Pergamon

SCIENCE DIRECT®

Bioorganic & Medicinal Chemistry 11 (2003) 4039–4047

BIOORGANIC &
MEDICINAL
CHEMISTRY

Topological Approach to Quantifying Molecular Lipophilicity of Heterogeneous Set of Organic Compounds

Vijay K. Agrawal,[a] Shahnaz Bano[a] and Padmakar V. Khadikar[b,*]

[a]*QSAR and Computer Chemical Laboratories, A.P.S. University, Rewa-486 003, India*
[b]*Research Division, Laxmi Fumigation and Pest Control Pvt. Ltd., 3 Khatipura, Indore-452 007, India*

Received 17 February 2003; accepted 26 May 2003

Abstract—The lipophilicity of the large set of organic compounds is investigated using distance-based topological indices. The results have shown that molecular lipophilicity can be modeled in multi-parametric model in that W, $^1\chi$, B, J and logRB along with indicator parameters are involved. The results are discussed critically.

Introduction

The use of lipophilicity (logP) as a correlating parameter in biological studies is now well established.[1,2] It is efficiently used as one of the molecular descriptors in structure–activity relationship (SAR) studies related to medicinal chemistry, toxicology, pharmaceutical sciences, biological chemistry and environmental research.[3] Furthermore, the wide-spread application of lipophilicity (logP) to bio-physical processes involving xenobiotic explains the urgent need for both valid and quick procedures to quantify molecular lipophilicity.[4,5]

It is worthy of mention that disadvantages and short comings in the experimental determination of logP and consequently hydrophobic parameter π of Hansch analysis provoked an intensive search for alternative lipophilicity descriptors. Our earlier study has shown that PI index, that is a distance-based topological index is a promising lipophilicity descriptor.[4] The lipophilicity and toxicity of nitrobenzene derivatives and polychlorinated biphenyl xenobiotics is well accounted for using this recently introduced topological index.[6] In addition, the hydrophobic fragmental constant approach of calculating logP is well known.[7]

Our success in using PI index for modeling lipophilicity (logP) prompted us to further investigate other distance-based topological indices or their combinations for modeling lipophilic behavior of organic compounds. This is, therefore, the primary objective of the present study. In fulfilling our objective, we have, therefore, used the following topological indices and their combinations for modeling lipophilicity: Wiener index[8] (W), Szeged index[9,10] (Sz), hyper-Wiener index[11] (HW), Balaban index[12] (J), first-order connectivity index[13] ($^1\chi$), branching index[13] (B) and logRB.[14] The results as discussed below indicate that lipophilicity of a large set of organic compounds (116), as presented in Table 1, can be modeled through multi-parametric regression in that topological indices along with indicator parameters are also involved. The results are discussed below.

At this stage, it is interesting to record that Mannhold and coworkers,[15] while updating the hydrophobic fragmental constant approach, have used different sets of organic compounds consisting of environmentally important chemicals, aliphatic alcohols, hydrocarbons, mono-halogenated alkanes, mono-halogenated *n*-alkanes, halogenated aliphatic hydrocarbons, alkyl benzoic acids and mono-substituted benzoic acids. In these individual cases they obtained excellent correlations between fragmental constants and lipophilicity (logP). However, no attempt is made to use fragmental constant approach to model lipophilicity (logP) of the combined (heterogeneous) set of organic compounds mentioned above. Success in the use of PI index prompted us to use the aforementioned topological indices for modeling lipophilicity (logP) of the combined set of 116 organic compounds (Table 1). In doing so we have adopted

*Corresponding author. Tel.: +91-731-253-1906; fax: +91-731-906; e-mail: vijay-agrawal@lycos.com (P.V.Khadikar); pvkhadikar@rediffmail.com (V.K. Agrawat).

0968-0896/03/$ - see front matter © 2003 Elsevier Ltd. All rights reserved.
doi:10.1016/S0968-0896(03)00370-5

Table 1. The compounds, their molecular lipophilicity, indicator parameters, and topological indices

