Estimation of the Aqueous Solubility of Organic Molecules by the Group Contribution Approach

Gilles Klopman* and Hao Zhu

Department of Chemistry, Case Western Reserve University, 10900 Euclid Avenue, Cleveland, Ohio 44106

Received November 2, 2000

Several group contribution methods to estimate the aqueous solubility of organic molecules are proposed and evaluated for their ability to predict the water solubility of new molecules. The learning set consisted of 1168 organic compounds with experimental data taken from the literature after critical evaluation. The best method, based on a new fragment atom scheme, leads to a squared correlation coefficient of 0.95 and an average absolute calculation error of 0.50 log unit, which is superior to other group contribution methods currently available. One of the advantages of this model is that it has upper and lower limits so that the predicted solubilities cannot be unrealistily high or low.

INTRODUCTION

The aqueous solubility of a chemical is an important molecular property that can profoundly affect its biological activity. The knowledge of the aqueous solubility could be crucial to the design of drugs with the proper transport properties. Being able to predict the aqueous solubility of drug candidates would therefore greatly assist the drug development process. Indeed, if the solubility could be estimated before the compound is synthesized, one could avoid synthesizing many unsuitable ones. The question, then, is what method is best for predicting the aqueous solubility of complex organic molecules.

Several different methods have been proposed for the prediction of solubility. They generally consist of multiple linear regression analysis using various molecular descriptors. These approaches can be categorized as follow (Table 1).[1]

(1) One class of techniques calculates aqueous solubilities using experimental physicochemical properties[2] such as partition coefficient, melting points, boiling points, or molar volumes (derived from liquid density). These methods require that a sufficient quantity of purified compound and relevant experimental results be available. The methods are not applicable to compounds not yet synthesized or isolated and are generally used only for small sets of compounds. Therefore they have had only limited applications.

(2) Some methods are based on properties that cannot be determined experimentally but can be calculated from the molecular structure[3] (e.g. molecular surface area, molecular volume, etc.). The practical superiority of this approach over type 1 method is that it does not require the knowledge of any experimental data of the compound whose solubility is to be predicted because all the necessary information is calculated directly from the molecular structure. However, the relationship between this kind of properties and the aqueous solubility is not straightforward, as the resultant equations always requires additional complex correction terms.

(3) Group contribution methods are based on a compilation of relevant structural features of the molecules.[4] The group contribution methods are particularly suitable because, as the group 2 methods, they do not require the knowledge of any experimental data of the compound whose solubility is to be predicted. On the other hand, they require a large learning set and use a large number of parameters.

Group contribution methods are the most practical means of estimating aqueous solubilities. They allow the approximate calculation of solubility by calculating the contribution of relevant substructural units of the compounds. The following literature methods are examples of previous group contribution approaches aimed at the estimation of aqueous solubility: Klopman et al.,[4] Nirmalakhandan et al.,[5] Wakita et al.,[6] Suzuki,[7] Kuhne et al.,[8] and Lee et al.[9] (Table 2). Among these methods, only the Klopman model is a pure group contribution model without using additional experimental parameters. Using this model, the solubility of simple organic compounds could be calculated very well. But the lack of water solubility data at the time this model was developed resulted in the fact that the solubility of complex molecules cannot be calculated very accurately because they contain functional groups not previously available to the model. Moreover, it fails to account for the interactions between functional groups. This model is restricted by these two disadvantages. Indeed, when used to estimate the solubility of the current test set of 1168 chemicals, the R square value is found to be 0.77 and the F value is 44. It is thus clear that it needs to be improved.

DATA SOURCES

The learning set of aqueous solubility data used in this study was compiled from several literature sources.[3,4,10−13] Values used were those reported to have been obtained at 25 °C. The data were critically evaluated by the following

* Corresponding author phone: +1-216-368-3618; fax: +1-216-368-5921; e-mail: gxk6@po.cwru.edu.

10.1021/ci000152d CCC: $20.00
Published on Web 03/09/2001

Table 1. Examples of Literature Methods for Estimating Solubility[a]

	R^2	F	S	N	ref	P	compounds covered
experimental physicochemical properties dependent method	0.970	-	0.386	41	2	1	alcohol and aromatic compounds
calculated properties dependent method	0.959	1861	0.386	241	3	3	alkanes, alkenes, alkynes, alcohol, aromatic compounds, and PCBs
group contribution method	0.948	245.3	0.502	483	4	46	general simple organic compounds
the current model	0.951	171.8	0.50	1168		118	general organic compounds

[a] *R:* correlation coefficient; *F: F* statistic; *S:* standard error; *N:* the number of molecules used to calculate the regression; ref: references; *P*: number of parameters used in the model.

Table 2. Characterization of Literature Group Contribution Approach Methods for the Estimation of the Aqueous Solubility of Diverse Organic Chemicals

model	no. of parameters	exptl terms	no. of compds	ref
Klopman et al.	46	0	469	4
Nirmalakhandan	12	2	365	5
Wakita et al.	45	2	307	6
Suzuki	10	1	497	7
Kuhne et al.	58	2	694	8
Lee et al.	22	1	379	9
the current model	118	0	1168	-

criteria before being accepted. None of the compounds that exist as gases under the conditions of measurement were included because of the added complexity of the process. This eliminated all hydrocarbons with fewer than five carbons and several one and two carbon halogenated hydrocarbons. Salts and mixtures were excluded as well because our programs cannot handle them. The resulting data set consisted of 1168 organic chemicals that cover a great variety of chemical classes, including some complex drug-like compounds. The solubility unit used in this paper is log-(M/m^3).

The compounds were entered into our program using the SMILES code.[14] A program, called **MPAR**, was developed to identify the occurrences of each parameter in the compounds from their connectivity matrix generated from the SMILES code. A standard multiple regression analysis was then used to find which parameters correlate best with aqueous solubility. After a relationship is found, it is encoded into the program, which can then be used to calculate the aqueous solubility of new molecule simply by entering its SMILES code.

