Estimation of Aqueous Solubility of Organic Molecules by the Group Contribution Approach. Application to the Study of Biodegradation

Gilles Klopman,* Shaomeng Wang, and D. M. Balthasar

Department of Chemistry, Case Western Reserve University, Cleveland, Ohio 44106-7078

Received March 2, 1992

A reliable and generally applicable aqueous solubility estimation method for organic compounds based on a group contribution approach has been developed. Two models have been established based on two different sets of parameters. One has a higher accuracy, while the other has a more general applicability. The prediction potentials of these two models have been evaluated through cross-validation experiments. For model I, the mean cross-validated r^2 and SD for 10 such cross-validation experiments were 0.946 and 0.503 log units, respectively. While for model II, they were 0.953 and 0.546 log units, respectively. Applying our models to estimate the water solubility values for the compounds in an independent test set, we found that model I can be applied to 13 out of 21 compounds with a SD equal to 0.58 log unit and model II can be applied to all the 21 compounds with a SD equal to 1.25 log units. Our models compare favorably to all the current available water estimation methods. A program based on this approach has been written in FORTRAN77 and is currently running on a VAX/VMS system. The program can be applied to estimate the water solubility of any organic chemical with a good or fairly good accuracy except for electrolytes. Applying our aqueous solubility estimation models to biodegradation studies, we found that although the water solubility was not the sole factor controlling the rate of biodegradation, ring compounds with greater solubilities were more likely to biodegrade at a faster rate. The significance of the relationship between water solubility and biodegradation activity has been illustrated by predicting the biodegradation activity of 27 new chemicals based solely on their estimated solubility values.

INTRODUCTION

The aqueous solubility of a drug is a key factor in determining its biological activity. Before an orally administered drug can become available to its receptor, it must dissolve in the GI fluid. Both the dissolution rate and the maximum amount of drug that can be dissolved are governed by the solubility of the drug in the medium.[1] The design of orally active drugs must account for the effects of the structural modification on solubility. The lack of sufficient aqueous solubility often causes a drug to appear inactive. The aqueous solubility may also be a factor that controls the rate of biodegradation[2–4] and bioaccumulation[5] processes. In our drug design and QSAR studies, we felt the need to have access to a fast and reliable water solubility estimation approach in order to correctly predict the biological activity of a proposed chemical. Recently, in our structure–biodegradation–activity study, we noticed that there was some correlation between water solubility and biodegradation activity for organic compounds.[6] However, due to the lack of water solubility data for some important compounds in our database, we were not able to evaluate the correlation for the entire database. Hence, it was necessary to develop an approach for the estimation of water solubility for organic compounds in order to assist our theoretical drug design and QSAR studies and also to address some interesting problems in our biodegradation study.

The aqueous solubility of a chemical is governed by three major factors: (1) the entropy of mixing; (2) the differences between the solute–water adhesive interaction and the sum of the solute–solute and water–water adhesive interactions; and (3) the additional solute–solute interactions associated with the lattice energy of crystalline solutes, which are applicable to solids but not to liquids.[7] In cases where some water molecules are entering into the solute phase, a partitioning process occurs instead of a simple solvation. A fundamental approach which will precisely calculate the water solubility of a chemical has to include calculations for all the above factors. Thus, for a simple solvation, calculation of factors 1 and 2 is required for liquids, while for solids, factor 3 also has to be considered. The task for such precise calculations is very difficult, time-consuming, and impractical at the present time.

Some approximations may be possible in the calculations without sacrificing too much accuracy in certain circumstances. Yalkowsky and Valvani[7] made approximations relating to factor 3 in their solubility study and linked factors 1 and 2 to the partition coefficient between *n*-octanol and water (log *P*). They were able to derive an equation relating aqueous solubility with the octanol–water partition coefficient, the melting point, and the entropy of fusion. This approach took into account the effects of crystallinity upon solubility through the use of the melting point and the entropy of fusion. Good results were obtained for a few of the classes of organic compounds which were examined. The applications of this approach are limited because it utilizes the melting point and log *P* data in the calculation of the water solubility for a chemical. In cases where the proposed chemical is not available, its melting point and log *P* value will have to be estimated through some calculation approaches. Although, in the past, a few reliable log *P* estimation methods have been developed, the estimation of the melting point of a chemical still remains a challenging problem. Thus, many investigators have taken the approach of using the partition coefficient between *n*-octanol and water (log *P*) alone for the estimation of the water solubility.[8–15] Good regression equations were found when the water solubility within a single class of compounds was studied. Even though one can obtain regression equations with mixed classes of compounds, much poorer results are usually obtained.[8] A solvatochromic approach has also been used by Kamlet et al.[16,17] to derive a fundamental, high-quality correlation for aqueous solubility. Recently, Nirmalakhandan and Speece[18–20] developed an

* All correspondence should be sent to this author.

0095-2338/92/1632-0474$03.00/0

Table I. Contribution Values to log S of Basic Group Set in Model I[a]

	parameter	no. of compds	frequency of use	contribution	remarks
1	$-CH_3$	249	449	-0.3361	
2	$-CH_2-$	200	565	-0.5729	
3	$-CH-(-)$	81	95	-0.6057	
4	$-C-(-)(-)$	30	35	-0.7853	
5	$=CH_2$	15	17	-0.6870	
6	$=CH-$	30	41	-0.3230	
7	$=C-(-)$	83	104	-0.3345	
8	$=C=$	0	0	undetermined	
9	$-C\equiv CH$	6	7	-0.6013	
10	$\equiv C-$	0	0	undetermined	not $-C\equiv N$, not in $-C\equiv CH$
11	$-C^*H_2-$	18	41	-0.4568	
12	$-C^*H-(-)$	10	40	-0.4072	
13	$-C^*-(-)(-)$	11	34	-0.3122	
14	$=C^*H-$	227	1180	-0.3690	
15	$=C^*-(-)$	234	758	-0.4944	
16	$=C^*=$	0	0	undetermined	
17	$\equiv C^*-$	0	0	undetermined	
18	-F	13	41	-0.4472	connecting to sp^3 carbon
19	-F	6	9	-0.1773	connecting to other atom
20	-Cl	59	156	-0.4293	connecting to sp^3 carbon
21	-Cl	92	266	-0.6318	connecting to other atom
22	-Br	25	38	-0.6321	connecting to sp^3 carbon
23	-Br	17	33	-0.9643	connecting to other atom
24	-I	6	7	-1.2391	connecting to sp^3 carbon
25	-I	4	4	-1.2597	connecting to other atom
26	-OH	27	27	1.4642	primary alcohol
27	-OH	19	19	1.5629	secondary alcohol
28	-OH	16	16	1.0885	tertiary alcohol
29	-OH	23	26	1.1919	connecting to a non-sp^3 carbon not in COOH
30	-OH	0	0	undetermined	connecting to a nitrogen
31	-OH	0	0	undetermined	connecting to an oxygen
32	-OH	0	0	undetermined	connecting to a phosphorus
33	-OH	0	0	undetermined	connecting to a sulfur
34	$-O^*-$	7	8	-0.2991	
35	$-O-$	23	35	0.8515	
36	-CHO	9	9	0.4476	aldehyde group
37	-COOH	25	26	0.2653	conjugated acid
38	-COOH	15	24	1.1695	nonconjugated acid
39	$-COO-$	32	35	0.8724	ester
40	$-CONH_2$	0	0	undetermined	
41	$-CONH-$	3	3	0.1931	
42	$-CON-(-)$	0	0	undetermined	
43	$-CON=$	0	0	undetermined	
44	$-CO-$	7	7	1.3049	
45	$-C^*O-$	3	3	1.5413	
46	$-NO-$	0	0	undetermined	not in NO_2
47	$-PO-$	0	0	undetermined	
48	$-SO-$	3	6	0.5826	
49	$-NH_2$	15	15	0.6935	
50	$-NH-$	7	8	0.9549	
51	$-N^*H-$	0	0	undetermined	
52	$-N-(-)$	0	0	undetermined	
53	$-N^*-(-)$	0	0	undetermined	
54	$-C\equiv N$	5	5	0.6262	
55	$HN=$	0	0	undetermined	
56	$-N=$	0	0	undetermined	
57	$-N^*=$	4	7	-0.3722	
58	$-NO_2$	17	19	-0.2647	
59	-SH	3	3	-0.5118	
60	$-S-$	0	0	undetermined	
61	$-S^*-$	0	0	undetermined	
62	$S=P-$	5	5	-2.4086	
63	$S=$	4	4	-1.3197	not in $S=P$
64	P	0	0	undetermined	sp^3 phosphorus
65	P	0	0	undetermined	non-sp^3 phosphorus, not in $S=P$
66	alkanes	6	6	-1.5387	
67	other hydrocarbons	76	76	-0.2598	any hydrocarbon except for alkanes
constant (C_0)				3.5650	