Compd	Structural details of the compounds	LogP	W	$^1\chi$ (= B)	J	Sz	logRB	HW	Ip_1	Ip_2
1	CH_3F	0.51	1	1.0000	1.0000	1	0.0000	1	0	0
2	n-C_4H_9F	2.00	20	2.4142	2.1906	20	5.6630	35	0	0
3	CH_3Cl	0.91	1	1.0000	1.0000	1	0.0000	1	0	0
4	C_2H_5Cl	1.43	4	1.4142	1.6330	4	0.6931	5	0	0
5	n-C_3H_7Cl	2.04	10	1.9142	1.9747	10	2.4849	15	0	0
6	n-C_4H_9Cl	2.64	20	2.4142	2.1906	20	5.6630	35	0	0
7	n-$C_5H_{11}Cl$	3.11	35	2.9142	2.3390	35	10.4505	70	0	0
8	n-$C_6H_{13}Cl$	3.66	56	3.4142	2.4475	56	17.0297	126	0	0
9	n-$C_7H_{15}Cl$	4.15	84	3.9142	2.5301	84	25.5549	210	0	0
10	n-$C_8H_{17}Cl$	4.73	120	4.4142	2.5950	120	36.1595	330	0	0
11	CH_3Br	1.19	1	1.0000	1.0000	1	0.0000	1	0	0
12	C_2H_5Br	1.61	4	1.4142	1.6330	4	0.6931	5	0	0
13	n-C_3H_7Br	2.10	10	1.9142	1.9747	10	2.4849	15	0	0
14	n-C_4H_9Br	2.75	20	2.4142	2.1906	20	5.6630	35	0	0
15	n-$C_5H_{11}Br$	3.37	35	2.9142	2.3390	35	10.4505	70	0	0
16	n-$C_6H_{13}Br$	3.80	56	3.4142	2.4475	56	17.0297	126	0	0
17	n-$C_7H_{15}Br$	4.36	84	3.9142	2.5301	84	25.5549	210	0	0
18	n-$C_8H_{17}Br$	4.89	120	4.4142	2.5951	120	36.1595	330	0	0
19	CH_3I	1.51	1	1.0000	1.0000	1	0.0000	1	0	0
20	C_2H_5I	2.00	4	1.4142	1.6330	4	0.6931	5	0	0
21	n-C_3H_7I	2.54	10	1.9142	1.9747	10	2.4849	15	0	0
22	n-C_4H_9I	3.08	20	2.4142	2.1906	20	5.6630	35	0	0
23	n-$C_5H_{11}I$	3.62	35	2.9142	2.3390	35	10.4505	70	0	0
24	n-$C_6H_{13}I$	4.16	56	3.4142	2.4475	56	17.0297	126	0	0
25	n-$C_7H_{15}I$	4.70	84	3.9142	2.5301	84	25.5549	210	0	0
26	i-C_3H_7Cl	1.90	9	1.7321	2.3238	9	2.0794	12	0	0
27	i-C_3H_7Br	2.14	9	1.7321	2.3238	9	2.0794	12	0	0
28	i-C_4H_9Cl	2.33	16	2.3939	2.0797	28	3.8712	23	0	0
29	$Cl–CH_2–CH_2–Cl$	1.48	10	1.9142	1.9747	10	2.4849	15	0	0
30	$Br–CH_2–CH_2–Br$	1.96	10	1.9142	1.9747	10	2.4849	15	0	0
31	$I–CH_2–CH_2–I$	2.71	10	1.9142	1.9747	10	2.4849	15	0	0
32	$Cl–CH_2–CH_2–CH_2–Cl$	2.00	20	2.4142	2.1906	20	5.6630	35	0	0
33	$Br–CH_2–CH_2–CH_2–Br$	2.37	20	2.4142	2.1906	20	5.6630	35	0	0
34	$I–CH_2–CH_2–CH_2–I$	3.02	20	2.4142	2.1906	20	5.6630	35	0	0
35	$Br–CH_2–CH_2–CH_2–Cl$	2.18	20	2.4142	2.1906	20	5.6630	35	0	0
36	$CH_2–F_2$	0.20	4	1.4142	1.6330	4	0.6931	5	0	0
37	$CH–F_2–CH_3$	0.75	9	1.7321	2.3238	9	2.0794	12	0	0
38	$CH_2–Cl_2$	1.25	4	1.4142	1.6330	4	0.6931	5	0	0
39	$CH–Cl_2–CH_3$	1.79	9	1.7321	2.3238	9	2.0794	12	0	0
40	$CH–Cl_2–CH_2–Cl$	1.89	18	2.2701	2.5395	18	4.9698	28	0	0
41	$CH_2–Br–Cl$	1.41	4	1.4142	1.6330	4	0.6931	5	0	0
42	$CH_2–I_2$	2.30	4	1.4142	1.6330	4	0.6931	5	0	0
43	$CH–F_3$	0.64	9	1.7321	2.3238	9	2.0794	12	0	0
44	$CH–Cl_3$	1.97	9	1.7321	2.3238	9	2.0794	12	0	0
45	$CH_3–C–Cl_3$	2.49	16	2.0000	3.0237	16	4.1589	22	0	0
46	$CH–Br_3$	2.67	9	1.7321	2.3238	9	2.0794	12	0	0
47	$CH–Cl–F_2$	1.08	9	1.7321	2.3238	9	2.0794	12	0	0
48	$CH–Cl_2–F$	1.55	9	1.7321	2.3238	9	2.0794	12	0	0
49	$CH_3–OH$	−0.77	1	1.0000	1.0000	1	0.0000	1	0	1
50	$CH_3–CH_2–OH$	−0.31	4	1.4142	1.6330	4	0.6931	5	0	1
51	$CH_3–CH_2–CH_2–OH$	0.25	10	1.9142	1.9747	10	2.4849	15	0	1
52	$(CH_3)_2–CH–OH$	0.05	9	1.7321	2.3238	9	2.0794	12	0	1
53	$CH_3–CH_2–CH_2–CH_2–OH$	0.88	20	2.4142	2.1906	20	5.6630	35	0	1
54	$(CH_3)_2–CH–CH_2–OH$	0.65	18	2.2700	2.5395	18	4.9698	28	0	1
55	$CH_3–CH_2–CH–(CH_3)–OH$	0.61	18	2.2700	2.5395	18	4.9698	28	0	1
56	$(CH_3)_3–C–OH$	0.35	16	2.0000	3.0237	16	4.1589	22	0	1
57	$CH_3–CH_2–CH_2–CH_2–CH_2–OH$	1.56	35	2.9142	2.3390	35	10.4505	70	0	1
58	$(CH_3)_2–CH–CH_2–CH_2–OH$	1.16	32	2.7701	2.6272	32	9.5342	58	0	1
59	$(CH_3–CH_2)_2–CH–OH$	1.21	31	2.8081	2.7542	31	9.2465	54	0	1
60	$CH_3–CH_2–C(CH_3)_2–OH$	0.89	28	2.5607	3.1685	28	8.1479	44	0	1
61	$CH_3–C(CH_3)_2–CH_2–OH$	1.31	28	2.5607	3.1685	28	8.1479	44	0	1
62	$(CH_3)_2–CH–CH–(CH_3)–OH$	1.28	28	2.5607	3.1685	28	8.1479	44	0	1
63	$CH_3–CH_3$	1.81	1	1.0000	1.0000	1	0.0000	1	0	0
64	$CH_2{=}CH_2$	1.13	1	1.0000	1.0000	1	0.0000	1	0	0
65	$CH{\equiv}CH$	0.37	1	1.0000	1.0000	1	0.0000	1	0	0
66	$CH_3CH_2CH_3$	2.36	4	1.4142	1.6330	4	0.6931	5	0	0
67	$CH_3\ CH_2\ CH_2CH_3$	2.89	10	1.9142	1.9747	10	2.4849	15	0	0
68	$CH_3\ CH_2\ CH_2\ H_2CH_3$	3.39	20	2.4142	2.1906	20	5.6630	35	0	0
69	Cyclo-propane	1.72	3	1.5000	2.2500	3	0.0000	3	0	0
70	Cyclo-pentane	3.00	15	2.5000	2.0833	20	3.4657	20	0	0
71	Cyclo-hexane	3.44	27	3.0000	2.0000	54	7.4547	42	0	0