COMPUTATIONAL MODEL

In a group contribution model,[4−8] the aqueous solubility values are calculated from an equation such as

$$\log S = C_0 + \sum_{i=1}^{N} C_i G_i \qquad (1)$$

where logS is the logarithm of the solubility, C_0 is a constant characteristic of the solvent [C_0 can be viewed as being a characteristic of the molarity of the solvent. For example, in model 3, the calculated solubility of water is obtained by adding the value of C_0 to the C_1 value of one OH group. This translates into a logS value of 4.56, approximately equal to the molarity (4.74) of water in 1 m^3 water.], C_i is the number of occurrence of the *i*th group in a molecule, and G_i is the contribution coefficient of the *i*th group. C_0 and C_is were calculated by the MPAR program.

Table 3. Comparison between the Results Obtained for the Three Group Contribution Methods Based on the Same Training Set of Molecules

	no. of parameters	R square	SE	F value
model 1	171	0.953	0.49	109.8
model 2	71	0.905	0.68	56.6
model 3	118	0.951	0.50	171.8

In our work, the major goals were the following: (1) to improve the accuracy and scope of the previous models by increasing the size and diversity of the learning set (model 1); (2) to investigate the effect of using LogP as one of the parameters (model 2); and (3) to investigate possible methodology enhancement (model 3). The calculated results for these three models are shown in Table 3.

Model 1 was obtained by fitting the solubility values of the 1168 molecules with the structural descriptors as described in a previous model.[4] Because of the structural diversity of these molecules, we found it necessary to increase the number of descriptors from 46 in our previous model[4] to 171. This set of descriptors resulted in a new model (model 1) with improved accuracy (Table 3).

Model 2 was obtained by using predicted LogP values calculated from the KlogP program[15] as a parameter. The "LogP" parameter was chosen because it shows good correlation with relevant aqueous solubility and KlogP is a reliable LogP model.[15] However, it was necessary to use 70 parameters besides KlogP in order to build a reasonable solubility model (Table 3).

Classic group contribution models, such as models 1 and 2, suffer from the fact that boundaries are often not well handled. Indeed, when a molecule has many hydrophilic groups, such as in sugars, their contribution to solubility tend to decrease as their number increase. If this is not taken into account, then the calculation will yield unrealistically high predicted solubility values. Similarly, for chemicals with low solubility, such as long alkane chains, the calculated solubility value may become unrealistically low. The reason is that the linear model used in the group contribution approach cannot change the contribution of every additional instance of a group. This is shown in Table 4 where we compare the contribution of the last methylene groups in increasing aliphatic chain (Table 4).

The problem thus is to find a mathematical tool capable to keep the extreme calculated solubility values within reasonable boundaries. To solve this problem, we modified the classical methodology by creating a model based on logS* instead of logS. Our objective was to have logS* equal or similar to logS within the normal range of logS but able to assume much larger values than logS for extremely high or low solubility values. The procedure, which is based on

Table 4. Table 4. Comparison of the Contributions of Last Methylene Groups in Increasing Oliphatic Chain

name of compd	molecular structure	exptl logS (M/m^3)	contribution of the last $-CH_2-$ group
n-heptane	$CH_3(CH_2)_5CH_3$	−1.57	
n-octane	$CH_3(CH_2)_6CH_3$	−2.35	−0.78
n-nonane	$CH_3(CH_2)_7CH_3$	−3.00	−0.65
n-decane	$CH_3(CH_2)_8CH_3$	−3.64	−0.64
n-undecane	$CH_3(CH_2)_9CH_3$	−4.04	−0.4
n-dodecane	$CH_3(CH_2)_{10}CH_3$	−4.51	−0.47
n-tridecane	$CH_3(CH_2)_{11}CH_3$	−4.72	−0.21
n-tetradecane	$CH_3(CH_2)_{12}CH_3$	−4.95	−0.23
n-pentadecane	$CH_3(CH_2)_{13}CH_3$	−5.14	−0.19
n-hexadecane	$CH_3(CH_2)_{14}CH_3$	−5.27	−0.13
n-heptadecane	$CH_3(CH_2)_{15}CH_3$	−5.30	−0.03

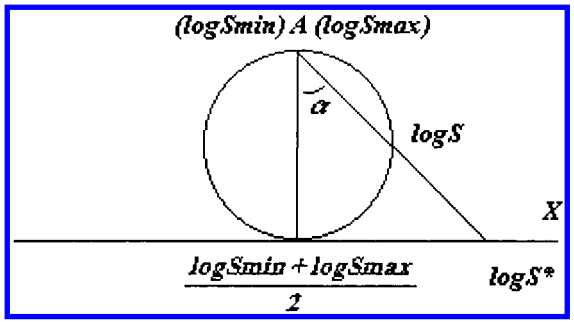

Figure 1. Mathematic model for the treatment of boundaries.

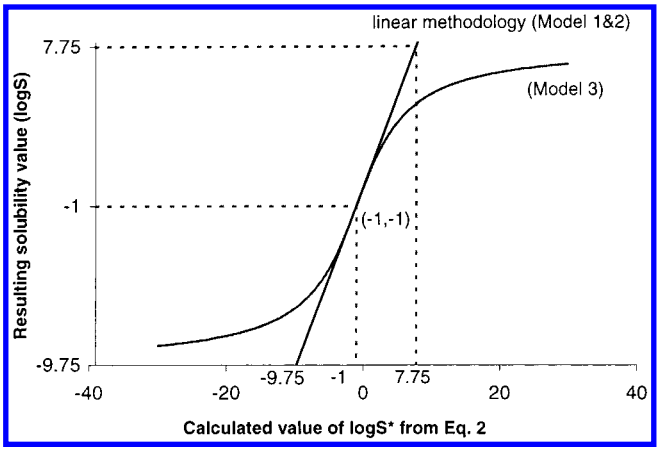

Figure 2. Relationship between the resulting solubility values and the value calculated from eq 1.

the theory of stereographic projection,[16] is as follows

$$\log S^* = f(\log S) = C_0 + \sum_{i=1}^{N} C_i G_i \qquad (2)$$

where logS* is a function of logS and is linearly correlated with parameters similar to those used in previous models. LogS is the experimental solubility value. The relationship between logS* and logS is defined as follows.

