J Mol Model (2005) 11: 89–96
DOI 10.1007/s00894-004-0224-7

ORIGINAL PAPER

Kunal Roy · Andrey A. Toropov

QSPR modeling of the water solubility of diverse functional aliphatic compounds by optimization of correlation weights of local graph invariants

Received: 2 April 2004 / Accepted: 2 November 2004 / Published online: 29 January 2005

Abstract The optimization of correlation weights scheme was used to model the water solubility (ln S) of diverse functional aliphatic compounds ($n = 193$). The optimized descriptor formulated based on the data of a training set ($n = 96$) generated statistically acceptable relations for the training set ($r^2 = 0.987$), test set ($n = 97$; $r^2 = 0.986$) and combined set ($r^2 = 0.987$). When the relation of ln S values with the optimized molecular descriptor formulated based on the data of the training set was used for the calculation of ln S values of the training set, r^2_{pred} value was found to be satisfactory (0.988), which is indicative of the predictive potential of the scheme. The results indicate the promising potential of the optimization of correlation weights scheme in modeling studies.

Keywords QSAR · QSPR · Optimization of correlation weights · Flexible descriptors · Nearest neighbouring codes · Water solubility

Introduction

Numerical representation of chemical structure and its relation with property or biological activity have lead to the fascinating fields of quantitative structure-activity/property/toxicity relationship (QSAR/QSPR/QSTR) studies. Among the different descriptors available, topological ones, formulated by graph theoretic approaches, [1–3] have been used extensively in modeling studies because of their ease of computation and low computational requirements [4–11]. Topological descriptors consider the arrangements of atoms in the (mostly hydrogen-suppressed) molecular graph, interatomic distance, kind of atoms, branching and cyclicity.

A huge number of topological descriptors are currently available for modeling studies. Although many such descriptors are highly intercorrelated, a large amount of chemical information can be decoded by the use of an appropriate combination of useful descriptors. Selection of appropriate descriptors from the plethora of available descriptors is a real problem in modeling studies. One has to take care that descriptors are chosen to extract the maximum amount of chemical information and, at the same time, the descriptors used in a multiple regression equation should not be inter-correlated. The concept of flexible topological descriptors, originally introduced by Randic, [12–14] is a major breakthrough in this regard as the difficulties of multiple regression are not present in such an approach. Flexible topological descriptors do not have a definite predetermined formalism, that can be applied to any sets of compounds for modeling biological activity or physicochemical properties. The formalism of such descriptors is defined based on an optimization procedure to obtain the best relation for a particular data set. Thus, the definition of the descriptors will vary from one data set to another and the ultimate objective of the iterative optimization procedure is to obtain the best predictive model. Several descriptors have been proposed in this line and their use has also been explored [15–22]. Among these descriptors, an interesting sort of flexible descriptors is based on the optimized correlation weights of the local graph invariants [19–21]. This scheme has been used successfully to model different sets of biological activity and physicochemical property data [23–32].

Like partition coefficient parameter in the n-octanol—water system, [33–39] water solubility is another

K. Roy (✉)
Drug Theoretics and Cheminformatics Lab,
Division of Medicinal and Pharmaceutical Chemistry,
Department of Pharmaceutical Technology,
Jadavpur University, Kolkata, 700 032, India
E-mail: kunalroy_in@yahoo.com
Tel.: +91-33-24146676
Fax: +91-33-24146677
http://www.geocities.com/kunalroy_in

A. A. Toropov
Uzbekistan Academy of Sciences Research Institute
'Algorithm-Engineering', 700125 F. Khodjaeva Street 25,
Tashkent, Uzbekistan

very important physicochemical parameter that can account for many properties of organic chemicals including the biopharmaceutical behavior of drugs [40]. Many attempts have been made to model water solubility using different indices, e.g., the Wiener and connectivity indices, [41] the PI index, [42] quantum chemical descriptors, [43, 44] dipole moment, surface area, volume, molecular weight, number of hydrogen bond acceptor/donor(s) and number of rotable bonds, [45] the TAU index, [46] the modified Wiener index, [47] etc., and different statistical and QSAR methods, e.g., genetic algorithm and partial least squares, [48] principal component analysis, [49] comparative molecular field analysis, [50] artificial neural network, [51] SIMCA, [52] etc.

In the present communication, we have applied the optimization of correlation weights scheme for modeling water solubility of diverse functional aliphatic compounds to show the usefulness of the scheme. Although mostly straight chain aliphatic compounds have been considered, the data set also contains a few alicyclic compounds.

Materials and methods

The molecular descriptor used in the present modeling studies was calculated based on the labelled hydrogen filled graph (LHFG) in the following manner:

$$\mathrm{DCW}(\mathrm{a}_k, \mathrm{LI}_k) = \sum_{k=1}^{n} \mathrm{CW}(\mathrm{a}_k) + \sum_{k=1}^{n} \mathrm{CW}(\mathrm{LI}_k). \quad (1)$$

In the above equation, the DCW term represents the molecular descriptor, the CW terms represent the correlation weights, a_k is the chemical element of the kth vertex of the LHFG and LI_k is the numerical value of a local invariant of the LHFG. As local invariants, we have used nearest neighboring codes (NNC). [32] The NNC of the kth vertex of the LHFG is calculated as

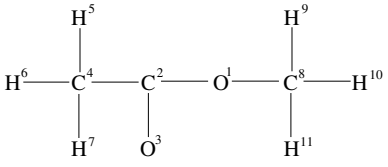
$$\mathrm{NNC}_k = 100N_\mathrm{T} + 10N_\mathrm{C} + N_\mathrm{H}. \quad (2)$$

In the above equation, N_T, N_C and N_H represent the total number of vertices, number of carbons and number of hydrogens, respectively, connected to the kth vertex. An example of the calculation of NNC for methyl acetate is shown in Table 1. NNC is a mathematical function of both the number and kind of neighbors for an atom.