(continued)

Table 1 (*continued*)

Compd	Structural details of the compounds	LogP	W	$^1\chi$ (=B)	J	Sz	logRB	HW	Ip_1	Ip_2
72	Benzene	2.13	27	3.0000	2.0000	54	7.4547	42	1	0
73	Tolune	2.73	42	3.3939	2.1230	78	12.4245	71	1	0
74	Naphthalene	3.30	109	4.9663	1.9253	243	34.424	215	1	0
75	Chlorobenzene	2.58	42	3.3939	2.1230	78	12.4245	71	1	0
76	Phenol	1.49	42	3.3939	2.1230	78	12.4245	71	1	1
77	Pentachlorophenol	4.90	174	5.4641	2.7603	282	57.0114	357	1	1
78	Hexachlorobenzene	5.27	174	5.4641	2.7603	282	57.0114	357	1	0
79	Biphenyl	3.91	198	5.9663	1.7997	360	62.3223	477	1	0
80	CF_4	1.18	16	2.0000	3.0237	16	4.1589	22	0	0
81	CCl_4	2.83	16	2.0000	3.0237	16	4.1589	22	0	0
82	CBr_4	3.42	16	2.0000	3.0237	16	4.1589	22	0	0
83	Benzoic acid	1.87	88	4.3045	2.2284	142	27.6625	176	1	0
84	4-Methyl-benzoic acid	2.36	120	4.6984	2.2599	192	37.8252	262	1	0
85	3-Methyl-benzoic acid	2.37	117	4.6984	2.3199	186	37.2374	245	1	0
86	2-Methyl-benzoic acid	2.18	114	4.7152	2.3960	180	36.5035	231	1	0
87	2,6-di-Methyl benzoic acid	2.21	144	5.1259	2.5572	224	46.7308	296	1	0
88	4-et-Benzoic acid	2.89	162	5.2364	2.2427	252	50.7812	390	1	0
89	4-Propyl benzoic acid	3.42	215	5.7364	2.2008	323	66.4610	571	1	0
90	4-*iso*-Propyl benzoic acid	3.40	206	5.6091	2.2951	314	64.4304	521	1	0
91	4–Butyl-benzoic acid	3.97	280	6.2364	2.1485	406	84.8612	817	1	0
92	4-*tert*-Butyl-benzoic acid	3.85	72	4.2694	1.5926	190	21.6710	120	1	0
93	3-F-benzoic acid	2.15	117	4.6984	2.3199	186	37.2374	245	1	0
94	4-F-benzoic acid	2.07	120	4.6984	2.2599	192	37.8252	262	1	0
95	2-F-benzoic acid	1.77	114	4.7152	2.3960	180	36.5035	231	1	0
96	3-Cl-benzoic acid	2.68	117	4.6984	2.3199	186	37.2374	245	1	0
97	4-Cl-benzoic acid	2.65	120	4.6984	2.2599	192	37.8252	262	1	0
98	2-Cl-benzoic acid	2.05	114	4.7152	2.3960	180	36.5035	231	1	0
99	3-Br-benzoic acid	2.87	117	4.6984	2.3199	186	37.2374	245	1	0
100	4-Br-benzoic acid	2.86	120	4.6984	2.2599	192	37.8252	262	1	0
101	2-Br-benzoic acid	2.20	114	4.7152	2.3960	180	36.5035	231	1	0
102	3-I-benzoic acid	3.13	117	4.6984	2.3199	186	37.2374	245	1	0
103	4-I-benzoic acid	3.02	120	4.6984	2.2599	192	37.8252	262	1	0
104	2-I-benzoic acid	2.40	114	4.7152	2.3960	180	36.5035	231	1	0
105	3-CH_3O-benzoic acid	2.02	156	5.2364	2.3303	240	49.7028	353	1	0
106	4-CH_3O-benzoic acid	1.96	162	5.2364	2.2427	252	50.7812	390	1	0
107	2-CH_3O-benzoic acid	1.59	150	5.2532	2.4430	228	48.3811	322	1	0
108	3-NO_2-benzoic acid	1.83	197	5.6091	2.4024	296	62.8613	464	1	0
109	4-NO_2-benzoic acid	1.89	206	5.6091	2.2951	314	64.4304	521	1	0
110	2-NO_2-benzoic acid	1.46	188	5.6259	2.5409	278	60.9518	416	1	0
111	3-CF_3-benzoic acid	2.95	240	5.9097	2.5158	354	76.7129	578	1	0
112	3-CN-benzoic acid	1.48	156	5.2364	2.3303	240	49.7028	353	1	0
113	4-CN-benzoic acid	1.56	162	5.2364	2.2427	252	50.7812	390	1	0
114	3-OH-benzoic acid	1.50	156	5.2364	2.3303	240	49.7028	353	1	1
115	4-OH-benzoic acid	1.58	162	5.2364	2.2427	252	50.7812	390	1	1
116	2-OH-benzoic acid	2.26	150	5.2532	2.4430	228	48.3811	322	1	1

$Ip_1 = 1$ if the compdound is aromatic, otherwise 0; $Ip_2 = 1$ if OH group is present in compd, otherwise 0.

lipophilicity (logP) of these compounds as reported by Mannhold and coworkers.[15]

Results and Discussion

The set of 116 organic compounds, their lipophilicity (logP) and indicator parameters Ip_1 and Ip_2 are presented in Table 1. The indicator parameter Ip_1 is used as unity if the compound is aromatic, otherwise its value is zero. Similarly, if –OH group is present in the molecules then Ip_2 is one, otherwise it is zero.

