniline (no. 94 in Table 1). Using the multiple linear regression, a 5-variable equation (Eq. 5) was developed and the model (M1) was then used to predict the log (BCF) values of 42 NOCs in the test set. The M1 gave a good estimated ability (r^2=0.9357, *RMSE*=0.39), a high stability (q^2=0.9211, *RMSV*=0.43), and a credible predictive potential (u^2=0.9028, *RMSP*=0.51).

The analytic equation corresponding to the model M1 was expressed as follows.

$$\begin{aligned}\log(BCF) &= (1.2006 \pm 0.0848) + (0.0621 \pm 0.0075) * x_{15} \\ &\quad -(0.1374 \pm 0.0340) * x_{17} + (0.0736 \pm 0.0189) * x_{25} \\ &\quad +(0.0679 \pm 0.0062) * x_{36} + (0.0651 \pm 0.0167) * x_{91} \\ n &= 80,\ m = 5,\ r^2 = 0.9357,\ RMSE = 0.39,\ F = 215.2 \\ &\qquad (\textit{modeling the training set}) \\ q^2 &= 0.9211,\ RMSV = 0.43 \quad (\textit{LOO validating the training set}) \\ t &= 42,\ m = 5,\ u^2 = 0.9028,\ RMSP = 0.51 \\ &\qquad (\textit{predicting the external test set}) \end{aligned} \tag{5}$$

Where *t*, *u*, and *RMSP* refer to the number of the test set samples, the predictive correlation coefficient, and the predictive root mean square error, respectively. *F* refers to the Fisher's statistic.

The above model M1 tests the applicability of the training set to predict the log(BCF) of NOCs. To improve the structural diversity of samples in modeling to make the QSAR model developed more widely available, the whole set of 122 NOCs was also modeled and validated by the

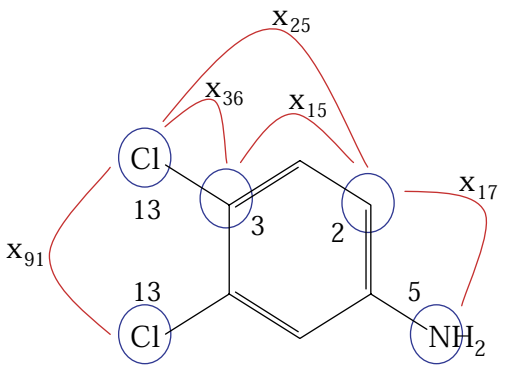

Fig. 2 Relation between some atomic types and MEDV descriptors of 3,4-dichloroaniline