Let us draw a ***X*** axis and a circle whose circumference is equal to the value of (logSmax − logSmin), in which logSmin is the minimum acceptable solubility value of logS and logSmax is the maximum acceptable solubility value (Figure 1). The circle, which starts and also ends at point **A,** is marked from logSmin to logSmax in a counterclockwise direction and touches the ***X*** axis at the point corresponding to logS = (logSmin + logSmax)/2.

Let us mark an experimental logS value on the circle. The line joining A to logS on the circle intercepts the ***X*** axis at a point we call logS*. We can calculate logS* from logS using eq 3

$$\log S^* = m + \mathrm{SIGN}[\log S - m] \times [\mathrm{TAN}(\angle\alpha) \times (\log S_{max} - \log S_{min}) \div \pi] \qquad (3)$$

where logS is the experimental aqueous solubility, *m* is equal to (logSmin + logSmax)/2, SIGN takes the sign of the term in the square bracket, and TAN is the tangent trigonometric function. The value of the angle ∠α can be obtained from the position of the experimental logS value on the ring using eq 4

$$\angle\alpha = \pi \times \mathrm{ABS}[\log S - m] \div (\log S_{max} - \log S_{min}) \qquad (4)$$

where ABS gives the absolute value of the term inside the

Table 5. Basic Parameter Set of the Group Contribution Method[a]

no.	parameter	freq of use	coeff	*t* value
1	-CH3 or CH4	2000	−1.177	55.361
2	-CH2-	1717	−0.706	56.941
3	-CH<	411	−0.185	4.754
4	=CH2	50	−1.193	13.264
5	=CH- not in HC(=O)-	99	−0.382	6.528
6	=C< not in -C(=O)-,-C(=S)-	51	0.263	3.006
7	=C= in S=C=N-	3	−1.579	4.934
8	C$CH (triple bond)	12	−0.784	5.128
9	-CH2- in a ring	437	−0.672	30.253
10	-CH< in a ring	231	−0.373	9.243
11	>C< in a ring	173	0.354	7.553
12	=CH- in a ring	2523	−0.526	48.104
13	F- aliphatic	21	−1.305	14.523
14	F- aromatic	97	−0.571	14.733
15	Cl- aromatic	405	−1.617	78.585
16	Cl- aliphatic	260	−0.988	33.417
17	Br- aromatic	39	−2.125	37.599
18	Br- aliphatic	43	−1.124	17.118
19	I- aromatic	15	−2.408	16.276
20	I- aliphatic	11	−1.711	13.349
21	-CH2-OH	49	1.972	21.176
22	-CH -OH	59	1.739	14.723
23	Ph -OH	50	0.935	7.316
24	-OH other	30	2.587	16.254
25	-O- in a ring	33	−0.575	4.963
26	-O- not in ester	226	0.71	12.563
27	-CH=O	8	0.927	3.997
28	-C(=O)OH aromatic	27	−0.42	2.6
29	-C(=O)OH aliphatic	49	0.954	12.419
30	-C(=O)O- ester	122	0.441	6.349
31	-C(=O)O- ester, in a ring	6	−1.261	4.703
32	-C(=O)N<	120	0.506	11.034
33	-C(O)N=	4	−0.846	2.929
34	-C(=O)-	28	0.55	2.789
35	-C(=O)- in a ring	37	0.267	1.784
36	-S=O	8	0.819	4.019
37	-CH2-NH2	2	2.461	6.061
38	Ph-NH2	21	−0.178	1.355
39	-NH2 other	15	−0.227	1.37
40	-NH-	60	0.971	12.2
41	-N-	83	2.44	31.037
42	-N$C aromatic	6	−0.414	2.086
43	-N$C aliphatic	14	1.313	6.28
44	=NH or -N=	14	0.112	0.703
45	-N= in a ring	250	−0.38	10.199
46	-NO2 aromatic	48	−1.393	16.871
47	-SH	7	−0.679	3.463
48	-S-	69	−0.352	4.184
49	-S- cyclic	17	−0.92	6.209
50	-C(=S)-	11	−2.795	16.992
51	P=O	18	1.686	8.476
52	P=S	35	−0.318	1.744

[a] The value of constant C_0 is 5.0924.

Table 6. Extended Parameters of the Group Contribution Approach[a]