The topological indices W, Sz, HW, $^1\chi$, B, J and logRB are calculated using methodology described in the Experimental and are summarized in Table 1. The correlatedness among the topological indices used and their correlation with the lipophilicity (logP) is demonstrated in Table 2.

A perusal of Table 2 indicates high collinearity exists among W, Sz, logRB, HW, $^1\chi$ and B indices. The Balaban index (J) does not correlate with any other topological index used. This shows that it is most appropriate topological index to be used in multi-parametric regression analysis. In addition, this Table 2 also shows that it is the indicator parameter $Ip_{1'}$ which except for the Balaban index J, correlates significantly with all other topological indices used. However, none of the topological indices, including indicator parameters, correlates significantly well with the lipophilicity (logP). It means that none of the molecular descriptors used is capable of modeling lipophilicity (logP) in mono-parametric regression.

Likewise of Mannhold and coworkers[15] observed that like topological indices, none of the single fragmental constant is capable of modeling lipophilicity (logP). It is the sum of such fragmental constants which resulted into statistically fruitful model for modeling lipophilicity (logP). It means that like-wise some combinations of topological indices from the larger pool (Table 1) will be useful for modeling lipophilicity logP. In view of this we have attempted several multi-parametric regressions.[16–19] The statistically significant results are recorded in Table 3.

Table 2. Correlation matrix for the correlation of molecular descriptors and their correlation with molecular lipophilicity (logP)

	LogP	W	$^1\chi = B$	J	Sz	log-RB	HW	Ip_1	Ip_2
LogP	1.0000								
W	0.3663	1.0000							
$^1\chi = B$	0.4466	0.9571	1.0000						
J	0.2164	0.2584	0.3803	1.0000					
Sz	0.3219	0.9840	0.9501	0.1903	1.0000				
log-RB	0.3558	0.9995	0.9586	0.2556	0.9852	1.0000			
HW	0.3824	0.9902	0.9196	0.2276	0.9620	0.9865	1.0000		
Ip_1	0.1650	0.8253	0.8564	0.1114	0.8845	0.8326	0.7635	1.0000	
Ip_2	−0.5320	−0.1493	−0.1278	0.2204	−0.1589	−0.1499	−0.1523	−0.15194	1.0000

Table 3. Regression parameters and quality of correlations for correlation of structural descriptors with logP

Model no.	Parameters used	$A_i = 1, 2, 3\ldots$	Intercept (B)	SE	R	R^2	F-ratio	$Q = R/SD$	Prob.
1.	W	0.0150(±0.0040)	0.5139	0.7869	0.7035	0.4949	34.623	0.8940	3×10^{-14}
	$^1\chi(=B)$	0.8975(±0.1734)							
	Ip_2	−1.5252(±0.2102)							
2.	$^1\chi(=B)$	0.8355(±0.0817)	0.5219	0.6619	0.8017	0.6427	63.543	1.2112	1×10^{-4}
	Ip_1	−2.0545(±0.2579)							
	Ip_2	−1.5684(±0.1766)							
3.	$^1\chi(=B)$	1.0200(±0.1721)	0.2770	0.7621	0.7212	0.5201	38.288	0.9463	0.00
	logRB	0.0563(±0.0125)							
	Ip_2	−1.5409(±0.2050)							
4.	$^1\chi(=B)$	1.4748(±0.1435)	−0.5452	0.5935	0.8458	0.7154	65.987	1.4251	0.00
	logRB	−0.0504(±0.0097)							
	Ip_1	−1.9685(±0.2319)							
	Ip_2	−1.6433(±0.1590)							
5.	W	−0.0142±0.0031)	−0.4111	0.6074	0.8378	0.7019	61.815	1.3793	0.00
	$^1\chi(=B)$	1.4118(±0.1467)							
	Ip_1	−2.0221(±0.2368)							
	Ip_2	−1.6361(±0.1628)							
6.	$^1\chi(=B)$	1.1628(±0.1304)	−0.0464	0.6358	0.8206	0.6734	54.120	1.29065	0.00
	HW	−0.0030(±9.6741×10^{-4})							
	Ip_1	−2.1542(±0.2498)							
	Ip_2	−1.6221(±0.1705)							
7.	$^1\chi(=B)$	0.8427(±0.1000)	0.5520	0.6650	0.8017	0.6427	47.219	1.2055	0.00
	J	−0.0219(±0.1740)							
	Ip_1	−2.0708(±0.2895)							
	Ip_2	−1.5617(±0.1853)							
8.	W	−0.0174(±0.0033)	−0.0824	0.5956	0.8463	0.7161	52.478	1.4209	0.00
	$^1\chi(=B)$	1.6691(±0.1828)							
	J	−0.3910(±0.1713)							
	Ip_1	−2.3047(±0.2631)							
	Ip_2	−1.5315(±0.1661)							
9.	$^1\chi(=B)$	1.5020(±0.1714)	0.2183	0.6101	0.8379	0.7021	49.019	1.3733	0.00
	J	−0.4282(±0.1829)							
	Sz	−0.0106(±0.0023)							