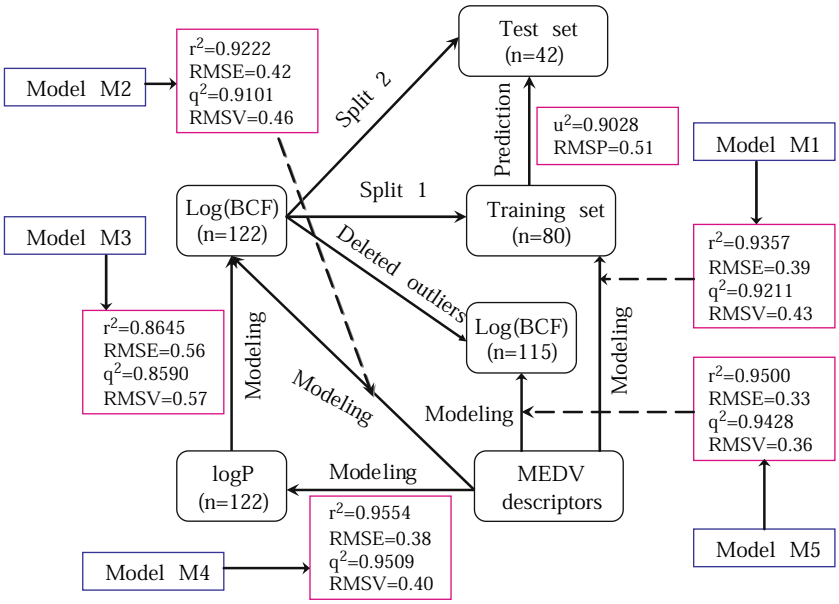

Fig. 3 Some statistical models between the log(BCF) and MEDV of 122 NOCs

same way as the training set. The optimal variable selection results from the VSMP program showed that the best QSAR model between the log(BCF) and MEDV descriptors of 122 NOCs was still a 5-variable model (M2) and the optimal descriptors were still nos. of $\boldsymbol{x}_{15}$, $\boldsymbol{x}_{17}$, $\boldsymbol{x}_{25}$, $\boldsymbol{x}_{36}$, and $\boldsymbol{x}_{91}$. The M2 can be expressed as follows.

$$\begin{aligned}\log(BCF) &= (1.2483 \pm 0.0732) + (0.0663 \pm 0.0069) * x_{15} \\ &\quad -(0.1343 \pm 0.0325) * x_{17} + (0.0542 \pm 0.0168) * x_{25} \\ &\quad +(0.0670 \pm 0.0056) * x_{36} + (0.0547 \pm 0.0173) * x_{91} \\ n = 122,\ m = 5,\ r^2 &= 0.9222,\ RMSE = 0.42,\ F = 275.0 \quad (modeling) \\ q^2 = 0.9101,\ RMSV &= 0.46 \quad (LOO\ validation)\end{aligned} \tag{6}$$

From Eq. (6), there are no significant differences between the qualities of the model M2 and M1, which denotes the applicability of the model M2, liking M1, in the estimation and prediction of the log(BCF) values for the external or unknown NOC compounds.

It is well known that there is a good linear relationship between the log(BCF) and logP. For 122 NOCs under study in this paper, the relation equation (M3) was as follows.

$$\begin{aligned}\log(BCF) &= -(0.4074 \pm 0.1311) + (0.7814 \pm 0.0282) * \log P \\ n = 122,\ m &= 1,\ r^2 = 0.8645,\ RMSE = 0.56,\ F = 765.9 \quad (modeling) \\ q^2 = 0.8590,\ RMSV &= 0.57 \quad (LOO\ validation)\end{aligned} \tag{7}$$

From Eq. (7), there is a good linear relationship (r^2= 0.8645) between the log(BCF) and logP, which implies that the logP determined easily through experiment could be used to estimate the log(BCF) obtained difficultly.

Comparing the results of Eq. (6) with Eq. (7), the statistical quality of the model M3 is obviously worse than

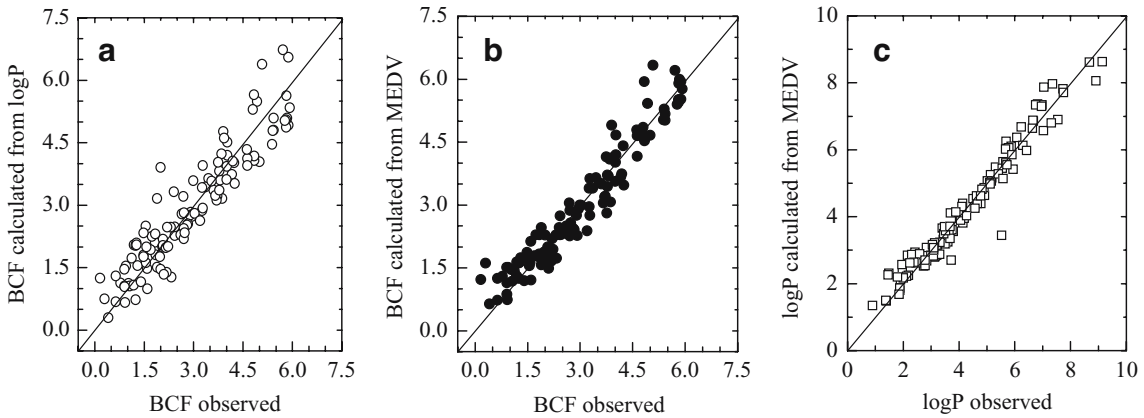

Fig. 4 The relationship graph of log(BCF) and logP (**a**) log(BCF) calculated from logP vs. observed; (**b**) log(BCF) calculated from MEDV vs. observed; (**c**) logP calculated from MEDV vs. observed