no.	parameter	freq of use	coeff	*t* value	no.	parameter	freq of use	coeff	*t* value
53(1)	cH -n =c -NH2	18	−1.518	11.182	81(3)	OH -C^H -C^H -NH -			
53(2)	n =c -OH				81(4)	OH -C^H -C^H -O -			
54	F- C -c =c - <2-F>	6	−0.352	2.237	81(5)	OH -C^H -C^ -OH			
55	N^ -C^H2-C^H2-N^ -	10	−0.872	5.518	82(1)	CH3-CO -CH2-	13	1.128	6.332
56	S -PS -O -CH3	14	−0.316	2.594	82(2)	CO -CH2-CH3			
57	NH -c -n =c -n =c -NH - <4-Cl >	4	−1.196	3.842	83	c =c -N -CO -CH2-	14	1.057	9.154
58(1)	cH =cH -c -O -PS -O -CH3 <5-O -CH3>	12	−1.456	11.625	84(1)	N^ -C^" -CH3	8	0.554	4.085
					84(2)	N^ =C^ -C^H2-			
58(2)	c -c -O -PS -O -CH3 <4-O -CH3>				84(3)	N^ =C^ -C -CH3			
58(3)	n =c -n =c -O -PS -O -CH3				84(4)	N^ =C^ -c =			
59(1)	Cl -C =CH - <2-Cl>	14	−1.029	6.325	85	O -CO -CH2-	7	2.14	9.407
59(2)	Cl -C =C - <2-Cl>				86	O -CH2-CH -	19	1.101	7.449
59(3)	Cl -C^ -C^ =C^ - <2-Cl>				87(1)	CH3-CO -O -CH2-	13	0.965	5.2
60(1)	Cl -c =c. -	25	−0.897	11.186	87(2)	CH3-CO -O -CH -			
60(2)	Cl -c =c -cH =c. -				87(3)	CH3-CO -O -CH3			
60(3)	Cl -c =c -c =cH -cH =cH - <4-Cl >				88	CO -CH -c =cH -cH =cH - <3-c ->	8	0.425	3.83
61(1)	O -CH2-CH2-	49	0.587	6.891	89	cH -c -CH2-CO -	8	−0.172	1.142
61(2)	O -CH2-C =				90	CO -O -c -cH =	21	−0.277	2.641
62	cH =cH -n =c. -cH =	7	2.413	9.827	91	C^O -C^H =C^ -	5	1.548	5.217
63(1)	CH -CH2-CH2-CH -	156	−0.175	4.994	92(1)	OH -C -CH3	91	0.235	4.014
63(2)	CH3-CH -CH -CH2-CH2-				92(2)	OH -CH -CH3			
63(3)	CH3-CH2-CH -CH -CH3				92(3)	OH -CH2-C -			
63(4)	CH2-CH2-CH -CH2-CH -				92(4)	OH -C -CH2-CH3			
63(5)	CH3-CH2-CH -CH2-CH -				92(5)	OH -CH -CH -CH3			
63(6)	CH3-CH2-CH -CH2-CH2-				92(6)	OH -CH -CH2-CH3			
63(7)	CH3-CH -CH2-CH2-CH2- <2-CH3>				92(7)	OH -CH2-CH -CH3			
63(8)	CH2-CH2-CH -CH2-CH2- <3-CH2->				92(8)	CH2-CH2-CH -CH2- <3-OH >			
63(9)	CH3-CH2-CH2-CH2-CH -CH3				92(9)	OH -CH -CH2-CH2-CH2-			
63(10)	CH2-CH2-CH -CH - <3-CH3>				93(1)	S -CH -CH3	28	0.126	1.633
64(1)	CH3-C^ -C^ -C^H2- <2-CH3>	21	−0.335	4.41	93(2)	S -CH2-CH2-			
64(2)	C^H2-C^H2-C^H -C^H2- <3-CH3>				93(3)	S^ -C^H2-C^H2-			
64(3)	CH2-CH2-C^H -C^H2-C^H2- <3-C^H2->				93(4)	S -CH2-CH2-CH3			
					94	PO -S -	9	0.519	2.724
64(4)	C^H2-C^H2-C^ -C^H2-				95	CH2-N -CH2- <2-CO ->	9	2.457	12.525
65(1)	cH =c. -cH =c. -cH =	28	−0.578	9.087	96	CH"-CH2-C^ -CH -	4	0.708	2.342
65(2)	c =cH -cH =c. -cH =c -				97	cH =c -c =cH - <2-OH >	11	−0.657	3.185
65(3)	cH =c. -cH =cH -cH =cH -c. =cH -				98	N -c =n -c -O -	4	−0.377	1.118
66(1)	N^H -c. =c. -	21	−1.311	12.761	99(1)	c =cH -cH =c -C^ -c =	5	0.355	2.263
66(2)	N^H -C^O -c. =				99(2)	cH =cH -c =c. -cH =c -			
66(3)	N^H -C^O -C^. -				100(1)	Cl -c =c -OH	9	1.092	5.452
67(1)	C^O -N^H -C^O -N^ -	4	−2.048	7.387	100(2)	Cl -c =c -CH2-			
67(2)	C^O -N^H -C^O -N^H -C^H =				101	C^O -C^ -CH2-CH2-	22	0.436	5.448
68	CO -C^ -C. - <3-OH>	9	−3.18	9.59	102	c -c =c - <2-OH>	4	−0.858	2.208
69(1)	N -c =cH -cH =c -	25	−0.294	2.604	103(1)	CH2-O -CH -	6	1.235	6.154
69(2)	N^ -c =cH -cH =c -				103(2)	CH2-O -C^H -			
69(3)	NH2-c =c -cH =c -				104	c. -C^H2-c. -	4	−1.001	3.636
69(4)	NH2-c =c. -				105	C^O -N^H -C^S -N^H -	4	0.937	4.322
69(5)	N -c =c -cH =c -c =				106	C^H -C^H2-C^H2-C.H -C.H -C^H2-C^H2-C. =	6	0.923	3.162
69(6)	cH =c -cH =c - <2-NH ->								
69(7)	NH2-c =c. -cH =				107	CH2-O -c =c -cH =c -cH = <6-Cl >	4	−0.737	2.306
70(1)	N^H -c. =n -	10	0.527	3.018					
70(2)	N -c =n -c =c -				108	NO2-c -cH =c -NO2	3	1.063	2.738
70(3)	N -c =n -c =n -				109(1)	c -O -CH -CH3 <3-CH3>	5	0.927	4.29
71	OH -CO -c =cH -cH = c - <4-cH=>	16	−0.336	2.953	109(2)	CH3-CH2-O -CH2-			
72	C^O -c. =cH -cH =cH -cH =c. −C^O -	6	−0.966	5.271	110	cH =cH -n =c. -c =	6	1.502	5.814
73(1)	OH -c =cH -cH =c -OH	9	−0.409	2.634	111	cH =cH -c −O -c -cH =c −CH -	16	−1.177	7.608
73(2)	OH -c =cH -cH =c -NH -				112(1)	CO -CH =CH -c =	4	−3.004	9.98
73(3)	OH -c =cH -cH =c -NH2				112(2)	CO -C^ =C^H -			
73(4)	OH -c =cH -cH =c -C" -				113	N^H -C^O -C^ -CH - CH2- <4-CH2->	12	0.391	3.034
74	cH =cH -cH =c -c =	169	0.159	4.484					
75	N -CH -CH3	14	0.468	5.049	114(1)	NH -CO -c =cH -	7	0.26	1.544
76	CH2-NH -CH2-CH2-	4	1.142	5.848	114(2)	CH3-NH -CO -CH2-			
77(1)	cH =c -c1 =cH -cH =c -c -	22	−0.346	4.857	115	I -c =c -	5	0.205	0.804
77(2)	cH =c -cH =c -c =c1 -c =				116(1)	O^ -C^ -CH2-	9	1.128	7.203
77(3)	cH =cH -c =cH -cH =c -c1 =cH -cH = <7-cH=>				116(2)	O^ -C^H -O^ -			
					116(3)	O^ -C^H2-C^H2-O^ -			
78(1)	cH =cH -cH =c -c1 =c -c = <6-Cl >	22	−0.316	3.826	117(1)	c. -c,	238	−0.702	28.037
					117(2)	c. -c. -			
78(2)	Cl -c =cH -c =c1 − <4-cH=>				117(3)	c,-c,			
78(3)	cH =cH -c =cH -c. =cH − <3-Cl >				118(1)	c -N -CO -NH -CH3	8	1.688	8.159
79	C^H =C^H -C^O -C^H =C. -	8	−0.818	2.765	118(2)	c -NH -CO -NH -SO2-			
80	CH2-CH2-O -c =	5	−0.607	2.118	118(3)	c -NH -CO -NH -C^H -			
81(1)	O -C^H -C^H -O -	11	−1.599	11.659	118(4)	c -NH -CO -N -CH2-			
81(2)	OH -C^H -C^H -OH								