(continued)

Table 3 (*continued*)

Model no.	Parameters used	A_i = 1, 2, 3...	Intercept (B)	SE	R	R^2	F-ratio	$Q = R$/SD	Prob.
	Ip_1	−1.7734(±0.2735)							
	Ip_2	−1.5180(±0.1703)							
10.	W	0.1784(±0.0287)	−0.8736	0.5092	0.8902	0.7925	79.430	1.7428	5×10^{-4}
	$^1\chi$(=b)	1.4545(±0.1232)							
	logRB	−0.6199(±0.0920)							
	Ip_1	−1.4037(±0.2187)							
	Ip_2	−1.6387(±0.1365)							
11.	$^1\chi$(=B)	1.7613 (±0.1777)	−0.1894	0.5778	0.8561	0.7329	57.065	1.4816	0.00
	J	−0.4337(±0.1664)							
	logRB	−0.0617(±0.0104)							
	Ip_1	−2.2707(±0.2537)							
	Ip_2	−1.5274(±0.1611							
12.	$^1\chi$(=B)	1.5153(±0.1456)	−0.6040	0.5906	0.8491	0.7209	53.723	1.4376	0.00
	Sz	0.0063(±0.0044)							
	logRB	−0.0784(±0.0219)							
	Ip_1	−2.2795(±0.3171)							
	Ip_2	−1.6366(±0.1583)							
13.	W	0.1792(±0.0276)	−0.5128	0.4889	0.9003	0.8106	73.456	1.8414	0.00
	$^1\chi$(=B)	1.7460(±0.1504)							
	J	−0.4415(±0.1408)							
	logRB	−0.6338(±0.0885)							
	Ip_1	−1.7089(±0.2315)							
	Ip_2	−1.5207(±0.1363)							
14.	$^1\chi$(=B)	2.0769(±0.1594)	−0.7950	0.4924	0.8988	0.8078	72.173	1.8253	0.00
	J	−0.3303(±0.1427)							
	logRB	−0.2296(±0.0279)							
	HW	0.0166(±0.0026)							
	Ip_1	−1.3646(±0.2592)							
	Ip_2	−1.5166(±0.1373)							
15.	W	−0.1094 (±0.0148)	−1.0440	0.4976	0.8976	0.8057	60.413	1.8038	0.00
	$^1\chi$(=B)	2.2597(±0.1753)							
	J	−0.2038(±0.1522)							
	Sz	0.0055(±0.0041)							
	HW	0.0261 (±0.0038)							
	Ip_1	−1.4112(±0.3111)							
	Ip_2	−1.5248(±0.1389)							

W, Wiener Index; $^1\chi$, first-order connectivity index = Branching Index; J, Balaban Index; Sz, Szeged Index; logRB; HW, Hyper Wiener; Ip_2, and Ip_2 are indicator parameters; A and B are the correlation coefficient; R, multiple correlation coefficient; R^2, coefficient of determination; SE, standard error of estimation; Q, quality factor (R/SE).

The regression parameters[16] and quality of correlations[17,18] (Table 3) indicate that statistically significant model starts coming from tri- to higher parametric regression expressions.

The data presented in Table 3 show that there are three tri-parametric-, four tetra-parametric-, five penta-parametric, two hexa-parametric and one hepta-parametric models which are statistically significant for modeling lipophilicity logP. Furthermore, the data presented in Table 4 show that the promising topological index is first-order connectivity index ($^1\chi$) and that the branching index (B) is found to be the same as $^1\chi$.

It is worthy to record that in obtaining statistically significant regressions (Table 3) we have to eliminate compounds **36**, **77**, **78**, **82**, **111** and **116** as outliers. They are, therefore, discarded in the regression procedure. We have, therefore, left with 110 compounds for further regression analysis. At present, we can not give any convincing reason for the occurrence of such compounds as outliers. Perhaps it is the outcome of the regression procedure for obtaining the statistically best model.

Out of the three tri-parametric models, the model containing $^1\chi$, Ip_1 and Ip_2 gave better results. This model is found as:

$$\log P = 0.5219 + 0.8355(\pm 0.0817)^1\chi - 2.0545 \times (\pm 0.2579)Ip_1 - 1.5684(\pm 0.1768)Ip_2$$

$$n = 110, SE = 0.7869, \quad R = 0.8017, \quad F = 57.120, \quad Q = 1.2112 \tag{1}$$

Table 4. Comparison of observed and estimated lipophilicity (logP) using model 13