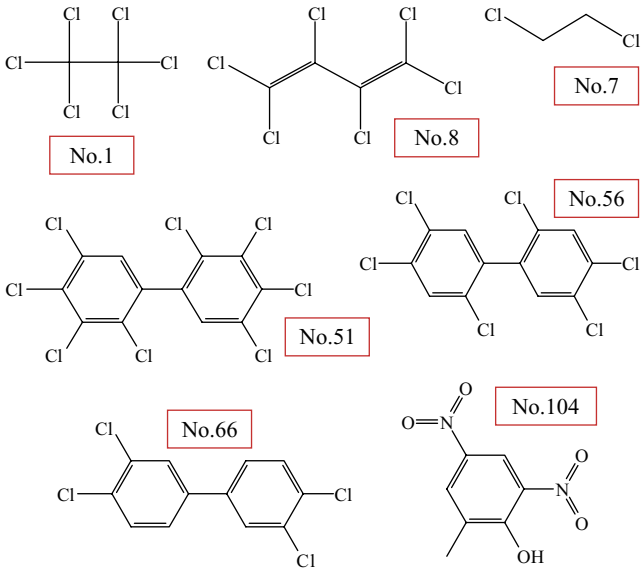

Fig. 5 The skeleton structures of seven NOC outliers having a larger residual than the double of RMSE

model M2. Then, whether is the logP integrated with the MEDV descriptors to more accurately predict the log(BCF) value? VSMP analysis on the integrated data set simultaneously including logP and MEDV descriptors gave a negative answer. However, the logP value can be accurately predicted using the optimal MEDV descriptors consisted of $\boldsymbol{x}_{15}$, $\boldsymbol{x}_{17}$, $\boldsymbol{x}_{36}$, $\boldsymbol{x}_{81}$, and $\boldsymbol{x}_{91}$. The predictive model (M4) with a high estimated quality and stability can be described as follows.

$$\begin{aligned}\log P &= (2.2516 \pm 0.0720) + (0.1003 \pm 0.0041) * x_{15} \\ &\quad -(0.1999 \pm 0.0300) * x_{17} + (0.0871 \pm 0.0049) * x_{36} \\ &\quad +(0.1121 \pm 0.0341) * x_{81} + (0.0831 \pm 0.0158) * x_{91} \\ n &= 122, m = 5, r^2 = 0.9554,\ RMSE = 0.38, F = 496.4 \\ &\qquad\qquad (modeling) \\ q^2 &= 0.9509,\ RMSV = 0.40 \quad (LOO\ validation)\end{aligned} \tag{8}$$

For the convenience of usage and analysis, the above model statistic results were summarized in Fig. 3. The relationship graphs between the log(BCF) or logP calculated and observed were shown in Fig. 4.

Error analysis

Some compounds had exhibited highly aberrant behavior in the model M2. The LOO validated results showed that there are 7 NOCs (their molecular structures are shown in Fig. 5) whose absolute residuals of the LOO cross-validation are higher than the double of *RMSE*. The LOO residuals are −1.26 (no. 1), 1.40 (no. 7), 1.02 (no. 8), 1.38 (no. 51), 1.18 (no. 56), 1.05 (no. 66), and 1.10 (no. 104), respectively.

The former three NOCs (nos. 1, 7, and 8) belong to one of 10 chlorinated aliphatic hydrocarbons (CAHs), each of them has a significantly different structure from the others located in CAHs and so has a higher residual. For instance, 1,1,2,3,4,4-hexachloro-1,3-butadiene (no. 8) is a unique CAH having a conjugated structure among 10 CAHs, 1,2-dichloroethane (no. 7) a unique CAH having no branch structure, and hexachloroethane (no. 1) a unique CAH whose carbons all have "$>$C$<$" structure. The other three NOCs (nos. 51, 56, and 66) belong to one of 36 polychlorinated biphenyls (PCBs) and are the structurally symmetrical PCB congeners. The last NOC (no. 104) is a unique compound containing two $-NO_2$ group among 19 substituted nitrobenzenes (SNBs).