[a] * indicates the atom is in a ring system. The open valences in parameters 1–65 are not filled by hydrogens.

approach based on graph theory (connectivity index) and polarizability. The practical superiority of this approach over the others is that it does not require any experimental data since all the necessary information can be calculated solely from molecular structure. However, its prediction potential has only been evaluated for a limited number of compounds. Irmann[21] used a group contribution approach to estimate the aqueous solubility of hydrocarbons and halo hydrocarbons. For different classes of compounds, different equations were employed. Wakita et al.[42] have taken a significant step in developing a group contribution scheme for the estimation of the water solubility values for organic compounds. Two different regression equations were obtained for liquid solutes and solid solutes. Bodor et al.[43,44] have recently attempted to correlate the water solubility of several classes of organic chemicals with a number of types of parameters, including molecular volume and surface, ovality of the molecule, atomic charges on hydrogen, carbon, nitrogen, and oxygen, dipole moment, molecule type indicators, and number of N–H single bonds in the molecule, through either a regression analysis or the neural network technique. Thus, all approaches currently available for the estimation of aqueous solubility either need experimental data or their use is restricted to limited classes of compounds.

We have recently developed a very reliable estimation methodology for the calculation of the partition coefficient between *n*-octanol and water (log *P*).[22] A fairly good correlation between water solubility and log *P* for a number of classes of organic compounds had been demonstrated by a number of groups.[7–13,15] This correlation was also shown to result from fundamental solute–solvent interactions as characterized by solvatochromic parameters.[23] Thus if the log *P* value is known for a molecule, the water solubility of that molecule can be estimated. Island and Lambert showed[13] that one can calculate the aqueous solubility value for a compound with a standard deviation of 0.63 log unit if its log *P* value is known. Currently the best available log *P* estimation methodologies can provide the log *P* value of an organic compound with a standard deviation around 0.4.[22] Therefore, the estimation of solubility will be poor if one uses this approach without further refinements.

It was previously found that log *P* can be calculated by the summation of the constitutional group's contribution.[24,25,42] This approach has been applied to the estimation of log *P* values of organic compounds in a number of laboratories[24–29] including ours,[22,30] and programs have also been developed in some laboratories.[22,26,27,31] Because of the good correlation between log *P* and water solubility and the successful application of the additive method to the calculation of log *P* values of organic compounds, we decided to explore the possibility of using the group contribution approach directly in the calculation of water solubility of organic compounds.

WATER SOLUBILITY CALCULATION

Solubility Estimation Model I. Methods and Materials. We assumed that water solubility values can be calculated by the following equation:

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

where S is the water solubility, C_0 is a constant, G_i is the ith group, and C_i is the contribution coefficient from the ith group. This equation implies, of course, that the effect of interactions between groups is negligible in the aqueous solubility estimation.

First, a set of 65 basic group parameters have been defined for our study, as shown in Table I. This set of group parameters essentially consists of two basic types of parameters, (1) basic fundamental groups (OH, CHO, COOH, COO, $CONH_2$, CONH, CON, CON=, CO, NO, NO_2, PO, SO, NH_2, NH, CN, SH) and (2) heavy atoms with both their hybridization and the number of hydrogen(s) attached to them specified. Sometimes the nearest heavy atom is also specified. Besides these, two other indicator parameters, indicating that the molecule is an alkane or any other hydrocarbon except for an alkane, are also used in our model since we found them to be important in our previous partition coefficient (*n*-octanol/water) study.[22] The quality of the final solubility model is essentially related to the set of group parameters which were defined. Currently, we use a relatively small set of such parameters in order to maintain the statistical significance of the parameters and of the model since the learning database is also relatively small. It should be emphasized though that the set of group parameters can be extended if one finds it necessary to include additional groups to accommodate uncommon molecules.

Quantitative solubility data in the literature are scarce.[18] Ideally, one would like to have a large learning set in order to produce a solubility estimation model as stable as possible and establish the statistical significance of each of the selected parameters. Unfortunately, for many organic compounds, the reported aqueous solubility is qualitative rather than quantitative and cannot be used to create a quantitative model. In some cases, the reported solubility data vary dramatically from different sources. We were nevertheless able to collect a database consisting of 483 compounds from four sources[13,18,36,37] for our study.

The compounds were entered into our program using the KLN methodology.[32] A program was developed to identify the occurrences of each parameter in the compounds from their connectivity matrix generated from KLN codes. A standard incremental regression analysis was then used to find which parameters correlate best with water solubility. The significance of each parameter is assessed by its *t* value, and the significance of the whole model is indicated by the *F* value.

After the relationship was found, it was coded into the program. Now this relationship can be used to calculate the aqueous solubility of any molecule by simply entering its molecular KLN code.