Compd	Observed logP	Estimated logP	Residual	Compd	Observed logP	Estimated logP	Residual
1	0.51	0.971	−0.461	**59**	1.21	1.348	−0.138
2	2.00	2.730	0.730	**60**	0.89	0.892	−0.002
3	0.91	0.971	−0.061	**61**	1.31	0.892	0.418
4	1.43	1.513	−0.083	**62**	1.28	0.892	0.388
5	2.04	2.174	−0.135	**63**	1.81	0.971	0.839
6	2.64	2.73	−0.091	**64**	1.13	0.971	0.159
7	3.11	3.191	−0.081	**65**	0.37	0.971	−0.601
8	3.66	3.609	0.051	**66**	2.36	1.513	0.847
9	4.15	4.060	0.09	**67**	2.89	2.174	0.716
10	4.73	4.633	0.097	**68**	3.39	2.730	0.66
11	1.19	0.971	0.219	**69**	1.72	1.650	0.07
12	1.61	1.513	0.097	**70**	3.00	3.424	−0.424
13	2.10	2.174	−0.075	**71**	3.44	3.955	−0.515
14	2.75	2.730	0.02	**72**	2.13	2.247	−0.117
15	3.37	3.191	0.179	**73**	2.73	2.418	0.312
16	3.80	3.609	0.191	**74**	3.30	3.313	−0.013
17	4.36	4.060	0.301	**75**	2.58	2.418	0.162
18	4.89	4.633	0.257	**76**	1.49	0.897	0.593
19	1.51	0.971	0.539	**77**	4.90		
20	2.00	1.513	0.487	**78**	5.27		
21	2.54	2.174	0.366	**79**	3.91	3.380	0.53
22	3.08	2.730	0.35	**80**	1.18	1.875	−0.695
23	3.62	3.191	0.429	**81**	2.83	1.875	0.955
24	4.16	3.609	0.551	**82**	3.42	—	—
25	4.70	4.060	0.641	**83**	1.87	2.546	−0.676
26	1.90	1.780	0.12	**84**	2.36	2.513	−0.153
27	2.14	1.780	0.36	**85**	2.37	2.322	0.048
28	2.33	3.162	−0.832	**86**	2.18	2.245	−0.065
29	1.48	2.174	−0.695	**87**	2.21	1.784	0.426
30	1.96	2.174	−0.215	**88**	2.89	2.774	0.116
31	2.71	2.174	0.536	**89**	3.42	3.225	0.195
32	2.00	2.730	−0.73	**90**	3.40	2.635	0.765
33	2.37	2.730	−0.36	**91**	3.97	4.106	−0.136
34	3.02	2.730	0.290	**92**	3.85	3.696	0.154
35	2.18	2.730	−0.55	**93**	2.15	2.322	−0.172
36	0.20	—	—	**94**	2.07	2.513	−0.443
37	0.75	1.780	−1.03	**95**	1.77	2.245	−0.475
38	1.25	1.513	−0.263	**96**	2.68	2.322	0.358
39	1.79	1.780	0.01	**97**	2.65	2.513	0.137
40	1.89	2.405	−0.515	**98**	2.05	2.245	−0.195
41	1.41	1.513	−0.103	**99**	2.87	2.322	0.548
42	2.30	1.513	0.787	**100**	2.86	2.513	0.347
43	0.64	1.780	−1.14	**101**	2.20	2.245	−0.045
44	1.97	1.780	0.19	**102**	3.13	2.322	0.808
45	2.49	1.875	0.615	**103**	3.02	2.513	0.507
46	2.67	1.780	0.89	**104**	2.40	2.245	0.155
47	1.08	1.780	−0.7	**105**	2.02	2.344	−0.324
48	1.55	1.780	−0.23	**106**	1.96	2.774	−0.814
49	−0.77	−0.550	−0.22	**107**	1.59	2.086	−0.496
50	−0.31	-7.8222×10^{-3}	−0.302	**108**	1.83	1.970	−0.14
51	0.25	0.654	−0.404	**109**	1.89	2.635	−0.745
52	0.05	0.260	−0.21	**110**	1.46	1.536	−0.076
53	0.88	1.209	−0.329	**111**	2.95	—	—
54	0.65	0.884	−0.234	**112**	1.48	2.344	−0.864
55	0.61	0.884	−0.274	**113**	1.56	2.774	−1.214
56	0.35	0.355	−0.005	**114**	1.50	0.823	0.677
57	1.56	1.670	−0.11	**115**	1.58	1.25	0.33
58	1.16	1.334	−0.174	**116**	2.26	—	—

Here and hereafter, n is the number of compounds used, SE is the standard error of estimation, R is the multiple correlation coefficient, F is the F-ratio and Q is the quality factor.

The positive sign associated with $^1\chi$ indicates favorable contribution of branching in the exhibition of lipophilicity (logP). The negative signs associated with Ip_1 and Ip_2 indicate their negative role towards lipophilicity.

The step-wise regression resulted into four tetra-parametric models (Table 3) showing that the models 4–6 give better results than the tri-parametric model discussed above and that the model-4 containing $^1\chi$, logRB, Ip_1, and Ip_2 is the best tetra-parametric model. This model is found as below:

$$\log P = -0.5412 + 1.4748(\pm 0.1435)\,^1\chi - 0.0504 \times (\pm 0.0097)\log RB - 1.9685 \times (\pm 0.2319) Ip_1 - 1.6433(\pm 0.1590) Ip_2 \quad (2)$$

$$n = 110, \quad SE = 0.5935, \quad Rr = 0.8456,$$

$$F = 65.987, \quad Q = 1.4251$$

The above mentioned model (eq 2) once again show that first-order connectivity index ($^1\chi$) is favorable for the modeling of lipophilicity (logP) of the organic compounds used.

Successive regression analysis resulted into five (8–12) penta-parametric models in that models 8, 10–12 gave better results than the tetra-parametric model discussed above.

Furthermore, the model-10 consisting of W, $^1\chi$, logRB, Ip_1, and Ip_2 is the best model among the five penta-parametric models (Table 3). This model is found as below:

$$\log P = -0.8736 + 0.1784(\pm 0.0287)W + 1.4545 \times (\pm 0.1232)\,^1\chi - 0.6199(\pm 0.0920)\log RB - 1.4037(\pm 0.2187) Ip_1 - 1.6387(\pm 0.1365) Ip_2 \quad (3)$$

$$n = 110, \quad SE = 0.5092, \quad Rr = 0.8902,$$

$$F = 79.430, \quad Q = 1.7428$$

Once again, we observe that first order connectivity index ($^1\chi$) is mainly responsible for the lipophilicity. The positive sign associated with Wiener index (W) indicates that the size and shape further help in enhancing lipophilicity of the compounds used.