If the above 7 outliers are deleted from the whole set of 122 NOCs, a new model (Eq. 9), called model M5, with r^2=0.9500, *RMSE*=0.33, q^2=0.9428, and *RMSV*=0.36, can be developed by VSMP program. Compared with model M2, the statistics of the model M5 make a significant improvement. Therefore, the model M5 is considered as the final predictive model.

$$\begin{aligned}\log(BCF) &= (1.2754 \pm 0.0607) + (0.0639 \pm 0.0057) * x_{15} \\ &\quad -(0.1475 \pm 0.0259) * x_{17} + (0.0626 \pm 0.0138) * x_{25} \\ &\quad +(0.0755 \pm 0.0052) * x_{36} + (0.0098 \pm 0.0195) * x_{91} \\ n &= 115, m = 5, r^2 = 0.9500,\ RMSE = 0.33,\ F = 413.8 \\ &\qquad\qquad (modeling) \\ q^2 &= 0.9428,\ RMSV = 0.36 \quad (LOO\ validation)\end{aligned} \tag{9}$$

Comparing Eq. (9) with Eq. (6) and Eq. (5), there are no significant differences for all regression coefficients but for the 5th one of $\boldsymbol{x}_{91}$, which shows a good stability of the models M1 and M2. The difference of $\boldsymbol{x}_{91}$ is because five NOCs out of seven outlier NOCs have more than 4 chloride atoms so as to make the regression coefficient of $\boldsymbol{x}_{91}$ more

Table 4 Comparison of some bioconcentration factor prediction models

No.	Method	*n*	*m*	r^2	*SE*	*F*	investigators
1	MEDV	115	5	0.950	0.330	413.8	This paper
2	Characteristic root index (CRI)	122	2	0.851	0.599	332.5	Sacan et al., 2004
3	The heuristic method (HM)	122	3	0.929	0.404	1575.6	Liu H.X. et al., 2006
4	Support vector machine (SVM)	122	3	0.953	0.331	2417.6	Liu H.X. et al., 2006

varied. According to the MEDV theory [19, 23], the descriptor x_{91} reflect the interaction between atomic type 13 and 13 which corresponds to the substituent, −Cl or −Br (Table 2).

MEDV related to the structures of NOCs

In order to go deep into the relationship between the structure of compounds and BCF, the relationship between logBCF and MEDV descriptors had been developed by VSMP. From the model M1, M2, and M5, the 5 optimal descriptors of nos. x_{15}, x_{17}, x_{25}, x_{36}, and x_{91} control the BCF values of organic chemicals. The atomic types were defined by the information about some substructures being $-CH_2-$ (type 2), −CH< or =C< (type 3), $-NH_2$ (type 5), and −Cl (type 13), respectively. So, the main structural factors determining the BCF values of NOCs under study are the benzene skeleton (type 2 and 3) and two substituents of −Cl or −Br (type 13) and $-NH_2$ (type 5).

Model comparisons

It is of interest to compare the results of the current study (the MEDV-based model) with those of recently published studies in which BCF models were developed (Table 4). The MEDV model is superior to the CRI-E_{HOMO} model in terms of the values of statistics. Moreover, the computation of the Characteristic root index (CRI) and the energy of the highest occupied molecular orbital (*E*HOMO) is far more time-consuming and complex than MEDV. Of the results reported in Table 4, the statistics of SVM model [32] are comparable with the MEDV model (M5). However, the development of SVM model is more time-consuming and complex than the multiple linear regression used in our MEDV model. On the other hand, only by using two softwares such as MOPAC and CODESSA can the descriptors used in SVM model be obtained.

Conclusions

The information presented in this study shows that a fairly good relationship existed between the MEDV descriptors and the log(BCF) of 122 nonionic compounds. It was shown that by using a small set of MEDV 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 $-NH_2$. Our 5-variable model is stable and has a high prediction power. This QSAR 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 QSAR are attractive because they can be calculated easily and rapidly. The optimal MEDV descriptors implied that the main structural factors affecting the BCF values were the substructures of $-CH_2-$, −CH< or =C<, $-NH_2$, and −Cl or −Br, respectively.