The descriptor (DCW), as defined in Eq. 1, is obtained from special correlation weights of local graph invariants, which are obtained by a Monte Carlo optimization procedure. The aim of this optimization procedure is to make the correlation coefficient between the property/activity of the training set under consideration and the descriptor (DCW) as large as possible. The predictive ability of the model should be validated using a test set.

The water solubility (ln S) values of diverse functional aliphatic compounds ($n = 193$) were taken from the literature. [8, 53] The data set was divided into a training set and a test set, as listed in Table 2. The starting value of each correlation weight was 1 and using a Monte Carlo iterative optimization procedure, [20, 21, 22] the best values of correlation weights [$\mathrm{CW}(\mathrm{a}_k)$ and $\mathrm{CW}(\mathrm{LI}_k)$] (which give largest possible correlation

Table 1 Example of the calculation the DCW value of methyl acetate based on the CWs listed in Table 3 [the adjacency matrix of methyl acetate is also shown]



Atom (a_k)	O1	C2	O3	C4	H5	H6	H7	C8	H9	H10	H11
O1	0	1	0	0	0	0	0	1	0	0	0
C2	1	0	1	1	0	0	0	0	0	0	0
O3	0	1	0	0	0	0	0	0	0	0	0
C4	0	1	0	0	1	1	1	0	0	0	0
H5	0	0	0	1	0	0	0	0	0	0	0
H6	0	0	0	1	0	0	0	0	0	0	0
H7	0	0	0	1	0	0	0	0	0	0	0
C8	1	0	0	0	0	0	0	0	1	1	1
H9	0	0	0	0	0	0	0	1	0	0	0
H10	0	0	0	0	0	0	0	1	0	0	0
H11	0	0	0	0	0	0	0	1	0	0	0
Values of NNC_k	220	310	110	413	110	110	110	403	110	110	110
CW(Atom)	1.696	−0.345	1.696	−0.345	−0.140	−0.140	−0.140	−0.345	−0.140	−0.140	−0.140
$\mathrm{CW}(\mathrm{NNC}_k)$	0.237	−0.259	−0.193	−0.327	−0.193	−0.193	−0.193	1.132	−0.193	−0.193	−0.193
$\mathrm{CW}(\mathrm{a}_k) + \mathrm{CW}(\mathrm{NNC}_k)$	1.933	−0.604	1.503	−0.672	−0.333	−0.333	−0.333	0.787	−0.333	−0.333	−0.333
DCW	$\mathrm{DCW}(\mathrm{a}_k, \mathrm{NNC}_k) = \sum \mathrm{CW}(\mathrm{a}_k) + \sum \mathrm{CW}(\mathrm{NNC}_k) = 0.949$										

Table 2 Optimized molecular descriptor and observed and calculated ln *S* values of diverse functional aliphatic compounds