[a] ^ element is included to one ring; . element is included to two rings; , element is included to three rings; c, o, n aromatic elements; "double bond attached.

square bracket, which is the length of the arc between logS and (logSmin + logSmax)/2 on the ring. Once the logS* values are calculated for all the chemicals, they can be used to identified the *C* parameters in eq 2 thus generating our model 3.

In the predictive model, the logS* value for a new molecule can be calculated from the relevant mutilinear relationship. The predicted solubility logS can then be obtained from logS* using eq 5

$$\log S = m + \mathrm{SIGN}[\log S^* - m] \times [\angle\alpha \div \pi \times (\log S_{max} - \log S_{min})] \quad (5)$$

where the term in the second square bracket calculates the length of the arc between (logSmin + logSmax)/2 and logS. The angle $\angle\alpha$ is calculated from logS* using the following equation

$$\angle\alpha = \mathrm{arctg}\{\mathrm{ABS}[\log S^* - m] \div [(\log S_{max} - \log S_{min}) \div \pi]\} \quad (6)$$

where arctg is the arctangent trigonometric function. With this procedure, a smooth curve is obtained between the maximum and minimum acceptable value (Figure 2) and whatever the value of logS*, the value of the intercept will be within the acceptable range of logS.

When a chemical is "miscible" with water, such as ethanol, it is difficult to distinguish which of the components is the solute. We therefore propose, somewhat arbitrarily, that when the number of moles of the solute is 1000 times that of the number of moles of water, the water becomes the solute. For this reason, we suggest that a log value of 7.75, which is 1000 times more than the number of moles of 1000 L of water, should be the upper limit of the solubility values of chemicals. The lower limit was set to −9.75 and is 2 log units lower than the solubility of the most insoluble compound we could identify, decachlorobiphenyl. Thus, the values of logSmax and logSmin were assigned the values of 7.75 and −9.75, respectively, in all equations above. We studied the variations produced by this and other selections and found that the proposed values produced the best results.

A comparison between the three models is made in Table 3. It is seen that model 3 is significantly better than the other two. The solubility of molecules that have extremely high or low aqueous solubility were not calculated very well in model 1. For this reason, model 1 needed many more parameters than model 3 so as to bring the solubility of extremely soluble and insoluble molecules into a reasonable range. Model 2 needed fewer parameters, but its accuracy (r square and standard error) is not very good. The solubility of many chemicals calculated with model 2 shows large errors, and we found no reasonable descriptor capable to improve the results. Furthermore, since the logP values were calculated from the KlogP program, which is also a group contribution model, it is obvious that the water solubility can as well be calculated directly from the molecular fragments.

RESULTS AND DISCUSSION

The quality of the final solubility model (model 3) is necessarily dependent on the set of group parameters used to create the model. First, a set of 52 basic group parameters

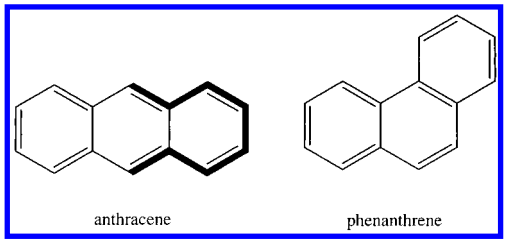

Figure 3. A parameter associated with polycyclic aromatic hydrocarbons.

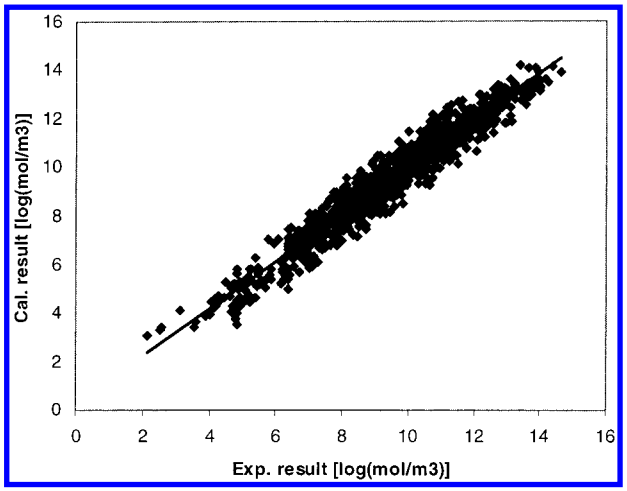

Figure 4. Correlation of predicted logS vs experimental logS values for 1168 chemicals.

derived from our previous work is used for our study (Table 5). This set of group parameters can be separated into two basic types of parameters, (1) general organic atoms (**C, N, O, P, S**) with their hybridization and the number of hydrogens attached to them and (2) basic functional groups (**OH, CHO, COOH, COO, $CONH_2$, CONH, COH, CON=, CO, NO, NO_2, PO, SO, NH_2, NH, CN, SH**).