Acknowledgments We are especially grateful to the Foundation for the Author of National Excellent Doctoral Dissertation of P. R. China (No. 200355) and Shanghai Basic Research Program (No. 06JC14067) and Guangxi Thousands of Talents Program (No. 2003208) for their financial supports.

References

1. Verweij F, Booij K, Satumalay K, van der Molen N, van der Oost R (2004) Chemosphere 54:1675–1689
2. Feijtel T, Kloepper-Sams P, den Haan K, van Egmond R, Comber M, Heusel R, Wierich P, Berge WT, Gard A, de Wolf W, Niesson H (1997) Chemosphere 34:2337–2350
3. Ivanciuc T, Ivanciuc O, Klein DJ (2006) Mol Divers 10:133–145
4. Voutsas E, Magoulas K, Tassios D (2002) Chemosphere 48:645–651
5. Fatemi MH, Jalali-Heravi M, Konuze E (2003) Anal Chim Acta 486:101–108
6. Olive BG, Nilm AJ (1983) Environ Sci Technol 17:287–291
7. Khadikar PV, Singh S, Mandloi D, Joshi S, Bajaj AV (2003) Bioorg Med Chem 11:5045–5050
8. Devillers J, Bintein S, Domine D (1996) Chemosphere 33:1047–1065
9. Bahadur NP, Shiu WY, Boocock DGB, Mackay D (1997) J Chem Eng Data 42:685–688
10. Bermudez-Saldana JM, Escuder-Gilabert L, Medina-Hernandez MJ, Villanueva-Camanas RM, Sagrado S (2005) J Chromatogr A 1063:153–160
11. Bermudez-Saldana JM, Escuder-Gilabert L, Medina-Hernandez MJ, Villanueva-Camanas RM, Sagrado S (2005) J Chromatogr A 1063:153–160
12. Sacan MT, Erdem SS, Ozpinar GA, Balcioglu IA (2004) J Chem Inf Comp Sci 44:985–992
13. Khadikar PV, Singh S, Mandloi D, Joshi S, Bajaj AV (2003) Bioorg Med Chem 11:5045–5050
14. Fatemi MH, Jalali-Heravi M, Konuze E (2003) Anal Chim Acta 486:101–108
15. Lu XX, Tao S, Hu HY, Dawson RW (2000) Chemosphere 41:1675–1688
16. Tao S, Hu HY, Lu XX, Dawson RW, Xu FL (2000) Chemosphere 41:1563–1568
17. Lin Z, Yu H, Gao S, Cheng J, Wang L (2001) Arch Environ Contam Toxicol 41:255–260
18. Roy K, Sanyal I, Roy PP (2006) SAR QSAR Environ Res 17:563–582
19. Liu SS, Yin CS, Cai SX, Li ZL (2001) J Chem Inf Comp Sci 41:321–329
20. Kier LB, Hall LH (1990) Pharm Res 7:801–807
21. Hall LH, Mohney B, Kier LB (1991) J Chem Inf Comp Sci 31:76–82
22. Liu SS, Yin CS, Shi YY, Cai SX, Li ZL (2001) Chin J Chem 19:751–756
23. Liu SS, Cui SH, Yin DQ, Shi YY, Wang LS (2003) Chin J Chem 21:1510–1516
24. Liu SS, Yin CS, Wang LS (2002) Chin Chem Lett 13:791–794
25. Liu SS, Liu Y, Yin DQ, Wang LS (2005) Chin Chem Lett 16:1559–1662

26. Zhang YH, Liu SS, Liu HY (2007) Chromatographia 65:319–324
27. Liu SS, Liu Y, Yin DQ, Wang XD, Wang LS (2006) J Sep Sci 29:296–301
28. Liu SS, Cui SH, Wang LS (2004) Chin Chem Lett 15:467–470
29. Liu SS, Liu HL, Yin CS, Wang LS (2003) J Chem Inf Comp Sci 43:964–969
30. Liu SS, Yin DQ, Cui SH, Wang LS (2005) Chin J Chem 23:622–626
31. Liu SS, Yin CS, Wang LS (2002) J Chem Inf Comp Sci 42:749–756
32. Liu HX, Yao XJ, Zhang RS, Liu MC, Hu ZD, Fan BT (2006) Chemosphere 63:722–733