Sl. no.	Compound name	Molecular Descriptor (DCW)	Water solubility (ln *S*)		
			Obs.[a]	Calc.[b]	Res.
Training set					
1	n-Butanol	−0.214	0.006	−0.215	0.221
2	2-butanol	0.282	0.066	0.284	−0.218
3	3-methylbutanol	−1.273	−1.167	−1.281	0.114
4	2-pentanol	−0.972	−0.635	−0.978	0.343
5	3-methyl-2-butanol	−0.777	−0.405	−0.782	0.377
6	2-Methyl-2-butanol	0.254	0.339	0.256	0.083
7	2-Hexanol	−2.226	−1.995	−2.239	0.244
8	3-Methyl-3-pentanol	−1.000	−0.830	−1.006	0.176
9	2-Methyl-3-pentanol	−2.031	−1.609	−2.043	0.434
10	2,3-Dimethyl-2-butanol	−0.805	−0.851	−0.810	−0.041
11	3,3-Dimethyl-1-butanol	−2.232	−2.590	−2.245	−0.345
12	3,3-Dimethyl-2-butanol	−1.736	−1.410	−1.746	0.336
13	2-Ethylbutanol	−2.527	−2.787	−2.542	−0.245
14	n-Hepatnol	−3.976	−4.166	−4.00	−0.166
15	3-Methyl-3-hexanol	−2.254	−2.263	−2.268	0.005
16	3-Ethyl-3-pentanol	−2.254	−1.917	−2.268	0.351
17	2,3-Dimethyl-3-pentanol	−2.059	−1.937	−2.071	0.134
18	2,4-Dimethyl-3-pentanol	−3.090	−2.801	−3.109	0.308
19	4-Heptanol	−3.480	−3.196	−3.501	0.305
20	n-Octanol	−5.230	−5.401	−5.262	−0.139
21	2,2,3-Trimethyl-3-pentanol	−3.018	−2.931	−3.036	0.105
22	2-Octanol	−4.734	−4.755	−4.763	0.008
23	2-Ethylhexanol	−5.035	−4.996	−5.065	0.069
24	n-Nonanol	−6.484	−6.907	−6.523	−0.384
25	2-Nonanol	−5.988	−6.319	−6.024	−0.295
26	4-Nonanol	−5.988	−5.952	−6.024	0.072
27	3,5-Dimethyl-4-heptanol	−5.598	−5.298	−5.632	0.334
28	1,1-Diethyl-pentanol	−4.762	−5.572	−4.791	−0.781
29	7-Methyloctanol	−6.289	−5.744	−6.327	0.583
30	3,5,5-Trimethylhexanol	−5.799	−5.769	−5.834	0.065
31	n-Decanol	−7.738	−8.517	−7.785	−0.732
32	n-Tetradecanol	−12.754	−12.772	−12.831	0.059
33	n-Pentadecanol	−14.008	−13.796	−14.092	0.296
34	n-Hexanol	−15.262	−14.603	−15.354	0.751
35	2,2-Dimethylpropanol	−0.978	−0.889	−0.984	0.095
36	1-penten-3-ol	0.104	0.035	0.105	−0.070
37	2-Hexen-4-ol	−1.034	−0.939	−1.040	0.101
38	Methyl formate	1.303	1.015	1.311	−0.296
39	Propyl formate	−1.265	−1.133	−1.273	0.140
40	Butyl formate	−2.519	−2.303	−2.534	0.231
41	Isobutyl acetate	−2.678	−2.849	−2.694	−0.155
42	Butyl acetate	−2.873	−3.154	−2.890	−0.264
43	Isopentyl acetate	−3.932	−4.398	−3.956	−0.442
44	Pentyl acetate	−4.127	−4.283	−4.152	−0.131
45	Hexyl acetate	−5.381	−4.721	−5.413	0.692
46	Isopropyl propionate	−2.377	−2.970	−2.391	−0.579
47	Isopentyl propionate	−5.186	−5.088	−5.217	0.129
48	Isopropyl butyrate	−3.631	−4.465	−3.653	−0.812
49	Ethyl heptanoate	−6.635	−6.303	−6.675	0.372
50	Ethyl hexanoate	−5.381	−5.425	−5.413	−0.012
51	Ethyl octanoate	−7.889	−7.799	−7.937	0.138
52	Ethyl nanoate	−9.143	−8.741	−9.198	0.457
53	Ethyl decanoate	−10.397	−9.434	−10.460	1.026
54	Methyl sec-butyl ether	−1.817	−1.704	−1.828	0.124
55	Butyl methyl ether	−2.313	−2.303	−2.327	0.024
56	Dipropyl ether	−3.627	−3.364	−3.649	0.285
57	Dibutyl ether	−6.135	−6.261	−6.172	−0.089
58	Methyl t-butyl ether	−0.591	−0.484	−0.595	0.111
59	Ethyl propyl ether	−2.373	−1.531	−2.387	0.856
60	1,3-Dichloropropane	−3.736	−3.716	−3.759	0.043
61	Chloroform	−2.114	−2.118	−2.127	0.009
62	2-Bromopropane	−3.870	−3.756	−3.893	0.137
63	Isobutyl bromide	−5.425	−5.600	−5.458	−0.142
64	Isoamyl bromide	−6.679	−6.645	−6.719	0.074
65	Iodomethane	−2.364	−2.303	−2.378	0.075

Table 2 (Contd.)