There are a significant number of compounds that show unacceptable errors if only the basic 52 parameters are used. For this reason, 66 more group parameters were added in the final equation (Table 6). The basic rule used to find the most important extended parameters is to analyze the relationship between substructures (for example, the interactions between some groups) and their relevant effect on solubility. To do this analysis, we used the CASE methodology.[17] Basically the CASE methodology is an artificial intelligence system capable of identifying structural descriptors that may be associated with the properties of the molecules that are examined, such as a biological or physicochemical property. The CASE methodology is also suitable to help find the fragments that affect aqueous solubilities. The procedure used to do so has been described by Klopman et al.[15] The significance of each of the parameter is indicated by its *t* value.

The main goal of using extended parameters is to account for the interaction of functional groups. For example, 1,2-dinitrobenzene, 1,3-dinitrobenzene, and 1,4-dinitrobenzene have the same formula and functional groups. Using the old model,[4] the same values for the aqueous solubility are obtained for them. In fact, because of the combined electron attraction of the appropriately located nitro groups in 1,3-dinitrobenzene, the hydrogen that is ortho to both nitro groups and the two hydrogens that are ortho to one and para to the

Table 7. Prediction of the Solubility of a Test Set of 120 Chemicals

no.	name of compd	exp. logS	cal. logS (new)	cal. logS (old)	no.	name of compd	exp. logS	cal. logS (new)	cal. logS (old)
1	carbamic acid, ethyl ester	3.85	2.88	1.78	61	4-heptanone	1.7	0.48	1.77
2	benzamide	2.04	2.07	0.67	62	2-butenal	3.32	3.09	3.49
3	glycine	3.52	4.52	2.58	63	1-butanol, 3-methyl-, acetate	1.08	1.73	0.68
4	L-serine	2.98	4.29	3.47	64	1-naphthalenamine	1.08	−0.04	−0.09
5	L-glutamine	2.45	3.26	1.2	65	2-naphthalenol	0.72	0.91	0.48
6	benz a anthracene, 7,12-dimethyl-	−4.02	−4.76	−5.6	66	D-glucopyranoside, 2-(hydroxymethyl) phenyl	2.15	3.26	5.3
7	lindane	−1.59	−3.04	−1.88					
8	L-leucine	2.2	2.09	1.13	67	2-propenoic acid, 3-phenyl-	0.52	−0.4	−0.79
9	L-methionine	2.58	2.38	0.72	68	2-propenoic acid, ethyl ester	2.26	1.76	1.17
10	L-phenylalanine	2.08	2	−0.23	69	4-pyrimidinone,2,3-dihydro-2-thioxo-	0.74	1.67	1.63
11	L-valine	2.7	2.54	1.65	70	acetic acid, hexyl ester	0.54	1.21	−0.16
2	endrin	−3.29	−3.04	−3.61	71	mercaptobenzothiazole, 2-	−0.15	0.97	−0.54
13	L-tryptophan	1.72	1.16	−1.76	72	benzoic acid, 4-amino-	2.6	1.48	0.4
14	L-isoleucine	2.41	2.09	1.13	73	acenaphthylene	−0.96	−1.27	−1.67
15	4-chlorobenzoic acid	−0.31	0.16	−0.85	74	dibenzo-p-dioxin	−2.31	−1.63	−0.78
16	L-arginine	3	3.42	1.25	75	1,1-ethanediol, 2,2,2-trichloro-	3.72	3.63	3.75
17	codeine	1.48	0.47	−0.76	76	DL-alanine	3.26	3.31	2.38
18	1,2,3-propanetricarboxylic acid, 2-hydroxy-	3.51	4.77	−0.02	77	decanoic acid	−0.44	−0.7	−2.09
19	2-propenamide	3.95	2.76	2.01	78	2-propanol, 1,1,1-trifluoro-	3.3	3.03	2.27
20	2-propenoic acid, 2-methyl-	3	2.83	0.7	79	guanidine, cyano-	2.69	4.02	4.58
21	2-propenoic acid, 2-methyl-, methyl ester	2.2	1.83	1.08	80	5-nonanone	0.41	−0.8	0.73
22	1,2-benzisothiazol-3(2H)-one, 1,1-dioxide	1.36	0.84	−1.5	81	1,2-dinitrobenzene	−0.1	0.45	−3.37
23	1-naphthalenesulfonic acid, 2-amino-	1.3	0.98	0.91	82	2,3-dichloro-2-methyl-butane	0.31	−0.78	0.75
24	9,10-anthracenedione	−2.19	−2.87	−1.67	83	1,2-diiodoethylene	−0.22	0.72	0.02
25	1,2-benzenedicarboxylic acid, butyl phenylmethyl ester	−2.64	−1.77	−4.88	84	3-methyl-3-hexanol	2	2.12	1.94
					85	ethane, 1,2-diethoxy-	2.23	3.16	1.7
26	9H-carbazole	−2.27	−1.62	−2.04	86	4-methylpentanol	1.86	1.84	2.22
27	benzenamine, 2-nitro-	1.04	1.57	−0.56	87	1-phenylethanol	2.08	2.44	1.69
28	1,2-benzenedicarboxylic acid	0.89	1.98	−1.46	88	1-hexen-3-one	2.17	1.27	2.06
29	phenol, 2-methoxy-	1.04	2.49	2.25	89	1,2,3,6,7,8-hexahydropyrene	−2.96	−4.25	−4.42
30	1-naphthalenol	0.78	1.41	0.48	90	dicamba	1.3	−0.17	−1.84
31	1,2-dicyanobenzene	0.62	2.03	3.03	91	dodine acetate	0.37	−2.34	−2.39
32	benzenamine, *N*,*N*-diethyl-	−0.03	1.02	0.42	92	biphenyl, 3,4-dichloro-	−4.44	−3.3	−2.93
33	biphenyl	−1.3	−0.24	−1.28	93	asulam	1.34	1.79	−1.76
34	1,1′-biphenyl −4-ol	−0.48	0.47	−0.28	94	O-*tert*-butyl carbamate	3.1	1.66	1.31
35	10H-phenothiazine	−2.1	−3.17	−1.51	95	3-methyl-3-heptanol	1.4	1.62	1.42
36	1,1′-biphenyl −4,4′-diamine	0.3	−0.84	−0.44	96	2,4′,5-PCB	−3.25	−3.33	−3.76
37	1,2-benzenediamine	2.58	2.5	2.26	97	2,3-dimethyl-1-butanol	2.61	2.17	2.53
38	2-propanol, 1,3-dichloro-	2.89	2.59	2.31	98	ditolyl ether	−1.85	−0.84	−2.06
39	2-propenoic acid, methyl ester	2.78	2.24	1.69	99	3-methyl-2-heptanol	1.28	0.55	1.49
40	2-imidazolidinethione	2.29	2.36	3.88	100	2′,3,4,4′,5′-PCB	−4.39	−5.25	−5.41
41	2-furancarboxaldehyde	2.9	3	3.39	101	2,3′,4′,5-PCB	−4.25	−4.63	−4.58
42	benzene, 1,3,5-trinitro-	0.11	1.23	−5.77	102	dichlorodibenzo-p-dioxin, 2,7-	−4.82	−3.59	−2.43
43	1,2,3-propanetriol, triacetate	2.4	3.73	−0.23	103	2,3,4,2′,4′,5′-PCB	−5.32	−5.62	−6.23
44	diazene, diphenyl-	0.25	0.1	−2.88	104	2,2′,3,3′,4,4′,5,5′-PCB	−6.16	−6.33	−7.88
45	acetamide, *N*-phenyl-	1.67	1.11	0.25	105	2,3,4,2′,5′-PCB	−4.91	−5.07	−5.41
46	diethylthiourea, *N*,*N*′-	1.54	0.44	3.05	106	2,3,3′,4,4′,5-PCB	−4.82	−5.76	−6.23
47	2-propenoic acid, 2-methylpropyl ester	1.79	1.13	0.45	107	2,3,4′-PCB	−3.26	−3.35	−3.76
48	ethanesulfonic acid, 2-amino-	2.91	4.07	4.49	108	2-chlorodibenzo-p-dioxin	−2.82	−2.66	−1.61
49	2-pentanone, 4-methyl-	2.26	1.95	2.6	109	2,2′,3,3′,4,4′,5,5′,6-nonachlorobiphenyl	−7.26	−6.65	−8.71
50	2-pentene	0.46	1.14	1.6	110	2,2′,3,5′-PCB	−3.47	−3.92	−4.58
51	butanedioic acid	2.8	3.65	0.2	111	2,2′,3,5,5′,6-hexachlorobiphenyl	−4.42	−5.45	−6.23
52	2,4-hexadienoic acid, (E,E)-	1.23	2.5	0.53	112	2,2′,3,4,4′,5′,6-heptachlorobiphenyl	−4.92	−5.93	−7.06
53	2-propanol, 1,1′-iminobis-	3.81	3.91	4.4	113	2,2′,3,3′,4,5,5′,6,6′-PCB	−7.41	−6.65	−8.71
54	endosulfan	−3.15	−2.25	−2.82	114	2,2′,3,4,5,5′-hexachlorobiphenyl	−4.68	−5.45	−6.23
55	anthranilic acid, o-	1.48	2.25	0.4	115	2,2′,3,4,5,5′,6-heptachlorobiphenyl	−5.94	−5.93	−7.06
56	2-naphthalenesulfonic acid, 5-amino-	0.65	0.47	0.91	116	2,2′,3,4,6-PCB	−4.43	−4.87	−5.41
57	dinitrotoluene, 2,4-	0.18	0.15	−3.99	117	2,3,4,5,2′,3′-PCB	−5.78	−5.45	−6.23
58	hydrazine, 1,2-diphenyl-	0.08	1.43	0.28	118	2,3,6-PCB	−3.29	−3.86	−3.76
59	benzaldehyde, 4-hydroxy-	2.04	3.48	3.17	119	2,2′,4,6,6′-PCB	−4.32	−5.55	−5.41
60	benzaldehyde, 4-methoxy-	1.51	2.73	2.01	120	2,3,3′,4,4′,6-hexachlorobiphenyl	−4.66	−5.61	−6.23