Results and Discussion. In order to evaluate the quality of our approach, we first used a relatively small database, which consisted of 200 compounds also selected by Nirmalakhandan and Speece in their solubility model study.[18]

Among the 67 group parameters selected as potential parameters of water solubility, 23 had nonzero occurrences in this learning database. The regression analysis found that these 23 parameters were all significant enough to be included in the correlation. The following equation was found correlating these parameters with solubility:

$$\begin{aligned}\log S = {} & 3.7968 - 1.7081P_1 - 0.4991P_2 + 0.7993P_3 + \\ & 1.9536P_4 - 1.7856P_5 - 0.2703P_6 + 0.9041P_7 - 0.6587P_8 - \\ & 2.4725P_9 - 0.7946P_{14} + 0.4245P_{15} - 1.5939P_{20} - \\ & 1.8126P_{21} - 1.6823P_{22} - 1.9195P_{23} + 0.2711P_{26} + \\ & 0.5178P_{27} + 2.8732P_{28} + 0.5807P_{29} + 1.2139P_{35} - \\ & 0.5244P_{36} + 0.0484P_{39} + 0.2115P_{67}\end{aligned}$$

$$n = 200,\ r^2 = 0.9692,\ F(23\ 176) = 240.69,\ \mathrm{SD} = 0.2227 \tag{2}$$

In eq 2, S refers to the aqueous solubility of an organic compound in grams/hundred grams of water (g/g%) and was measured at 25 °C; P_i (i = 1–65) indicates the number of times the corresponding fragment exists in a compound. P_{67} is equal to 1 for nonalkane hydrocarbons and 0 otherwise.

Our results are comparable to, if not better than, those reported by Nirmalakhandan and Speece using a graph-

Table II. Cross-Validation Test Results for 200-Compound Database

	learning set			test set		
	no. of compds	r^2	SD	no. of compds	r^2	SD
1	156	0.97	0.22	44	0.97	0.27
2	159	0.96	0.23	41	0.98	0.22
3	166	0.97	0.21	34	0.90	0.36
4	157	0.97	0.22	43	0.94	0.33
5	158	0.97	0.21	42	0.93	0.36
mean	159	0.97	0.22	41	0.94	0.31

Table III. Cross-Validation Results for 461-Compound Database

	learning set			test set		
	no. of compds	r^2	SD	no. of compds	r^2	SD
1	442	0.961	0.461	27	0.924	0.426
2	440	0.961	0.457	29	0.933	0.530
3	441	0.962	0.458	28	0.905	0.520
4	448	0.960	0.456	21	0.967	0.546
5	443	0.961	0.461	26	0.949	0.438
6	441	0.962	0.454	28	0.928	0.544
7	445	0.959	0.458	24	0.972	0.507
8	443	0.961	0.452	26	0.949	0.584
9	451	0.960	0.458	18	0.968	0.504
10	444	0.960	0.461	25	0.964	0.426
mean	443.8	0.961	0.458	25.2	0.946	0.503

theoretical and polarizability-based approach developed with a subset of 145 compounds ($n = 145$, $r^2 = 0.924$, SD = 0.318).[18] In their approach, two graph indices were used as well as the calculated polarizability, which was derived using the number of hydrogen, carbon, chlorine, and bromine atoms and the number of double bonds.

A very efficient and powerful approach for evaluating the validity of a model in QSAR and QSPR studies is the cross-validation test.[33] To do so, 20% of compounds from the original database were randomly selected as a test set while the remaining 80% were used as a learning database. A model is developed based on the learning set, and a prediction is performed for the test set. The r^2 and standard deviation are calculated for both the learning and test sets. The results for five such cross-validation experiments are shown in Table II.

It can be seen that the average r^2 value for the learning sets is 0.97 and that the average standard deviation is 0.22. For the test sets, the average cross-validation r^2 is 0.94 and the average standard deviation is 0.31. We feel that these results are satisfactory.

Having established that the methodology was adequate, we turned our attention to developing the model for the largest possible number of compounds which would include many different classes of organic compounds as well. It should be noted though that an attempt to include every possible class of organic compounds is not realistic at this time due to the lack of proper aqueous solubility data.

More aqueous solubility data were found in Island and Lambert's recent publication.[13] However, among the 300 compounds in their compilation,[34] some are redundant with our previous 200 compounds. After the redundant compounds were removed, 230 new compounds could be added to the learning set, thus expanding our database to 430 compounds. Fifty-three parameters were now found to have nonzero occurrences in this 430-compound set; however, we found that 15 of these 53 parameters occur only once or twice in the entire set. Therefore, the derived values for these 15 parameters through the regression analysis will not be reliable since the values heavily depend on the solubility data of one or two compounds. In order to make the whole correlation stable and reliable, more data were found and entered into the database.[36,37] A set of 483 compounds was compiled.[35] A quick evaluation found that there were 54 parameters having nonzero occurrences in the 483-compound database. Among them, nine parameters occur once or twice. Since we were unable to obtain more data to increase the occurrences of those nine parameters, we decided to remove those compounds (14 total) containing these parameters in order to obtain a reliable correlation. Thus, a database of 469 compounds was formed to establish the solubility estimation model. Forty-five parameters were found to have nonzero occurrences, and each of them occurred in at least three compounds. A regression analysis found all of them to be significant enough to be included in the model. This regression analysis also derived the contribution values for each parameter. The occurrences of each parameter, the total number of compounds they appear in, and their contributions to the aqueous solubility are listed in Table I.

With these 45 parameters, a linear equation (model I) with a r^2 equal to 0.961, a $F(45\ 423)$ value equal to 229.1, and a standard deviation equal to 0.458 log unit is obtained.

In order to test the reliability and the prediction potential of model I, cross-validation experiments were performed again. About 5% of the compounds were randomly removed from the original database as a test set. The remaining 95% of the compounds formed the learning set, and models were established based on the reduced learning sets. The solubility of the compounds in the test set were then calculated using the corresponding model. Table III shows the r^2 values and the standard deviations of 10 such cross-validation experiments for both the learning and the test sets.

It can be seen that the r^2 and standard deviation values almost remain the same for each learning set. This shows that the correlations are very stable. The average r^2 value and the average standard deviation for the 10 cross-validation test sets is 0.946 and 0.503 log units, respectively. This indicates that our model can be used as a reliable tool to estimate the water solubility values for most organic chemicals.

However, one will encounter a problem when one uses this model to predict the solubility values for new compounds which contain one or more parameters not included in model I. In these cases, a big error is normally expected for the estimation. It can be seen that only 45 out of 67 parameters were evaluated, although a fairly large set (469 compounds) was used to establish model I. And indeed, these 45 parameters covered many classes, but not every class of organic chemicals. As mentioned earlier, this problem can possibly be solved in the future if more solubility data become available to us. The learning database could be expanded and the parameter set can be extended, thus updating model I. Eventually the model could cover every class of organic chemicals. Nevertheless, at the current stage, it will be very useful to find an alternative to solve this problem since we often feel the need to estimate the solubility values for newly proposed chemicals in our drug design process. In many cases, some unknown parameters occur in the proposed chemicals.

Solubility Estimation Model II. Examining the defined parameters in Table I, it can be seen that they were constructed in a very restricted and precise manner. This might be essential to achieve a high accuracy for the model, but it also sacrifices some applicability of the model. Thus, it is possible to design a less restricted and less precise set of parameters and to utilize them to establish a model having a wider range of applicability. And of course, as a tradeoff, some accuracy will be lost.