Sl. no.	Compound name	Molecular Descriptor (DCW)	Water solubility (ln *S*)		
			Obs.[a]	Calc.[b]	Res.
66	Diiodomethane	−5.354	−5.388	−5.386	−0.002
67	Dichloroethsulfide	−5.456	−5.457	−5.489	0.032
68	n-Butane	−5.850	−6.020	−5.885	−0.135
69	n-Pentane	−7.104	−7.530	−7.147	−0.383
70	2,2-Dimethylpropane	−6.614	−7.198	−6.654	−0.544
71	2,4-Dimethylpentane	−9.222	−10.109	−9.278	−0.831
72	2,2,4-Trimethylpentane	−10.181	−9.501	−10.242	0.741
73	2,2,5-Trimethylhexane	−11.435	−11.624	−11.504	−0.120
74	Cyclohexane	−7.524	−7.322	−7.569	0.247
75	1,2-Dimethylcyclohexane	−9.642	−9.830	−9.700	−0.130
76	Cycloheptane	−8.778	−8.095	−8.831	0.736
77	n-Hexane	−8.358	−9.106	−8.408	−0.698
78	n-Octane	−10.866	−12.059	−10.931	−1.128
79	3-Methylpentane	−8.163	−8.819	−8.212	−0.607
80	1-Pentyne	−4.251	−3.707	−4.277	0.570
81	1-Heptyne	−6.759	−6.931	−6.800	−0.131
82	1-Nonanyne	−9.267	−9.694	−9.323	−0.371
83	1,8-Nonadiyne	−6.414	−6.862	−6.453	−0.409
84	1,6-Heptadiyne	−3.906	−4.030	−3.930	−0.100
85	2-Heptene	−8.420	−8.796	−8.471	−0.325
86	4-Methyl-1-pentene	−7.087	−7.460	−7.130	−0.330
87	1,5-Hexadiene	−6.206	−6.194	−6.243	0.049
88	1,4-Pentadiene	−4.952	−4.789	−4.982	0.193
89	Cyclopentene	−5.078	−4.835	−5.109	0.274
90	3-Methyl-2-butanone	−0.478	−0.286	−0.481	0.195
91	3-Hexanone	−1.927	−1.904	−1.939	0.035
92	3-Methyl-2-pentanone	−1.732	−1.545	−1.742	0.197
93	4-Methyl-2-pentanone	−1.732	−1.637	−1.742	0.105
94	4-Methyl-3-pentanone	−1.732	−1.870	−1.742	−0.128
95	4-Heptanone	−3.181	−3.325	−3.200	−0.125
96	5-Nonanone	−5.689	−5.929	−5.723	−0.206
Test set					
1	2-methylpropanol	−0.019	0.023	−0.019	0.042
2	n-Pentanol	−1.468	−1.347	−1.477	0.130
3	2-Methylbutanol	−1.273	−1.058	−1.281	0.223
4	3-pentanol	−0.972	−0.486	−0.978	0.492
5	n-Hexanol	−2.722	−2.790	−2.738	−0.052
6	3-Hexanol	−2.226	−1.832	−2.239	0.407
7	2-Methyl-2-pentanol	−1.000	−1.117	−1.006	−0.111
8	3-Methyl-2-pentanol	−2.031	−1.639	−2.043	0.404
9	4-Methylpentanol	−2.527	−2.282	−2.542	0.260
10	4-Methyl-2-pentanol	−2.031	−1.814	−2.043	0.229
11	Cyclohexanol	−1.392	−0.960	−1.400	0.440
12	2-Methyl-2-hexanol	−2.254	−2.473	−2.268	−0.205
13	2,3-Dimethyl-2-pentanol	−2.059	−2.002	−2.071	0.069
14	2,4-Dimethyl-2-pentanol	−2.059	−2.145	−2.071	−0.074
15	2,2-Dimethyl-3-pentanol	−2.990	−2.643	−3.008	0.365
16	3-Heptanol	−3.480	−3.194	−3.501	0.307
17	3-Nonanol	−5.988	−6.119	−6.024	−0.095
18	5-Nonanol	−5.988	−5.744	−6.024	0.280
19	2,6-Dimethyl-3-heptanol	−5.598	−5.776	−5.632	−0.144
20	4-Penten-1-ol	−0.392	−0.355	−0.394	0.039
21	3-Penten-2-ol	0.220	0.127	0.221	−0.094
22	1-Hexen-3-ol	−1.150	−1.354	−1.157	−0.197
23	2-Methyl-4-penten-3-ol	−0.955	−1.156	−0.961	−0.195
24	Ethyl formate	−0.011	0.174	−0.011	0.185
25	Ethyl formate	−0.011	−0.345	−0.011	−0.334
26	Propyl formate	−1.265	−1.174	−1.273	0.099
27	Butyl formate	−2.519	−2.733	−2.534	−0.199
28	1-Pentyl formate	−3.773	−3.500	−3.796	0.296
29	Methyl acetate	0.949	1.191	0.955	0.236
30	Methyl acetate	0.949	0.924	0.955	−0.031
31	Ethyl acetate	−0.365	−0.092	−0.367	0.275
32	Ethyl acetate	−0.365	−0.069	−0.367	0.298
33	Isopropyl acetate	−1.123	−1.194	−1.130	−0.064
34	Isopropyl acetate	−1.123	−1.245	−1.130	−0.115
35	Propyl acetate	−1.619	−1.704	−1.629	−0.075

Table 2 (Contd.)