[a] The calculation results were obtained from current model (new) and previous Klopman group contribution model (old).[4]

other are highly acidic making 1,3-dinitrobenzene much more soluble than the other two isomers. Thus the presence of a group consisting of two nitro groups in *meta*- in an aromatic ring should be used as a parameter.

Some other situations are more complex. For example, the solubility of anthracene is 10 orders of magnitude less than its isomer, phenanthrene. This difference is also seen in their melting points. Indeed, the melting point of phenanthrene (99.5 °C) is much lower than that of anthracene (217.5 °C). The difference in melting point is clearly related to the difference in crystallinities. Similar solubility differences were also found in a number of other isomeric Polycyclic

Aromatic Hydrocarbons (PAH). We used the CASE program to help resolve which molecular fragment would help us assign the difference to. We found that the bolded fragment in Figure 3 appropriately resolves the problem.

All the parameters found to be relevant are listed in Table 6. Each parameter is generated after analysis of a series of compounds with similar molecular structure. Some parameters have similar structures and their coefficients are quite close. Such parameters are grouped together. For example, all the fragments of parameter 64 are molecular fragments of polycyclic compounds that affect the solubility, and their coefficients are all defined as −0.175 after analysis.

Using the new extended parameter set to calculate the solubility values of the 1168 chemicals, we obtained a linear relationship between the experimental and calculated values with the following statistical characteristics: (1) multiple linear regression between logS* obtained from experimental value and logS* calculated ($\boldsymbol{R}^2 = 0.86$, $\boldsymbol{F} = 63.6$, $\boldsymbol{s} = 0.61$, $\boldsymbol{N} = 1168$) and (2) linear relationship between experimental and calculated logS values ($\boldsymbol{R}^2 = 0.95$, $\boldsymbol{s} = 0.50$, $\boldsymbol{N} = 1168$).

The clear improvement from calculated logS* to logS indicates the importance of the addition of boundary restrictions.