Table IV. Contribution Values to log *S* of Group Parameters in Model II[a]

	parameter	no. of compds	frequency of use	contribution	remarks
1	CH_3	258	458	−0.4169	sp^3
2	CH_2	223	628	−0.5199	sp^3
3	CH	93	137	−0.3057	sp^3
4	C	58	88	−0.1616	sp^3
5	$=CH_2$	15	17	−0.7788	sp^2
6	CH=	258	1262	−0.3843	sp^2
7	C=	286	890	−0.5085	sp^2
8	C or CH	13	22	−0.4711	sp
9	NH_2	15	15	0.6184	sp^3
10	NH	11	12	0.7796	sp^3
11	N*H	1	1	0.7974	sp^3
12	N	3	3	1.0734	sp^3
13	N*	2	2	0.3906	sp^3
14	N=	2	2	−0.8015	sp^2
15	N*=	5	8	−0.3677	sp^2
16	N	5	5	1.0026	sp
17	NO_2	18	21	−2.2003	
18	OH	122	138	1.0910	sp^3
19	O	67	93	0.4452	sp^3
20	O=	112	158	0.9545	sp^3
21	S	8	8	−0.6161	sp^3
22	S	8	8	−0.3648	sp^2
23	S	3	3	−1.9783	other sulfur
24	P	9	9	−0.9139	any phosphorus
25	F	20	53	−0.5862	
26	Cl	143	430	−0.6292	
27	Br	42	71	−0.9190	
28	I	10	11	−1.4676	
29	COO	36	40	−0.4537	ester group
30	COOH	40	50	−1.2440	acid group
31	$CONH_n$ (n = 0, 1, 2)	6	6	−0.5531	
32	alkanes	6	6	−1.8549	
33	other hydrocarbons	77	77	−0.2168	except for alkanes
constant (C_0)				3.7253	

[a] * indicates the atom is in a ring system.

The following set of parameters were designed, as shown in Table IV. A 483-compound learning database (i.e., the 469 compounds used in model I, plus another 14 chemicals which were removed due to the occurrence of unique parameters) was used to establish model II.

The regression analysis found that all of the parameters were significant enough to be included in model II. The occurrences of each parameter, the total number of compounds they appear in, and their contributions to the aqueous solubility are also shown in Table IV. The statistical data for model II are n = 483, r^2 = 0.948, F(33 449) = 245.3, and SD = 0.526 log units.

Again, the reliability and the prediction potential of this model are evaluated through the cross-validation experiments. Table V shows the results for 10 such cross-validation experiments.

Comparisons between Tables III and V show that both models are very stable. Their r^2 and standard deviation values remain almost constant in both cases after 5% of compounds are removed from the learning databases. Model I has a better accuracy than model II. The mean values of the standard deviations of the 10 test sets in the cross-validation experiments for models I and II are 0.46 and 0.55 log units, respectively. These results indicate that one can use model I to estimate the water solubility values with a higher accuracy (±0.46 log unit) for new proposed compounds if there are no unknown parameters in the new compounds. In cases where the newly proposed compounds contain some unknown parameter, model II can be used and fairly good results will be obtained (with an accuracy ±0.55 log units). Since the designed parameter set in model II covers almost every class of organic chemicals, it has a very wide applicability. The only remaining potential parameters such as N^+(sp3), N^+(sp2) are not included in model II because it has been shown that electrolytes have very different behavior in terms of their water solubility compared to neutral organic chemicals.

Table V. Cross-Validation Results for 483-Compound Database

	learning set			test set		
	no. of compds	r^2	SD	no. of compds	r^2	SD
1	457	0.947	0.533	26	0.955	0.425
2	459	0.946	0.530	24	0.967	0.483
3	453	0.948	0.525	30	0.933	0.596
4	462	0.948	0.523	21	0.938	0.646
5	454	0.947	0.528	29	0.935	0.576
6	460	0.947	0.526	23	0.947	0.646
7	463	0.948	0.526	20	0.946	0.589
8	463	0.946	0.531	20	0.979	0.450
9	456	0.945	0.528	27	0.967	0.560
10	455	0.948	0.530	28	0.958	0.493
mean	458.2	0.947	0.528	24.8	0.953	0.546

Both models I and II are coded in our water solubility estimation program. When the program is applied to estimate the solubility value for an organic chemical, it will first check if there is(are) unknown parameter(s) in the chemical for model I. If there is no unknown parameter in the chemical, model I is used to calculate the water solubility values; otherwise, model II will be used. Appendix I (see Supplementary Material) provides the chemical names of all the 483 compounds in our study, as well as their experimental and calculated water solubilities through models I and II.

Near the end of this work, a test set of compounds designed by Yalkowsky and Banerjee[41] became available to us. This test set contains 21 commonly used substances of environmental or pharmaceutical interest, including a wide variety of complex structures, often with multiple functionalities. This test set has been used by Yalkowsky and Banerjee[41] to evaluate the prediction potential of all the current available aqueous solubility estimation methods. In order to compare our models to all the current available methods, we estimated the solubility values of the 21 compounds in the test set, as shown in Table VI.

Model I can only be applied to estimate the water solubility values for 13 out of 21 compounds in the test set since the other compounds contain new parameters. The standard deviation between their estimated and observed water solubility values for these 13 compounds is 0.58 log unit. Actually, there is only one compound (diazinon) with a larger error between its estimated and observed water solubility, being −1.56 log units. The standard deviation will be 0.37 log unit if this outlier is eliminated. Model II was applied to estimate the water solubility values for all the 21 compounds in the test set. The standard deviation was found to be 1.25 log units for these 21 compounds. There are two compounds (antipyrine and diazepam) with very large errors between their estimated and observed water solubility values, being −3.15 and −2.78 log units, respectively. If these two outliers are eliminated from the data set, the standard deviation will be 0.86 log unit for the remaining 19 compounds.

Comparison of our models to all the current available water solubility estimation methods[41] shows that our models rank the best. The most accurate method can be applied to estimate

Table VI. Predicted and Experimental Water Solubility Values of 21 Compounds in Test Set[a]

	name	$(\log S)_I$	$(\log S)_{II}$	$(\log S)_{exp}$
1	2,2′,4,5,5′-PCB	−4.90	−4.90	−4.89[b]
2	benzocaine	1.42	1.29	0.68
3	aspirin	1.23	1.48	1.39
4	theophylline		1.93	1.63
5	antipyrine		0.24	3.39
6	atrazine	0.06	−0.05	−0.55
7	phenobarbital		0.92	0.66
8	diuron		0.15	−0.76
9	nitrofurantoin		0.81	−0.38
10	phenytoin		−0.47	−0.99
11	diazepam		−3.54	−0.76
12	testosterone	−1.27	−2.17	−1.07
13	lindane	−1.45	−1.88	−1.60
14	parathion	−0.84	−0.94	−1.29
15	diazinon	−2.32	−2.29	−0.76
16	phenolphthalein	−0.14	−1.48	0.10
17	malathion		0.06	−0.36
18	chlorpyrifos	−2.72	−2.77	−2.67
19	prostaglandin E2	−0.08	−1.21	0.53
20	DDT	−5.31	−5.00	−5.08
21	chlordane	−4.29	−4.55	−3.86[c]