Sl. no.	Compound name	Molecular Descriptor (DCW)	Water solubility (ln *S*)		
			Obs.[a]	Calc.[b]	Res.
36	Propyl acetate	−1.619	−1.726	−1.629	−0.097
37	Methyl propionate	−0.305	−0.345	−0.307	−0.038
38	Methyl propionate	−0.305	−0.390	−0.307	−0.083
39	Ethyl propionate	−1.619	−1.474	−1.629	0.155
40	Ethyl propionate	−1.619	−1.666	−1.629	−0.037
41	Propyl propionate	−2.873	−3.086	−2.890	−0.196
42	Propyl propionate	−2.873	−2.992	−2.890	−0.102
43	Butyl propionate	−4.127	−4.305	−4.152	−0.153
44	Pentyl propionate	−5.381	−5.181	−5.413	0.232
45	Methyl butyrate	−1.559	−1.945	−1.568	−0.377
46	Methyl butyrate	−1.559	−1.988	−1.568	−0.420
47	Ethyl butyrate	−2.873	−2.936	−2.890	−0.046
48	Propyl butyrate	−4.127	−4.423	−4.152	−0.271
49	Propyl butyrate	−4.127	−4.390	−4.152	−0.238
50	Ethyl valerate	−4.127	−4.069	−4.152	0.083
51	Dimethyl ether	1.509	1.772	1.518	0.254
52	Isopropyl methyl ether	−0.563	−0.138	−0.566	0.428
53	Isopropyl methyl ether	−0.563	−0.065	−0.566	0.501
54	Diethyl ether	−1.119	−0.550	−1.126	0.576
55	Diethyl ether	−1.119	−0.254	−1.126	0.872
56	Methyl propyl ether	−1.059	−0.620	−1.065	0.445
57	Methyl propyl ether	−1.059	−0.877	−1.065	0.188
58	Ethyl isopropyl ether	−1.877	−1.291	−1.888	0.597
59	Methyl isobutyl ether	−2.118	−2.071	−2.131	0.060
60	Isopropyl propyl ether	−3.131	−3.086	−3.150	0.064
61	Chloroethane	−2.912	−2.420	−2.930	0.510
62	Chloropropane	−4.166	−3.516	−4.191	0.675
63	2-Chloropropane	−3.670	−3.127	−3.692	0.565
64	Chlorobutane	−5.420	−4.934	−5.453	0.519
65	Isobutyl chloride	−5.225	−4.605	−5.256	0.651
66	Bromoethane	−3.112	−2.429	−3.131	0.702
67	Bromopropane	−4.366	−3.990	−4.392	0.402
68	Bromobutane	−5.620	−5.448	−5.654	0.206
69	1,3-Dibromopropane	−4.136	−4.792	−4.161	−0.631
70	Iodoethane	−3.678	−3.684	−3.700	0.016
71	Iodopropane	−4.932	−5.273	−4.962	−0.311
72	Iodobutane	−6.186	−6.816	−6.223	−0.593
73	Isobutane	−5.655	−5.867	−5.689	−0.178
74	2-Methylbutane	−6.909	−7.322	−6.951	−0.371
75	2,2-Dimethylbutane	−7.868	−8.45	−7.915	−0.535
76	Methylcyclohexane	−8.583	−8.867	−8.635	−0.232
77	Cyclooctane	−10.032	−9.560	−10.092	0.532
78	n-Heptane	−9.612	−10.438	−9.670	−0.768
79	2-Methylpentane	−8.163	−8.727	−8.212	−0.515
80	2,2-Dimethylpentane	−9.122	−8.450	−9.177	0.727
81	Cyclopentane	−6.270	−6.102	−6.308	0.206
82	Methylcyclopentane	−7.329	−7.599	−7.373	−0.226
83	1-Hexyne	−5.505	−5.434	−5.538	0.104
84	1-Octyne	−8.013	−8.427	−8.061	−0.366
85	1-Pentene	−6.028	−6.148	−6.064	−0.084
86	2-Pentene	−5.912	−5.849	−5.948	0.099
87	1-Hexene	−7.282	−7.437	−7.326	−0.111
88	1-Octene	−9.790	−10.638	−9.849	−0.789
89	1,6-Heptadiene	−7.460	−7.691	−7.505	−0.186
90	Cyclohexene	−6.332	−5.941	−6.370	0.429
91	Cycloheptene	−7.586	−7.276	−7.632	0.356
92	2-Butanone	0.581	1.561	0.584	0.977
93	2-Pentanone	−0.673	−0.389	−0.677	0.288
94	3-Pentanone	−0.673	−0.534	−0.677	0.143
95	2-Hexanone	−1.927	−1.794	−1.939	0.145
96	2-Heptanone	−3.181	−3.274	−3.200	−0.074
97	2,4-Dimethyl-3-pentanone	−2.791	−2.991	−2.808	−0.183

[a]From Ref. [8] and [53]
[b]From Eq. 4 (using optimized correlation weights listed in Table 3)

Table 3 Optimized correlation weights for different local invariants (obtained by the Monte Carlo optimization procedure)

Invariant type	local invariant	Optimized weight
a_k	H	−0.140
	C	−0.345
	O	1.696
	S	−3.501
	Cl	−1.193
	Br	−1.393
	I	−1.959
NNC_k	0100	1.550
	0110	−0.193
	0211	0.714
	0220	0.237
	0301	−1.243
	0310	−0.259
	0312	−0.180
	0320	4.019
	0321	0.020
	0401	2.722
	0402	−0.039
	0403	1.132
	0412	1.156
	0413	−0.327
	0421	1.736
	0422	−0.243
	0430	3.046
	0431	0.036
	0440	0.415

coefficient between the ln *S* values of the training set and the molecular descriptor [DCW]) were found. Based on the optimized correlation weights, the molecular descriptor was finally defined and this was then used to derive all the relations with ln *S* values of both the training and test sets using the least squares method of regression.

$$\ln S = \alpha + \beta * \mathrm{DCW}(a_k, LI_k) \quad (3)$$

The correlation weights were optimized using a PASCAL program developed by one of the authors (AAT). [54] Least squares linear regression analyses were performed using a GW-BASIC program *RRR98* developed by the other author (KR) [55]. The statistical quality of the equations [56] was judged by examining the parameters r_a^2 (adjusted r^2, i.e., explained variance), *r* (correlation coefficient), *F* (variance ratio) with *df* (degree of freedom), *s* (standard error of estimate) and *AVRES* (average of absolute values of residuals). The significance of the regression coefficients was judged by the corresponding standard errors and '*t*' test. A compound was considered as an outlier for a particular equation when the residual exceeded twice the standard error of estimate of the equation. Predicted residual sum of squares (PRESS) statistics were calculated for the training set by the "leave-one-out" (LOO) technique [57, 58] using the programs *KRPRES1* and *KRPRES2* [55] and q^2 (cross-validation r^2 or predicted variance) along with *SDEP* (standard deviation of error of predictions) values were reported. The predictive capacity of the model was determined by applying it to the test set and the value of r^2_{pred} was reported.