Finally, two hexa-parametric models, both having better statistics than the penta-parametric model given above are obtained (Table 3). Out of these two models the model consisting of W, $^1\chi$, J, logRB, Ip_1 and Ip_2, as given below, was found better for modeling lipophilicity (logP) of the compounds used.

$$\log P = -0.5128 + 0.1792(\pm 0.0270)W + 1.7460 \times (\pm 0.1504)\,^1\chi - 0.4415(\pm 0.1408)J - 0.6338(\pm 0.0885)\log RB - 1.7089 \times (\pm 0.2315) Ip_1 - 1.5207(\pm 0.1363) Ip_2 \quad (4)$$

$$n = 110, \quad SE = 0.4889, \quad R = 0.9003,$$

$$F = 73.456, \quad Q = 1.8414$$

Further, multi-parametric regression resulted into statistically significant hepta-parametric model-15. However, it has to be discarded on the grounds that its quality is poor than the hexa-parametric model discussed above.

The aforementioned results, therefore, indicate that lipophilicity of the large and heterogeneous set of 116 organic compounds can be modeled successfully by the combination of topological indices and indicator parameters. Also, that such a combined model as given by (eq 4) is the best for this purpose.

In order to confirm our findings we have estimated lipophilicity (logP) of the organic compounds used and compared them with the observed lipophilicity. Such a comparison (Table 4) and the observed residue, that is difference between observed and estimated lipophilicity are in favor of our findings.

In order to confirm our finding we have estimated predictive correlation coefficient by correlating observed and estimated lipophilicity (logP) (Fig. 1). The $R^2 = 0.7476$ obtained from Figure 1 is in favor of our proposed model expressed by (eq 4).

The quality factor Q is a useful parameter to be used in deciding predictive potential of the model. The higher the value of Q the better is the predictive potential of the models. In our case highest value of Q is observed for the model expressed by (eq 4). Thus, amongst all the proposed model this is the best model for modeling the lipophilicity of the compounds used. Recently, the use of Q factor has been criticized;[19] however, we found this to be a useful parameter in deciding predictive potential of the model.

Looking to the reservation for the use of Q for estimating predictive ability of the model, we have used cross-validation method for this purpose. Consequently, we have estimated various cross-validation parameters and recorded them in Table 5.

PRESS (predictive residual sum of squares) is a good estimate of the real prediction error of the model. If PRESS is smaller than sum of the squares of response value that is SSY the model predict better than chance

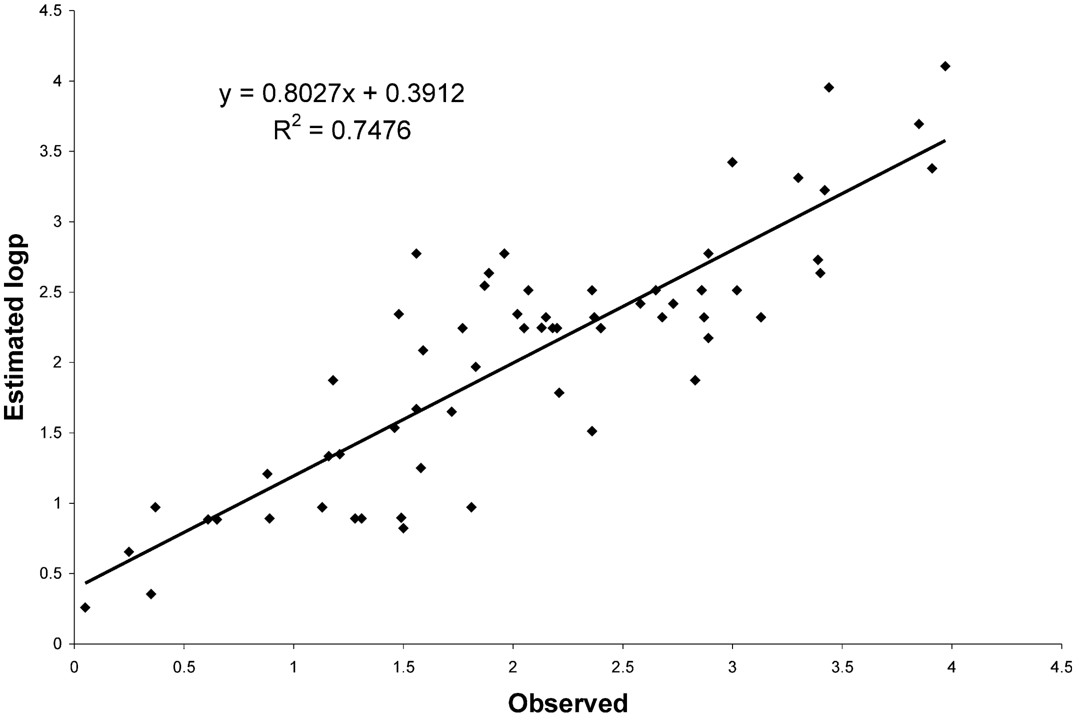

Figure 1. Correlation of observed and estimated lipophilicity (logP) using model 13.

Table 5. Cross-validated parameters for the proposed models

Model	Number of parameters	PRESS	SSY	PRESS/SSY	R^2_{CV}	S_{PRESS}	PSE
1	3	46.4410	83.5193	0.5561	0.4439	0.6619	0.6498
2	4	36.1857	92.9746	0.3980	0.6022	0.5871	0.5799
3	5	26.9698	102.9905	0.2619	0.7381	0.5092	0.4952
4	6	24.6185	105.3418	0.2337	0.7663	0.4889	0.4731

can be considered statistically significant. In this regard, all the models (Table 5) are statistically significant. Furthermore, the ratio PRESS/SSY should be smaller than 0.4, and value of this ratio smaller than 0.1 indicates an excellent model with high predictive potential.

