Prediction of Aqueous Solubility for a Diverse Set of Heteroatom-Containing Organic Compounds Using a Quantitative Structure−Property Relationship

Jon M. Sutter and Peter C. Jurs*

Department of Chemistry, The Pennsylvania State University, 152 Davey Laboratory, University Park, Pennsylvania 16802

Received October 3, 1995[⊗]

The primary goal of a quantitative structure−property relationship (QSPR) is to identify a set of structurally based numerical descriptors that can be mathematically linked to a property of interest. The types of descriptors fall into three categories: topological, electronic, and geometric. In this study, 140 organic compounds with diverse structures were split into a training set, a cross-validation set, and a prediction set. The training set was used to build multiple linear regression and computational neural network models, the cross-validation set was used to prevent overtraining of the neural network, and the prediction set was used to validate the mathematical models. A set of nine descriptors was found that effectively linked the aqueous solubility to each structure. However, the polychlorinated biphenyls (PCBs) had a large root-mean-square (rms) error associated with them. Therefore models were also built using a training set that contained no PCBs. A set of nine descriptors was found with a significant improvement of the rms error of the training set as well as the prediction set.

INTRODUCTION

Aqueous solubility is a very important molecular property, which plays an integral role in many different biological and physical processes. A knowledge of the aqueous solubility could be crucial for such things as environmental impact studies and drug transport studies. Obtaining experimental values of aqueous solubility is not a difficult task, but there is justification for creating methods that can predict this property. Being able to predict the aqueous solubility of a compound would greatly assist drug development, because a solubility could be estimated before the compound was synthesized, which would eliminate the need to synthesize many of the unsuitable compounds. The ability to predict aqueous solubility would also be useful for people working on environmental impact studies, because the solubility could be estimated without handling compounds that are highly toxic, carcinogenic, or undesirable for some other reason. The question is not whether there is a need for predicting aqueous solubility, but rather what is a good method to use? In this study the method of quantitative structure−property relationships (QSPRs) was investigated as a solution to the problem.

Simply stated, a QSPR relates the structure of a compound to the physical property of interest. Many structural features of a compound are quantified to numerically represent the molecule. The numerical values are referred to as descriptors and can be classified into three categories: topological, geometric, and electronic. The goal is to find a small subset of the large number of calculated descriptors that can effectively predict the property of interest. QSPRs seem to be a logical choice for predicting aqueous solubility since the solubility is almost exclusively dependent on the intermolecular forces that exist between the solute molecules and the solvent molecules. Those forces are dependent on the structure of the compound. Therefore the solute−solute, solute−solvent, and solvent−solvent interactions, which are largely responsible for the amount of the compound dissolving in water should be numerically represented by the descriptors derived from the structure of the molecule. Theoretically, QSPRs should be able to accurately predict the solubility, and this is proven by the numerous successful QSPR models found in the literature.

There are many methods that have been developed over the years to predict solubility.[1−7] In the small but representative selection of studies referenced here, many authors used molecular descriptors and multiple linear regression to predict the solubility. Some descriptors were complicated, semiempirically derived charge descriptors,[1] while others were simple, molecular connectivity descriptors.[2] One model used a group contribution method to predict solubility.[5] Although such a model has attractive features, such as the ease of calculating the solubility, there are limitations as well. A group contribution method is not able to determine the difference in solubility between isomers and is not able to predict the solubility of a compound that contains a functional group not present in the model. Another model used a theoretical approach to predict the aqueous solubility.[4] In this approach the activity coefficient, γ, was estimated from the water−octanol partition coefficient and used in conjunction with some experimental parameters to calculate the aqueous solubility. The result of this model was a very good fit; however, several experimental values were needed to estimate the aqueous solubility. The studies that corrected for the differences between the solubility of a liquid and a solid compound showed an improvement in the overall standard error. Isnard and Lambert[8] surveyed aqueous solubility values correlated to the *n*-octanol/water partition coefficients in the literature and observed a 5−12% improvement in the standard error when the effects of crystalline interactions were accounted for using the melting point of the solid compounds.

In the present study, an empirical method was used that did not rely on a correction term for the solid compounds. The justification for not correcting for solid compounds was

[⊗] Abstract published in *Advance ACS Abstracts,* January 1, 1996.

0095-2338/96/1636-0100$12.00/0

as follows: (1) when experimental properties are predicted it would be beneficial not to use other experimental values, (2) in past studies from the literature only a slight improvement was observed in the standard error when the correction term was employed, and (3) it is possible that some of the properties needed to determine the differences between liquids and solids will be present in the large pool of descriptors calculated from the structure. A method that could predict the aqueous solubility based solely on the structure of the compound would allow a chemist to estimate the solubility before the compound was even made, which, as stated earlier, would have obvious advantages in fields such as drug design.

EXPERIMENTAL SECTION

The quality of the data plays a crucial role in the development of a good QSPR. If the data are erroneous, the descriptors found to be important may contain information useless in predicting the property of interest but may contain information that is able to predict the erroneous values effectively. Such models would have no ability to generalize and therefore would not be able to accurately predict the aqueous solubility of unknown compounds. Heller et al. recently developed an expert system designed to alleviate the problems associated with inaccurate aqueous solubility data.[9] In Heller's method, the quality of the data is evaluated and given a rating from 1 to 4, 1 being the best rating. Only data that had been given a satisfactory rating (1−3) by this method was used in this study.

Since a QSPR can only predict properties for unknown compounds that are structurally similar to those used to build the model, a diverse data set was obtained. Out of 140 compounds, 13 were selected randomly for the prediction set, 11 were selected randomly for the cross-validation set, and the remaining compounds comprised the training set. The training set was used to build both multiple linear regression and computational neural network models, the cross-validation set was used to prevent overtraining of the neural network, and the prediction set was used to validate both types of models. Table 1 provides a list of all the compounds used with their corresponding reference, solubility rating, experimental aqueous solubility, and solubility predicted from a 9:3:1 computational neural network.