Results and discussion

The values of the optimized correlation weights of local invariants (a_k and NNC_k) are shown in Table 3. Based on the correlation weights as listed in Table 3, the molecular descriptors (DCW) were calculated for all the compounds as listed in Table 2. The calculation of the descriptor for methyl acetate is shown in Table 1.

The results of the relations of ln *S* values of different subsets of the training set with the molecular descriptor (DCW) are given in Table 4. It is observed that the descriptor could explain the variance of ln *S* values to the extent of 99.3% for alcohols ($n = 37$), 98.1% for esters ($n = 16$), 96.5% for ethers ($n = 6$), 99.7% for halocarbons ($n = 8$), 95.7% for hydrocarbons ($n = 22$) and 99.6% for ketones ($n = 7$). The average of the absolute values of the residuals is lowest for halocarbons (0.057) and highest for hydrocarbons (0.359). When all compounds of the training set ($n = 96$) were considered (Table 4), the following relation was obtained:

$$\ln S = 1.006 * \mathrm{DCW}(a, \mathrm{NNC}) \quad (4)$$

The insignificant intercept in Eq. 4 was set to zero.

Table 4 Relations of water solubility (ln *S*) of different subsets of the training set with the optimized molecular descriptor (DCW)[a]

Type of compound	Regression coefficient		Statistics		
	β (se)	α (se)	r_a^2 (*r*)	r^2 (*s*)	*F* (AVRES)
alcohols ($n = 37$)	0.993 (0.010)	–[b]	0.993 (0.996)	0.993 (0.320)	10187.5 (0.242)
esters ($n = 16$)	0.903 (0.032)	−0.435 (0.172)	0.981 (0.991)	0.983 (0.371)	794.2 (0.279)
ethers ($n = 6$)	0.962 (0.047)	–[b]	0.965 (0.982)	0.965 (0.381)	417.8 (0.240)
Halocarbons ($n = 8$)	1.002 (0.006)	–[b]	0.997 (0.999)	0.997 (0.084)	24231.4 (0.057)
Hydrocarbons ($n = 22$)	1.030 (0.013)	–[b]	0.957 (0.978)	0.957 (0.473)	6385.0 (0.359)
Ketones ($n = 7$)	1.083 (0.028)	0.191[c] (0.078)	0.996 (0.998)	0.997 (0.114)	1513.0 (0.076)
All[d] ($n = 96$)	1.006 (0.007)	–[b]	0.987 (0.994)	0.987 (0.380)	22062.6 (0.284)

[a] Model Equation: ln $S = \alpha + \beta$*DCW (a, NNC)
[b] Intercept set to zero
[c] Significant at 90% level
[d] Leave-one-out cross-validation statistics: $q^2 = 0.987$, SDEP = 0.386

Table 5 Relations of water solubility (ln S) of different subsets of the test set with the optimized molecular descriptor (DCW)[a]

Type of compound	Regression coefficient		Statistics		
	β (se)	α (se)	r_a^2 (r)	r^2 (s)	F (AVRES)
alcohols ($n=23$)	0.973 (0.018)	–[b]	0.981 (0.990)	0.981 (0.239)	2904.1 (0.201)
esters ($n=27$)	1.024 (0.016)	–[b]	0.986 (0.993)	0.986 (0.206)	4066.0 (0.164)
ethers ($n=10$)	1.047 (0.074)	0.445 (0.118)	0.957 (0.981)	0.962 (0.269)	199.8 (0.203)
Halocarbons ($n=12$)	0.966 (0.032)	–[b]	0.847 (0.920)	0.847 (0.511)	890.9 (0.430)
Hydrocarbons ($n=19$)	1.021 (0.013)	–[b]	0.928 (0.963)	0.928 (0.419)	6628.4 (0.327)
Ketones ($n=6$)	1.259 (0.065)	0.581 (0.128)	0.987 (0.995)	0.989 (0.210)	370.8 (0.150)
All[c] ($n=97$)	1.041 (0.013)	0.193 (0.054)	0.986 (0.993)	0.986 (0.342)	6715.7 (0.274)

[a]Model Equation: ln $S=\alpha+\beta$*DCW (a, NNC)
[b]Intercept set to zero
[c]Prediction statistics: $r^2_{pred}=0.988$

From Table 4, it can be observed that the above equation could predict and explain 98.7% of the variance of the ln S values of the training set. Out of 96 compounds, 1,1-diethylpentanol, isopropyl butyrate, ethyl decanoate, ethyl propyl ether, 2,4-dimethylpentane and n-octane acted as outliers in the case of modeling of all compounds (training set) with the molecular descriptor. Equation 4 was applied to the compounds of the training set and test set to calculate the ln S values as shown in Table 2.

The results of relations of ln S values of different subsets of the test set with the molecular descriptor (DCW) are given in Table 5. It is observed that the descriptor could explain the variance of ln S values to the extent of 98.1% for alcohols ($n=23$), 98.6% for esters ($n=27$), 95.7% for ethers ($n=10$), 84.7% for halocarbons ($n=12$), 92.8% for hydrocarbons ($n=19$) and 98.7% for ketones ($n=6$). The average of the absolute values of the residuals is highest for halocarbons (0.430) and lowest for ketones (0.150). When all compounds of the test set ($n=97$) were considered (Table 5), the molecular descriptor could explain 98.6% of the variance. Out of 97 compounds, diethyl ether, cyclooctane, 2,2-dimethylpentane and 2-butanone acted as outliers while modeling all compounds (test set) with the molecular descriptor. When Eq. 4 was used to predict the ln S values of the compounds of the test set (Table 2), the r^2_{pred} value was found to be 0.988 (Table 5).