In view of the above, we observed that (Table 5) models 4–6 are the excellent models for modeling lipophilicity of heterogeneous set of organic compounds chosen by us. Furthermore, model 6 has the lowest value of the aforementioned ratio establishing its superiority over the others. The highest value of cross-validation correlation coefficient (R^2_{cv}) further confirms this finding.

Another useful cross-validation parameter is the uncertainty of prediction (S_{PRESS}). However, in the present case this parameter is of no use as its magnitude is the same as that of standard error of estimation (SE). In such cases an important cross-validation parameter named as predictive square error (PSE) is available (Table 5). This parameter is more directly related to the uncertainty of the predictions. The lowest value of PSE for model 6 finally confirms its excellent predictive potential.

Conclusion

The aforementioned results and discussion show that the combination of topological indices is useful for quantifying molecular lipophilicity (logP) and that it is similar to hydrophobic fragmental constant approach. Unlike the latter approach in our approach (based on topological indices) it is not necessary to work with a set of compounds of similar nature and family.

Experimental

Lipophilicity

The molecular lipophilicity (logP) used by Mannhold et al.[15] are adopted in the present study.

Molecular graphs

The hydrogen suppressed molecular graphs[20] were used for the calculation of topological indices W, Sz, $^1\chi$, B, J and logRB (Table 1).

Topological indices

The details for the calcculations of Wiener(W),[8] Szeged (Sz),[9,10] first-order connectivity index ($^1\chi$),[13] branching index(B),[11] Hyper-Wiener (HW),[14] logRB,[14] and Balaban (J)[12] indices, quality factor (Q)[16–19] and maximum R^2 improvement method are given in our earlier communications and they are thus not repeated here.

Multiple regression analyses for correlating tadpole narcosis of the present set of compounds with the aforementioned molecular descriptors were carried out using *Regress-1* software as supplied by Professor I. Lukovits, Hungarian Academy of Sciences, Budapest, Hungary. Several multiple regressions were attempted using correlation matrix from this program and the best results were considered and discussed in developing QSAR and hence, for modeling lipophilicity of heterogeneous set of organic compounds.

Computations

All the computations were carried out in Power Macintosh 9600/233.

Acknowledgements

The authors are thankful to Professor Istvan Lukovits, Hungarian Academy of Sciences, Budapest, Hungary for providing software to carryout regression analysis and to Prof. Ivan Gutman, Faculty of Science, University of Kragujevac, Yugoslavia for introducing one of the authors (P.V.K.) to this fascinating field of chemical graph theory and topology. Authors are also thankful to CSIR, New Delhi, India for sanctioning a research scheme.

References and Notes

1. Agrawal, V. K.; Singh, J.; Khadikar, P. V. *Bioorg. Med. Chem.* **2002**, *10*, 3981.
2. Khadikar, P. V.; Agrawal, V. K.; Karmarkar, S. *Bioorg. Med. Chem.* **2002**, *10*, 3499.
3. Singh, J. *On Topological Modeling of QSAR: A Multivariate Analysis*. PhD Thesis, A.P.S. University, Rewa, India, 2003.
4. Khadikar, P. V.; Singh, S.; Shrivastava, A. *Bioorg. Med. Chem.* **2002**, *12*, 1125.
5. Khadikar, P. V.; Phadnis, A.; Shrivastava, A. *Bioorg. Med. Chem.* **2002**, *10*, 1181.
6. Khadikar, P. V.; Karmarkar, S.; Agrawal, V. K. *J. Chem. Inf. Comput. Sci.* **2000**, *41*, 934.
7. Hansch, C.; Leo, A.; Hockman, D., Eds. *Exploring QSAR: Hydrophobic, Electronic and Steric Constatnts*; ACS Professional Reference Book, Washington, DC, 1995; p 110.
8. Wiener, H. *J. Am. Chem. Soc.* **1947**, *69*, 17.
9. Gutman, I. *Graph Theory Notes New York* **1994**, *27*, 9.
10. Khadikar, P. V.; Deshpande, N. V.; Kale, P. P.; Dobrynin, A.; Gutman, I.; Domotor, G. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 547.
11. Devillers, J.; Balaban, A. T. *Topological Indices and Related Descriptors in QSAR and QSPR*; Gorden & Breach: Williston, VT, 2000; p 40, 245.
12. Balaban, A. T. *Chem. Phys. Lett.* **1982**, *89*, 399.
13. Randic, M. *J. Am. Chem. Soc.* **1975**, *97*, 6609.
14. Diudea, M. V., Ed. *QSPR/QSAR Studies by Molecular Descriptors*. Babes-Bolyai University: Cluj, Romania 2000 p 31.
15. Mannhold, R.; Rekker, R. F.; Dross, K.; Bijloo, G.; de Vries, G. *Quant. Struct-Act. Relat* **1998**, *17*, 517.
16. Chaterjee, S.; Hadi, A. S.; Price, B. *Regression Analysis by Examples,* 3rd Ed; Wiley: New York, 2000.
17. Pogliani, L. *Amino Acids* **1994**, *6*, 141.
18. Pogliani, L. *J. Phys. Chem.* **1996**, *100*, 18065.
19. Todeschini, R. *Chemometrics Web News*; Milano Chemometric & QSAR Research Group: file: ///C/Windows/Desktop/Web news on chemometrics.html Feb, 2001.
20. Gutman, I.; Popovil, L.; Khadikar, P. V.; Karmarkar, S.; Joshi, S.; Mandloi, M. *Commun. Math. Comput. Chem. (MATCH)* **1997**, *35*, 95, and references therein.