[a] The water solubility unit is mol/M^3. $(\log S)_I$ and $(\log S)_{II}$ denote the estimated water solubility values by model I and model II, respectively. $(\log S)_{exp}$ denotes the experimental water solubility values. [b] Based on ref 19, the water solubility value (log *S*, mol/M^3) for 2,2′,4,5,5′-PCB should be −4.89 instead of −3.77 as from ref 41. [c] Based on ref 34, the water solubility value (log *S*, mol/M^3) for chlordane should be −3.86 instead of −2.35 as from ref 41.

only 6 out of the 21 compounds with a standard deviation of 0.83 log unit, in contrast to our model I, which can be applied to estimate 13 out of the 21 compounds with a standard deviation of 0.58 log unit. The method developed by Wikita et al.[42] can be applied to estimate 8 out of 21 compounds with a standard deviation equal to 1.27 log units. These results show that model I not only has a wider applicability but also has a better accuracy than any available method. The only two methods which can be applied to estimate the aqueous solubility for most of the compounds in the test set and also show a fairly good accuracy are PCLOGP and REK.LOGP. It was shown[41] that the PCLOGP approach can be used to estimate the aqueous solubility values for 20 out of 21 compounds in the test set with a standard deviation equal to 1.32 log units, while the REK.LOGP approach can be applied to estimate 16 out of 21 with a standard deviation equal to 1.68 log units. Model II compares favorably to both.

The major advantage of this approach is its wide applicability and its simplicity. It can give a good or fairly good estimation of the aqueous solubility for most organic chemicals. The water solubility estimation program only requires molecular structures (KLN code) as its input, no additional experimental data is needed. Thus, it can serve as a general tool for estimating the aqueous solubility of proposed compounds in a drug design process and in QSAR studies.

The shortcomings of this approach are also obvious. The effect of interactions between polar functional groups in the aqueous solubility estimation is not explicitly shown in the models. The success of this approach may be in part due to the fact that most of the chemicals in the databases are monofunctional. However, the models gave a fairly good estimation of the aqueous solubility for the limited number of multiple functional compounds available in the database. This indicates that the interactions between functional groups for these compounds do not play an important role. How the interactions between functional groups affect the aqueous solubility for multiple functional compounds is not clear. It remains a further issue to be studied if more aqueous solubility data become available. The effect of crystalline forces upon the aqueous solubility of solid polar compounds is currently not addressed in our approach. However, as shown by Yalkowsky and Valvani,[7] the problem can be partly solved if the melting point of a molecule is known. The effort of developing a melting point estimation approach is currently under way in our laboratory. As soon as the melting point estimation approach becomes available, it can be used to improve the aqueous solubility estimation. Figure 1 illustrates the implementation of our program.

APPLICATION TO BIODEGRADATION STUDIES

Man-made chemicals used as refrigerants, fire retardants, paints, solvents, herbicides, and pesticides can cause considerable environmental pollution and human health problems as a result of their persistence, toxicity, and transformation into hazardous metabolites. Degradation, therefore, is important because it can lead to detoxification and reduction of chemical accumulation in the environment. Among the chemical, microbial, and photolytic mechanisms of degradation, microbial degradation or biodegradation is the most important in pollution control because many populations of microorganisms which are capable of breaking down organic chemicals are naturally occurring throughout the environment. The study of the relationships between the structure and biodegradability and the determination of the factors which play key roles in the biodegradation process are essential for us to be able to predict the fate of chemicals released into the environment.

Previous biodegradation studies have found that many physicochemical properties such as water solubility, volatility, and oil–water partitioning can influence the rate of biodegradation.[2–4] In our recent study of the structure–biodegradation activity relationship, we found that water-soluble chemicals are usually more biodegradable than insoluble ones.[6] For a total of 120 chemicals, 67.5% of water-soluble compounds (27/40) biodegrade rapidly while only 37.5% of the water-insoluble compounds (30/80) do so. It was noted, however, that water solubility values are not available for a number of important compounds. Now it is possible to apply our aqueous water solubility estimation models to calculate the water solubility values for all the compounds in the database and to study the relationship between the aqueous solubility and biodegradation activity.

A database consisting of 283 organic compounds for which biodegradation activity was measured was obtained from the literature.[38] Of these, 119 compounds were active (biodegrade rapidly), while 164 compounds were inactive (biodegrade slowly or do not biodegrade at all).

The 283 molecules were coded using the KLN code,[32] and their aqueous solubility values were calculated through the program. It was found that model I was utilized to estimate the aqueous solubility values for 244 out of the 283 compounds (86%) in the biodegradation database, while model II was used to estimate the aqueous solubility values for the other 39 compounds (14%). The calculated aqueous solubility values and biodegradation activities are compiled and available upon request.

A cutoff value had to be determined for deciding whether a chemical is to be classified as soluble or insoluble in water. As no absolute cutoff value can be assessed, we decided to examine the database using different cutoff values. The results are shown in Table VII.

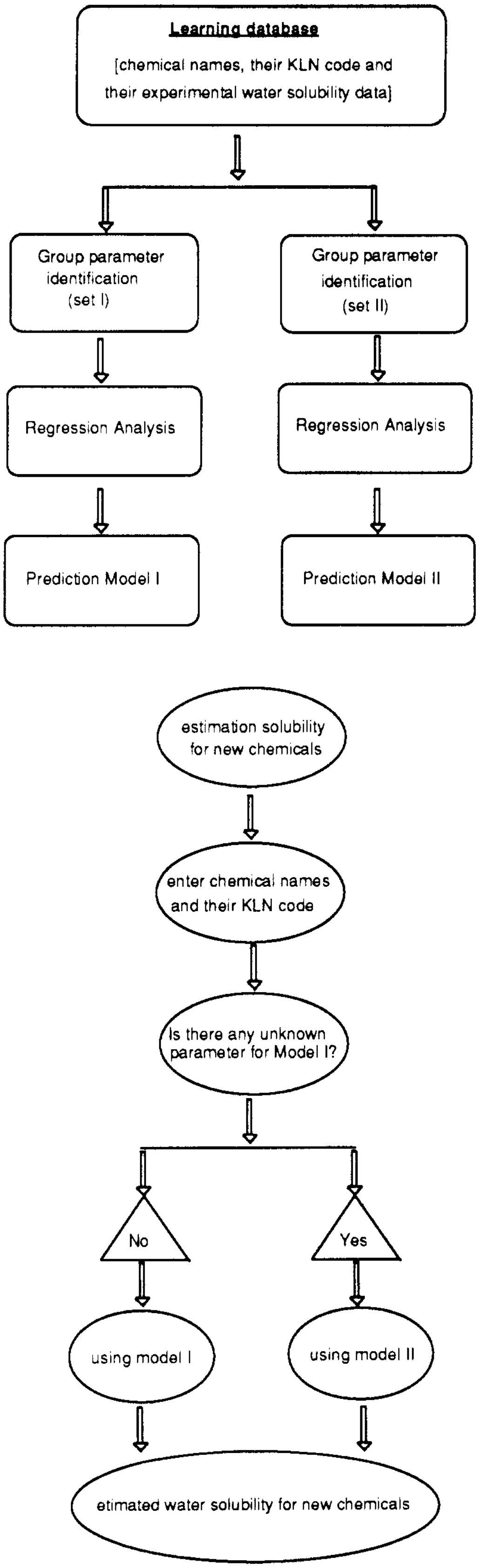
Learning database
[chemical names, their KLN code and their experimental water solubility data]
Group parameter identification (set I)
Group parameter identification (set II)
Regression Analysis
Regression Analysis
Prediction Model I
Prediction Model II
estimation solubility for new chemicals
enter chemical names and their KLN code
Is there any unknown parameter for Model I?
No
Yes
using model I
using model II
etimated water solubility for new chemicals

Figure 1. Implementation of the program.