The results of relations of ln S values of different subsets of the combined set with the molecular descriptor (DCW) are given in Table 6. It is observed that the descriptor could explain the variance of ln S values to the extent of 99.2% for alcohols ($n=60$), 98.5% for esters ($n=43$), 97.5% for ethers ($n=16$), 92.0% for halocarbons ($n=20$), 94.8% for hydrocarbons ($n=41$) and 98.8% for ketones ($n=13$). The average of the absolute values of the residuals was lowest for ketones (0.132) and highest for hydrocarbons (0.344). When all compounds of the combined sets ($n=193$) were considered (Table 6), the molecular descriptor could explain 98.7% of the variance. Out of 193 compounds, 1,1-diethylpentanol, n-hexanol, isopropyl butyrate, ethyl decanoate, ethyl propyl ether, 2,4-dimethylpentane, 2,2,4-trimethylpentane, cycloheptane, n-octane, diethyl ether, 2,2-dimethylpentane and 2-butanone acted as outliers in case of modeling of all compounds (combined set) with the molecular descriptor.

The same data set was modeled previously [46] using molecular connectivity ($^1\chi^v$), molecular negentropy and TAU indices. The statistical quality of the QSPR relation obtained in the present paper considering all the compounds ($n=193$) is better than the relations obtained previously [46].

The present analysis shows that the optimization of correlation weights scheme can generate statistically acceptable models for water solubility of diverse functional aliphatic compounds. Moreover, the scheme does

Table 6 Relations of water solubility (ln S) of different subsets of the combined set with the optimized molecular descriptor (DCW)[a]

Type of compound	Regression coefficient		Statistics		
	β (se)	α (se)	r_a^2 (r)	r^2 (s)	F (AVRES)
alcohols ($n=60$)	0.991 (0.008)	–[b]	0.992 (0.996)	0.992 (0.291)	14257.1 (0.229)
esters ($n=43$)	0.954 (0.018)	−0.160 (0.069)	0.985 (0.993)	0.985 (0.303)	2779.6 (0.229)
ethers ($n=16$)	1.066 (0.044)	0.435 (0.105)	0.975 (0.988)	0.977 (0.285)	588.4 (0.227)
Halocarbons ($n=20$)	0.981 (0.020)	–[b]	0.920 (0.959)	0.920 (0.401)	2519.7 (0.306)
Hydrocarbons ($n=41$)	1.026 (0.009)	–[b]	0.948 (0.974)	0.948 (0.445)	13114.7 (0.344)
Ketones ($n=13$)	1.157 (0.037)	0.398 (0.092)	0.988 (0.994)	0.989 (0.203)	961.3 (0.132)
All[c] ($n=193$)	1.025 (0.008)	0.121 (0.042)	0.987 (0.994)	0.987 (0.364)	14879.2 (0.277)

[a]Model Equation: ln $S=\alpha+\beta$* DCW (a, NNC)
[b]Intercept set to zero

not require complex calculation of diverse descriptors and statistical analysis for proper selection of descriptors and intercorrelation among them. Furthermore, as each 'elementary' molecular fragment has been provided with a 'personal' numerical local descriptor, one can identify vertices that increase/decrease the property under analysis. Thus, the scheme merits further assessment on exploring QSPR/QSAR of different physicochemical properties/biological activity data using different local invariants to justify its suitability in modeling studies. Furthermore, the present study shows the successful use of nearest neighboring codes as useful local invariants in the optimization of correlation weights scheme, which warrants extensive evaluation.