After this work was finished, we started searching for additional water solubility data for molecules that were not presents in our set. We found a total of 120 compounds from a number of references.[8,18−20] This set was then used as an unbiased test of the accuracy of our new model. The results were then compared with those obtained from the previous model,[4] as shown in Table 7. The standard deviation was found to be 0.79 log unit with our new model, and only one compound (dodine acetate) showed a large error (−2.71 log unit) between the estimated and observed water solubility. But with the previous model, the standard deviation was found to be twice as large, i.e., 1.43. The predicted values of the water solubility of PCBs, most of which are mostly insoluble, were particularly bad. It is clear that our new model is significantly better than the previous one.

To use additional parameters in our model resulted in more precise results but lower *F* statistic values. It indicates that the additional parameters were not reliable. The occurrence of these parameters was very low, and it is probable that they are erroneous. Nevertheless, as more experimental data becomes available, one should be able to improve the model some more. One must however also remain aware that, as with any group contribution model, the calculation of the solubility of new compounds, which contain fragments not encountered in the model, may yield unanticipated errors. Nevertheless, the parameters defined in this model do cover a great many classes of organic compounds, and therefore the model should be quite reliable for the calculation of the water solubility of most new molecules.

CONCLUSION

A modified group contribution approach was used to correlate the aqueous solubility of 1168 organic compounds. The model was based solely on information derived from molecular formula. No experimental values were used to generate any of the parameters in the model. Key to this effort was the attempt to improve the methodology by establishing boundaries as well as to cover the conformational influences of functional groups. Compared with previous models, the current equation can adequately cover many more types of chemicals and even very complex compounds. It is shown that, as with other calculations of the solvent effect, the influence of interactions between functional groups cannot be ignored. After inclusion of our boundary treatment model, the results were found to be more reliable for the very soluble and insoluble chemicals than with previous models. It should be emphasized that the set of parameters could still be extended if one finds it necessary to include additional groups to accommodate uncommon interactions in molecules.

REFERENCES AND NOTES

(1) Yalkowsky, S. H.; Banerjee, S. Aqueous Solubility Methods of Estimation for Organic Compounds; Marcel Dekker: New York, 1992; p 41.
(2) Yalkowsky, S. H. Estimation of the aqueous solubility of complex organic compounds. *Chemosphere* **1993**, *26*, 1239−1261.
(3) Huibers, P. T.; Katritzky A. R. Correlation of the aqueous solubility of hydrocarbons and halogenated hydrocarbons with molecular structure. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 283−292.
(4) Klopman, G.; Wang, S.; Balthasar, D. M. Estimation of aqueous solubility of organic molecules by the group contribution approach. Application to the study of biodegradation. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 474−482.
(5) Nirmalakhandan, N. N.; Speece, R. E. Prediction of aqueous solubility of organic chemicals based on molecular structure. 2. Application to PNAs, PCBs, PCDDs, etc. *Environ. Sci. Technol.* **1989**, *23*, 708−713.
(6) Wakita, K.; Yoshimoto, M.; Miyamoto, S.; Watanabe, H. A method for calculation of the aqueous solubility of organic compounds by using new fragment solubility constant. *Chem. Pharm. Bull. (Tokyo)* **1986**, 34, 4663−4681.
(7) Suzuki, T. Development of an automatic estimation system for both the partition coefficient and aqueous solubility. *J. Comput.-Aided Mol. Design* **1991**, 5, 149−166.
(8) Kühne, R.; Ebert, R.-U.; Kleint, F.; Schmidt, G.; Schüürmann, G. Group Contribution Method to Estimate Water Solubility of Organic Chemicals. *Chemosphere* **1995**, *30*, 2061−2077.
(9) Lee, Y.; Myrdal, P. B.; Yalkowsky, S. H. Aqueous Functional Group Activity Coefficients (AQUAFAC) 4: Applications to Complex Organic Compounds. *Chemosphere* **1996**, *33*, 2129−2144.
(10) Cao, C.; Li, Z. Molecular Polarizability. 1. Relationship to Water Solubility of Alkanes and Alcohols. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 1−7.
(11) Prankerd, R. J.; Mckeown, R. H. Physico-Chemical Properties of Barbituric Acid Derivatives: IV. Solubilities of 5,5-disubstituted Barbituric Acids in Water. *Int. J. Pharm.* **1994**, *112*, 1−15.
(12) Fini, A.; Fazio, G.; Feroci, G. Solubility and Solubilization Properties of Non-Steroidal Antiinflammatory Drugs. *Int. J. Pharm*. **1995**, *126*, 95−102.
(13) Wauchope, R. D.; Buttler, T. M.; Hornsby, A. G.; Augustijin-Beckers, P. W. M.; Burt, J. P. The SCS/ARS/CES Pesticide Properties Database for Environmental Decision-Making. *Rev. Environ. Contamination Toxicol*. **1992**, *123*, 1−36.
(14) Weininger, D. Smiles, a Chemical Language and Information-System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.
(15) Klopman, G.; Li, J.; Wang, S.; Dimayuga, M. Computer Automated logP Caculations Based on an Extended Contribution Approach. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 752−781.
(16) Burkhardt, H. Theory of Functions of a Complex Variable; Rasor, S. E. (Translation); D. C. Heath & CO: Boston, New York, Chicago, 1913; p 48.
(17) Klopman, G. Artificial Intelligence Approach to Structure−Activity Studies. Computer Automated Structure Evaluation of Biological Activity of Organic Molecules. *J. Am. Chem. Soc.* **1984**, *106*, 7315.
(18) Herman, R. A. Quantitative Structure-Pharmacokinetic Relationships for Systemic Drug Distribution Kinetics not Confined to a Congeneric Series. *J. Pharm. Sci.* **1994**, *83*, 423−428.
(19) Miithani, S. D.; Bakatselou, V.; tenHoor, C. N.; Dressman, J. B. Estimation of the Increase in Solubility of Drugs as a Function of Bile Salt Concentration. *Pharm. Res*. **1996**, *13*, 163−167.
(20) Schwarzenbach, R.; Gschwend, P. M.; Imboden, D. M. *Environmental Organic Chemistry*; J. Wiley: New York, 1993; p 107.

CI000152D