Table VII. Biodegradation Activity and Solubility of Organic Compounds[a]

cutoff value of	soluble range			insoluble range		
log S (mol/M^3)	+	−	+%	+	−	+%
1.00	85	100	46	34	64	35
1.50	73	79	48	46	85	35
2.00	57	62	48	62	102	38
2.50	40	49	45	79	115	41
3.00	29	30	49	90	134	40

[a] In this table, + and − refer to the numbers of fast and slowly biodegradable compounds, respectively. +% refers to the percentage of the active compounds in the soluble or insoluble range. A compound is considered as soluble if its solubility is larger than the cutoff value; a compound is considered as insoluble if its solubility is less than or equal to the cutoff value.

Table VIII. Relationship between Biodegradation Activity and Solubility for Chain Compounds[a]

cutoff value of	soluble range			insoluble range		
log S (mol/M^3)	+	−	+%	+	−	+%
1.00	47	43	52	19	12	61
1.50	44	37	54	22	18	55
2.00	41	35	54	25	20	56
2.50	33	32	51	33	23	59
3.00	23	26	47	43	29	60

[a] All symbols have the same meanings as in Table VII.

Examination of this table clearly shows that if a compound is water soluble, it almost has an equal chance of being biodegradable and nonbiodegradable (the probability for biodegradation varied from 45% to 49% for different cutoff values, the average value being 47%). However, if a molecule is insoluble, it will only have on average a 38% chance of being biodegradable (the value varying from about 34% to 41% as the solubility cutoff values change from 1.00 to 3.00). This result shows that a soluble compound generally has a higher probability of being biodegradable than an insoluble one.

Previous studies had shown that chain and ring compounds might undergo different biodegradation mechanisms.[39] Therefore, we thought it would be interesting to evaluate these compounds separately. Thus, the 283 compounds were separated into two sets, one set consisting of 121 chain compounds, of which 55% were active. The other set contains 162 ring compounds, of which 33% were active and 148 contained aromatic ring(s).

Table VIII shows the evaluation results for the chain compounds.

As can be seen, the trend or "rule of thumb", that a compound with higher water solubility will have a higher probability to be biodegradable, is not observed for the chain compounds. Overall the average probability for a soluble compound to be biodegradable is 52%, while the average probability for an insoluble compound to be biodegradable is 58%. This indicates that there is no relation between water solubility and biodegradation activity for chain compounds. This probably means that other factors, such as the substructures identified by our CASE methodology in a previous study,[6] play the key role in the biodegradation process of chain compounds. For example, one of these substructures was the -$(CH_2)_{10}$- fragment which would lower solubility but provide a linear molecule which can be readily degraded by aerobic pathways.

The results of a similar evaluation for the 162 ring compounds are shown in Table IX.

Table IX shows that among soluble compounds, 39% (varying from 37% to 41%) are biodegradable. This is substantially greater than the database average of 33%. Furthermore, among insoluble compounds, only 24% (varying from 16% to 31%) are fast biodegradable. Thus, in this case, soluble compounds have a higher probability of being biodegradable than insoluble ones.

Table IX. Relationship between Biodegradation Activity and Solubility for Ring Compounds[a]

cutoff value of log S (mol/M^3)	soluble range			insoluble range		
	+	–	+%	+	–	+%
0.50	46	73	39	7	36	16
1.00	38	57	40	15	52	22
1.50	29	42	41	24	67	26
2.00	16	27	37	37	82	31

[a] All symbols have the same meanings as in Table VII.

Table X. Distribution of Active and Inactive Compounds within Each Solubility Range for Ring Compounds

solubility range	total no. of compds within the range	active/ inactive	active/ total, %
(<0.0)	27	7/20	26
(0.0, 2.0)	92	30/62	33
(>2.0)	43	16/27	37

The comparison between Tables VII and IX shows that, overall, chain compounds have a higher probability to be biodegradable than ring compounds. However, aqueous solubility apparently plays no significant role in the biodegradability of the chain compounds in the database.

In order to examine how the biodegradability changes as the solubility increases for ring compounds, we evaluated the active/inactive compound distribution within different ranges. Table X shows the results.

Table X clearly shows that as the solubility increases, the probability for a ring compound to be biodegradable increases, going from 26% to about 37%. However, if solubility was the only factor controlling the biodegradation rate, we would expect a much higher percentage of active compounds in the highest solubility range.

We further applied our approach to the prediction of biodegradation activity for 27 new chemicals obtained from the Howard et al. biodegradation study.[40] Since 20 of these chemicals are ring compounds, their biodegradability can possibly be predicted based solely on their aqueous solubility. Using 1.75 log units (mol/M^3) as the soluble/insoluble cutoff value and assuming that soluble compounds will be easily biodegradable, we correctly predicted the biodegradability of 74.1% (20/27) of the compounds, as is shown in Table XI. Interestingly, we correctly predicted every inactive compound (a total of 14 existed in the database) based on their lack of solubility only. This result, we believe, shows the significance of aqueous solubility; however, it also shows that other factors have to be considered in the biodegradation study. In the Howard et al. study,[40] these 27 new chemicals had been used to test their model-relating structural fragments to biodegradation activity. They correctly predicted every active compound in the database, but five of the inactive compounds were incorrectly predicted to be biodegradable. Comparing their results to ours, we find that every compound incorrectly predicted in their study was correctly predicted in ours, while every compound incorrectly predicted in our study was correctly predicted in theirs. This suggests that including the water solubility in the Howard et al. model may lead to a more powerful model for the prediction of biodegradation activity of organic chemicals.

Table XI. Prediction of Biodegradation Activity of New Compounds Only[a] Using Solubility

	name	solubility		pred[b]	pred[c]	exp
1	3-methylcholanthrene	–5.71	I[d]	–	–	–
2	1-arginine	2.86	II[e]	+	+	+
3	hydroxyl acetic acid	5.29	I	+	+	+
4	1,2-dibromo-3-chloropropane	0.12	I	–	+	–
5	4-nitroaniline	1.53	I	–	–	–
6	butanoic acid	2.92	I	+	+	+
7	dibromochloromethane	1.27	I	–	+	–
8	dibenzofuran	–0.55	II	–	+	+
9	acenaphthylene	–1.62	I	–	+	+
10	δ-hexachlorocyclohexane	–1.45	I	–	–	–
11	neburon	–1.41	II	–	–	–
12	2,3-dimethylnaphthalene	–1.56	I	–	+	+
13	terbutryn	–0.73	II	–	–	–
14	pebulate	–0.29	II	–	+	+
15	4-hydroxyl-3-methoxy-cinnamic acid	1.97	I	+	+	+
16	2,3,7,8-tetrachlorodibenzo-*p*-dioxin	–4.99	I	–	–	–
17	2,4-dichloro-1 (4-nitrophenoxy)benzene	–2.17	I	–	–	–
18	dodecyl benzenesulfonate	–4.72	I	–	+	+
19	chloroxuron	–1.21	II	–	+	–
20	2,6-dichlorobenzamide	0.30	II	–	+	–
21	carboxin	0.26	II	–	+	+
22	ethyl 3-hydroxybenzoate	1.92	I	+	+	+
23	oryzalin	–0.95	II	–	–	–
24	acetate	3.02	II	+	+	+
25	2,2′,5,5′-tetrachlorobiphenyl	–4.14	I	–	–	–
26	diclofopmethyl	–1.79	I	–	+	+
27	fluridone	–4.55	II	–	+	–