References

1. Harary F (1971) Graph theory. Addison-Wesley, Reading, MA
2. Balaban AT (ed) (1976) Chemical application of graph theory. Academic Press, London
3. Trinajstic N (1992) Chemical graph theory, 2nd edn. CRC Press, Boca Raton
4. Devillers J, Balaban AT (eds) (1999) Topological indices and related descriptors in QSAR and QSPR. Gordon and Breach Science Publishers, The Netherlands
5. Ivanciuc O (1998) Structural similarity measures for database searching. In: Schleyer PvR, Allinger NL, Clark T, Gasteiger J, Kollman PA, Schaefer HF, Schniener PR (eds) Encyclopedia of computational chemistry. Wiley, Chichester
6. Boncher D, Rouvray DH (eds) (1991) Chemical graph theory. Introduction and Fundamentals. Academic Press, New York
7. Balaban AT (ed) (1997) From chemical topology to three-dimensional geometry. NewYork
8. Kier LB, Hall LH (1976) Molecular connectivity in chemistry and drug research. Academic Press, New York
9. Kier LB, Hall LH (1986) Molecular connectivity in structure-activity analysis. Research Studies Press, Letchworth
10. Todeschini R, Consonni V (2000) Handbook of molecular descriptors. Wiley-VCH, Weinheim, Germany
11. Kier LB (1989) Quant Struct-Act Relat 8:218–223
12. Randic M (1991) J Comput Chem 12:970–980
13. Randic M (1991) Chemom Intell Lab Syst 10:213–227
14. Randic M (1991) J Chem Inf Comput Sci 31:311–320
15. Randic M (1992) J Chem Inf Comput Sci 32:686–692
16. Estrada E (1995) J Chem Inf Comput Sci 35:1022–1025
17. Amic D, Beslo D, Lucic D, Nikolic S, Trinajstic N (1998) J Chem Inf Comput Sci 38:819–822
18. Randic M, Basak SC (1999) J Chem Inf Comput Sci 39:261–266
19. Sinha DK, Basak SC, Mohanty RK, Basumallick IN (1999) Some aspects in mathematical chemistry. Visva-Bharati University Press, Santiniketan
20. Toropov AA, Toropova AP (1998) Russ J Coord Chem 24:81–85
21. Toropov AA, Toropova AP, Voropaeva NL, Ruban IN, Rashidova SS (1998) J Coord Chem 24:525–529
22. Toropov AA, Voropaeva NL, Ruban IN, Rashidova SS (1999) Polym Sci Ser A 41:975–985
23. Krenkel G, Castro EA, Toropov AA (2001) J Mol Struct (THEOCHEM) 542:107–113
24. Mercader A, Castro EA, Toropov AA (2001) J Mol Model 7:1–5
25. Mercader A, Castro EA, Toropov AA (2000) Chem Phys Lett 330:612–623
26. Krenkel G, Castro EA, Toropov AA (2001) J Mol Sci 2:57–65, http://www.mdpi.org/ijms
27. Marino DJG, Perruzo PJ, Castro EA, Toropov AA (2002) Internet Electron J Mol Des 1:115–133, http://www.biochempress.com
28. Duchowicz P, Castro EA, Toropov AA (2002) Computers and Chemistry 26:327–332
29. Toropov AA, Duchowicz P, Castro EA (2003) Int J Mol Sci 4:272–283, http://www.mdpi.org/ijms
30. Perruzo PJ, Marino DJG, Castro EA, Toropov AA (2003) Internet Electron J Mol Des 2:334–347, http://www.biochempress.com
31. Toropov AA, Schultz TW (2003) J Chem Inf Comput Sci 43:560–567
32. Toropov AA, Roy K (2004) J Chem Inf Comput Sci 44:179–186
33. Kubinyi H (1995) Quantitative structure-activity relationships. In: Wolff ME (ed) Burger's medicinal chemistry and drug discovery, 5th edn, vol 1. John Wiley New York, pp 497–571
34. Ghose AK, Crippen GM (1987) J Chem Inf Comput Sci 27:21–35
35. Ghose AK, Viswanadhan VN, Wendoloski JJ (1998) J Phys Chem 102:3762–3772
36. Bodor N, Gabanyi Z, Wong C-K (1989) J Am Chem Soc 111:3783–3786
37. Klopman G, Wang S (1991) J Comput Chem 12:1025–1032
38. Moriguchi I, Hirono S, Liu Q, Nakagome I, Matsushita Y (1992) Chem Pharm Bull (Tokyo) 40:127–130
39. Saxena AK (1995) Quant Struct-Act Relat 14:142–150
40. Benet LZ, Kroetz DL, Sheiner LB (1996) In: Hardman JG, Limbard LE, Molinoff PB, Ruddon RW, Goodman Gilman A (eds) Goodman and Gilman's The pharmacological basis of therapeutics. Mc-Graw Hill, New York, pp 3–27
41. Ferreira MM (2001) Chemosphere 44:125–146
42. Khadikar PV, Mandloi F, Bajaj AV, Joshi S (2003) Bioorg Med Chem Lett 13:419–422
43. Katritzky AR, Wang Y, Sild S, Tamm T (1998) J Chem Inf Comput Sci 38:720–725
44. Yin C, Liu X, Guo W, Lin T, Wang X, Wang L (2002) Water Res 36:2975–2982
45. Chen XQ, Cho SJ, Li Y, Venkatesh S (2002) J Pharm Sci 91:1838–1852
46. Roy K, Saha A (2003) Internet Electron J Mol Des 2:475–491, http://www.biochempress.com
47. Yang F, Wang ZD Huang YP (2004) J Comput Chem 25:881–887
48. Wanchana S, Yamashita F, Hashida M (2002) Pharmazie 57:127–129
49. Gao H, Shanmugasundaram V, Lee P (2002) Pharm Res 19:497–503
50. Puri S, Chickos JS, Welsh WJ (2003) J Chem Inf Comput Sci 43:55–62
51. Liu R, So SS (2001) J Chem Inf Comput Sci 41:1633–1639
52. Butina D, Gola JM (2003) 43:837–841
53. Hansch C, Quinlan JE, Lawrence GL (1968) J Org Chem 33:347–350
54. The program for optimization of correlation weights was developed in PASCAL by Toropov AA
55. The GW-BASIC programs *RRR98*, *KRPRES1* and *KRPRES2* were developed by Kunal Roy (1998) and standardized using known data sets.
56. Snedecor GW, Cochran WG (1967) Statistical Methods. Oxford& IBH Publishing Co. Pvt. Ltd., New Delhi, pp 381–418
57. Kier LB, Hall LH (1992) Atom description in QSAR models: development and use of an atom level index. In: Testa B (ed) Advances in drug research, vol 22. Academic Press, New York, pp 1–38
58. Wold S, Eriksson L (1995) Validation tools. In: van de Waterbeemd H (ed) Chemometric methods in molecular design. VCH, Weinheim, pp 309–318