[a] +, biodegradable fast; –, biodegradable slowly or nonbiodegradable. [b] Prediction based on solubility only. [c] Prediction based on the Howard et al. model, obtained from ref 40. [d] I indicates that the solubility value was calculated by model I. [e] II indicates that the solubility value was calculated by model II.

CONCLUSION

A reliable and generally applicable method for the estimation of aqueous solubility has been developed. Two models have been established based on two different sets of parameters. Model I has a higher accuracy, while model II has a wider applicability. The mean cross-validated r^2 value and standard deviation of 10 cross-validation experiments for model I were found to be 0.96 and 0.46 log units, respectively. And the mean cross-validated r^2 value and standard deviation of 10 such cross-validation experiments for model II were found to be 0.95 and 0.53 log units, respectively. Applying our models to estimate the water solubility values for the compounds in an independent test set, we found that model I can be applied to 13 out of 21 compounds with a standard deviation equal to 0.58 log unit and that model II can be applied to all the 21 compounds with an standard deviation equal to 1.25 log units. Our models compare favorably to all the current available water solubility estimation methods. A program based on this approach has been written in FORTRAN77 and is currently running on a VAX/VMS system. The program can be applied to estimate the water solubility of any organic chemical with a good or fairly good accuracy except for electrolytes. Even though, one can expect better results, should more experimental data become available, we feel that the current program is satisfactory as a general tool to estimate aqueous solubility values of organic chemicals. We applied our methodology to calculate the solubility of a set of compounds whose biodegradation activity was known. Evaluation of the correlation between the water solubility and the biodegradation activity showed that although the water solubility was not the sole factor controlling the rate of biodegradation, ring compounds with greater solubilities were more likely to biodegrade at a faster rate. The difference observed for chain and ring compounds suggests that chemical structure (i.e., reactivity), the major factor controlling the rate of biodegradation and the role of the water solubility, only becomes important when reactivity is lowered. Nevertheless, the significance of the aqueous solubility is shown through the prediction of the biodegradation activity of 27 new chemicals.

The solubility estimation program is available for distribution from DSI, care of G. Klopman at CWRU.

ACKNOWLEDGMENT

We are grateful to the National Institutes of Health for their support of this work through a grant under the NCDDG-OI program.

Supplementary Material Available: Appendix I giving names of the compounds used in the study and their calculated and experimental water solubility values (7 pages). Ordering information is given on any current masthead page.

REFERENCES AND NOTES

(1) Yalkowsky, S. H.; Morozowich, W. In *Drug Design*; Ariens, E. J., Ed.; Academic: New York, 1980; Vol. 9, p 121.

(2) Mudder, T. I. Development of empirical structure–biodegradability relationships and testing protocol for slightly soluble and volatile priority pollutants. University Microfilms International Order No. 8123345, Ann Arbor, MI, 1976.

(3) Vaishnav, D. D.; Boethling, R. S.; Babeu, L. Quantitative structure–biodegradability relationships for alcohols, ketones and alicyclic compounds. *Chemosphere* **1987**, *16*, 695–703.

(4) Lyman, W. J.; Reehl, W. F.; Rosenblatt, D. H. *Handbook of Chemical Property Estimation Methods*; McGraw-Hill: New York, 1982; Chapter 9.

(5) For example, see Zaroogian, G. E.; Heltshe, J. F.; Johnson, M. *Environ. Toxicol. Chem.* **1985**, *4*, 3.

(6) Klopman, G.; Balthasar, D. M.; Rosenkranz, H. S. Application of the Computer Automated Structure Evaluation (CASE) program to the Study of Structure-Biodegradation Relationships of Miscellaneous Chemicals. *Environ. Toxicol. Chem.*, in press.

(7) Yalkowsky, S. H.; Valvani, S. C. Solubility and Partitioning I: Solubility of Non-electrolytes in water. *J. Pharm. Sci.* **1980**, *69* (8), 912–922.

(8) Hansch, C.; Quinlan, J. E.; Lawrence, G. L. The Linear Free-Energy Relationships between Partition Coefficients and Aqueous Solubility of Organic Liquids. *J. Org. Chem.* **1968**, *33*, 347–350.

(9) Yalkowsky, S. H.; Orr, R. J.; Valvani, S. C. Solubility and Partitioning: 3 the Solubility of Halobenzenes in Water. *Ind. Eng. Chem. Fundam.* **1979**, *18*, 351–353.

(10) Yalkowsky, S. H.; Valvani, S. C. Relationships between Aqueous Solubilities, Partition Coefficients, and Molecular Surface Areas of Rigid Aromatic Hydrocarbons. *J. Chem. Eng. Data* **1979**, *24*, 127–129.

(11) Banerjee, S.; Yalkowsky, S. H.; Valvani, S. C. Water Solubility and Octanol/Water Partition Coefficients of Organics, Limitations of the Solubility Partition Coefficient Correlation. *Environ. Sci. Technol.* **1980**, *14*, 1227–1229.

(12) Briggs, G. G. Theoretical and Experimental Relationships between Sole Adsorption, Octanol–Water Partition Coefficients, Water Solubilities, and Their Relationship to Respective Water Solubility(s) Values. *J. Agric. Food Chem.* **1981**, *29*, 1050–1059.

(13) Isnard, P.; Lambert, S. Aqueous Solubility and *n*-Octanol/Water Partition Coefficients Correlations. *Chemosphere* **1989**, *18* (9/10), 1837–1853.

(14) Warne, M. St. J.; Connell, D. W.; Hawker, D. W.; Schuurmann, G. Prediction of Aqueous Solubility and the Octanol–Water Partition Coefficient for Lipophilic Organic Compounds Using Molecular Descriptors and Physico-chemical Properties. *Chemosphere* **1990**, *21* (7), 877–888.

(15) Patil, G. S. Correlation of Aqueous Solubility and Octanol–Water Partition Coefficient Based on Molecular Structure. *Chemosphere* **1991**, *22* (8), 723–738.

(16) Kamlet, M. J.; Abraham, M. H.; Carr, P. W.; Doherty, R. M.; Taft, R. W. Solute–Solvent Interactions in Chemistry and Biology. Part 7, An analysis of Mobile Phase Effects on High Pressure Liquid Chromatography Capacity Factors and Relationships of the Latter with Octanol-Water Partition Coefficients. *J. Chem. Soc. Perkin Trans. 2* **1988**, 1087–1097.

(17) Kamlet, M. J.; Dolerty, R. M.; Abraham, M. H.; Marcus, Y.; Taft, R. W. Linear Solvation Energy Relationships. 46. An Improved Equation for Correlation and Prediction of Octanol/Water Partition Coefficients of Organic Nonelectrolytes (Including Strong Hydrogen Bond Donor Solutes). *J. Phys. Chem.* **1988**, *92*, 5244–5255.

(18) Nirmalakhandan, N. N.; Speece, R. E. Prediction of Aqueous Solubility of Organic Chemicals Based on Molecular Structure. *Environ. Sci. Technol.* **1988**, *22*, 328–338.

(19) 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.

(20) Speece, R. E. Comment on Prediction of Aqueous Solubility of Organic Chemicals Based on Molecular Structure. 2. Application to PNAs, PCBs, PCDDs, etc. *Environ. Sci. Technol.* **1990**, *24*, 927–929.

(21) Irmann, F. A Simple Correlation Between Water Solubility and Structure of Hydrocarbons and Halohydrocarbons. *Chem.-Ing.-Tech.* **1965**, *37*, 789–798.

(22) Klopman, G.; Wang, S. A Computer Automated Structure Evaluation (CASE) Approach to Calculation of Partition Coefficient. *J. Comput. Chem.* **1991**, *12*, 1025–1032.

(23) Hawker, D. The Relationship between octan-1-ol/Water Partition Coefficient and Aqueous Solubility in Terms of Solvatochromic Parameters. *Chemosphere* **1989**, *19*, 1585–1593.

(24) Rekker, R. F. *The Hydrophobic Fragmental Constant*; Elsevier: New York, 1977.

(25) Leo, A.; Hansch, C.; Elkins, D. *Chem. Rev.* **1971**, *71*, 533.

(26) Chou, J.; Jurs, P. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 172.

(27) Leo, A. In *Comprehensive Medicinal Chemistry*; Hansch, C., Ed.; Pergaman: Oxford, 1990; Vol. 4, pp 295–319.

(28) Ghose, A. K.; Pritchett, A.; Crippen, G. M. *J. Comput. Chem.* **1988**, *9*.

(29) Broto, P.; Moreau, G.; Vandycke, C. *Eur. J. Med. Chem.* **1984**, *19*, 71.

(30) Klopman, G.; Namboodiri, K.; Schochet, M. *J. Comput. Chem.* **1985**, *6*, 28.

(31) Suzuki, T.; Kudo, Y. Automated log *P* Estimation Based on Combined Additive Modeling Methods. *J. Comput.-Aided Mol. Design* **1990**, *4*, 155–198.

(32) Klopman, G.; McGonigal, M. *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 48.

(33) Wold, S. *Technometrics* **1978**, *20*, 397.

(34) Compound under the name Toxaphene is a very complex mixture, it was excluded from the database.

(35) The solubility values of all 483 compounds were converted to the same unit, mol/M^3.

(36) (a) Yaws, C. L.; Yang, H.-C.; Hopper, J. R.; Hansen, K. C. *Chem. Eng.* **1990**, *April*, 177–183. (b) Yaws, C. L.; Yang, H.-C.; Hopper, J. R.; Hansen, K. C. *Chem. Eng.* **1990**, *July*, 115–118.

(37) Stephen, H.; Stephen, T., Eds. *Solubilities of Inorganic and Organic Compounds, Vol. 1, Binary Systems, Part 1*; Macmillan: New York, 1963; pp 55–79.

(38) Howard, P. H.; Hueber, A. E.; Boethling, R. S. Biodegradation data evaluation for structure/biodegradation relations. *Environ. Toxicol. Chem.* **1987**, *6*, 1–6. For criteria of determining a chemical to be active or inactive, also see ref 40.

(39) For example, see Chaudhry, G. R.; Chapalmadugu, S. *Microbiol. Rev.* **1991**, *55* (1), 59–79.

(40) Howard, P. H.; Boethling, R. S.; Stiteler, W. M.; Merlan, W. M.; Hueber, A. E.; Beauman, J. A.; Larosche, M. E. Predictive Model for Aerobic Biodegradability Developed From a File of Evaluated Biodegradation Data. *Environ. Toxicol. Chem.*, in press.

(41) Yalkowsky, S. H.; Banerjee, S. *Aqueous Solubility, Methods of Estimation for Organic Compounds*; Marcel Dekker: New York, Basel, and Hong Kong, 1992; Chapter 4, pp 128–148.

(42) Wakita, K.; Yoshimoto, M.; Miyamoto, S.; Watanabe, H. A Method for Calculation of the Aqueous Solubility of Organic Compounds by Using New Fragment Solubility Constants. *Chem. Pharm. Bull. (Tokyo)* **1986**, *34*, 4663–4681.

(43) Bodor, N.; Huang, M.-J. Submitted for publication.

(44) Bodor, N.; Harget, A.; Huang, M.-J. Neural Network Studies. 1. Estimation of the Aqueous Solubility of Organic Compounds. *J. Am. Chem. Soc.* **1991**, *113*, 9480–9483.

Registry No. 2,2′,4,5,5′-PCB, 37680-73-2; DDT, 50-29-3; benzocaine, 94-09-7; aspirin, 50-78-2; theophylline, 58-55-9; antipyrine, 60-80-0; atrazine, 1912-24-9; phenobarbital, 50-06-6; diuron, 330-54-1; nitrofurantoin, 67-20-9; phenytoin, 57-41-0; diazepam, 439-14-5; testosterone, 58-22-0; lindane, 58-89-9; parathion, 56-38-2; diazinon, 333-41-5; phenolphthalein, 77-09-8; malathion, 121-75-5; chlorpyrifos, 2921-88-2; prostaglandin E2, 363-24-6; chlordane, 12789-03-6; 3-methylcholanthrene, 56-49-5; arginine, 74-79-3; hydroxyl acetic acid, 79-14-1; 1,2-dibromo-3-chloropropane, 96-12-8; 4-nitroaniline, 100-01-6; butanoic acid, 107-92-6; dibromochloromethane, 124-48-1; dibenzofuran, 132-64-9; acenaphthylene, 208-96-8; hexachlorocyclohexane, 319-86-8; Neburon, 555-37-3; 2,3-dimethylnaphthalene, 581-40-8; Terbutryn, 886-50-0; Pebulate, 1114-71-2; 4-hydroxy-3-methoxycinnamic acid, 1135-24-6; 2,3,7,8-tetrachlorodibenzo-*p*-dioxin, 1746-01-6; 2,4-dichloro-1-(4-nitrophenoxy)benzene, 1836-75-5; dodecyl benzenesulfonate, 1330-69-4; chloroxuron, 1982-47-4; 2,6-dichlorobenzamide, 2008-58-4; carboxin, 5234-68-4; ethyl 3-hydroxybenzoate, 7781-98-8; oryzalin, 19044-88-3; acetate, 64-19-7; 2,2′,5,5′-tetrachlorobiphenyl, 35693-99-3; Diclofopmethyl, 51338-27-3; fluridone, 59756-60-4.