Each of the sources reported a remarkable precision in the repeated analysis of their samples (approximately 0.02 log units). The experimental solubility values obtained by different sources for the same compound did not reflect this high degree of precision. Shown in Table 2 is a list of all the compounds that were found in more than one source. The mean of the absolute values of the differences was 0.11 log units, and the range was 0−0.34 log units. The discrepancies can be attributed to several factors including, temperature control, purification of materials, equilibration, separation of phases, and analysis of the saturated solution.[14] In other words, slight variations in the experimental methods amongst the different laboratories resulted in slight variations in the aqueous solubility. For instance, two of the references[11,12] used the generator column method,[15] and the other two did not. Therefore, the observed differences in the aqueous solubility were attributed to the normal experimental errors expected between different methods. When two or more sources reported a solubility value for the same compound, the decision on which value to use was based on either which had the better solubility rating or which choice would maintain consistency. Therefore the values in refs 10, 11, and 12 were preferentially selected over the values in ref 13 since those sources had a better solubility rating, with the exception of 1-pentyne and 1-hexyne. The two alkynes were taken from ref 13 to maintain consistency, since the remaining alkynes were only found in ref 13.

Once the compounds were selected and their aqueous solubility values were assigned, the Automated Data Analysis and Pattern Recognition Toolkit[16,17] (ADAPT) method was used to build a QSPR. ADAPT is a combination of computer programs that have been created over the years to develop QSPRs and quantitative structure-activity relationships (QSARs). A flow diagram of the ADAPT methodology used in this study is shown in Figure 1. All computations related to the ADAPT method were performed on a DEC 3000 AXP Model 500 workstation.

As indicated by Figure 1, the first step in developing a QSPR using the ADAPT methodology was structure entry. This was done by sketching the compounds on a graphics terminal and then storing them as connection tables. After entry, the geometry of each compound was optimized using the semiempirical molecular orbital program MOPAC[18] with the PM3 Hamiltonian.[19]

A total of 144 descriptors were generated for each compound using the ADAPT software. The descriptors can be grouped into three categories, electronic, geometric, or topological, or a combination of these groups. The charged partial surface area (CPSA)[20] descriptors are generated by combining geometric and electronic information. Topological descriptors can be calculated directly from the two-dimensional connection tables. Some examples of topological descriptors that were found in the pool of descriptors are the molecular weight, atom counts, and connectivity descriptors. Geometric descriptors are dependent on the three-dimensional representation of the molecule. Examples of these descriptors that were members of the pool of descriptors in this study are the length-to-breadth ratio, molecular volume, and radius of gyration. Electronic descriptors contained the electrical information of the compound and included such things as the most negative or the most positive partial atomic charge. A semiempirical method for calculating atomic charges (PKACHG)[21] was used to characterize the electronic structure of the compound. PKACHG partitioned the atomic charges into separate σ and π schemes. The atomic charge for each atom was the sum of these σ and π charges.

Not all of the descriptors could be used to build a statistically sound model. Therefore, the next step to build the QSPR was descriptor reduction. The descriptor pool was first reduced by analyzing the data without regard to the dependent variable, aqueous solubility. Descriptors that contained the same value for every compound were discarded, and one of two descriptors that were highly pairwise correlated was also discarded.

Using these methods, the original pool of 144 descriptors was reduced to 53 descriptors. To reduce the descriptors to an acceptable subset size that contained the most important information, the genetic algorithm[22] and simulated annealing[23] optimization routines were employed. Both routines are iterative improvement optimization techniques with a small degree of randomness that allowed them to avoid local

Table 1. Compounds and Their Corresponding Experimental and Predicted Solubility Values Used in This Study[f]

no.	compd name	−log(*S*)[a] exp	−log(*S*) pred	rating[b]	ref	no.	compd name	−log(*S*)[a] exp	−log(*S*) pred	rating[b]	ref
1	ethane	2.70	2.75	3	13	71	2-pentene[c]	2.54	2.90	3	13
2	propane	2.85	2.68	3	13	72	3-methyl-1-butene	2.73	2.78	3	13
3	*n*-butane	2.98	2.85	3	13	73	2-methyl-1-pentene[d]	3.03	3.16	3	13
4	isobutane[c]	3.08	2.83	3	13	74	4-methyl-1-pentene	3.24	3.22	3	13
5	*n*-pentane	3.18	3.26	2	10	75	1-hexene	3.08	3.28	2	12
6	isopentane	3.18	3.17	3	13	76	1-heptene	3.73	3.95	2	12
7	2,2-dimethylpropane[d]	3.33	3.12	3	13	77	2-heptene	3.82	4.08	3	13
8	*n*-hexane	3.84	3.84	2	10	78	1-octene	4.44	4.59	2	12
9	*n*-heptane	4.47	4.44	2	10	79	1-nonene[d]	5.05	5.24	2	12
10	*n*-octane	5.12	5.05	2	10	80	1,3-butadiene	2.13	1.98	3	13
11	2-methylpentane	3.74	3.69	2	10	81	2-methyl-1,3-butadiene[c]	2.03	1.90	3	13
12	2-ethyl-1,3-hexanediol[c]	2.81	2.75	2	10	82	1,4-pentadiene	2.09	1.96	3	13
13	3-methylpentane	3.68	3.69	2	10	83	1,5-hexadiene	2.67	2.57	3	13
14	3-methylhexane	4.31	4.30	2	10	84	1,6-heptadiene	3.34	3.30	3	13
15	2,2-dimethylbutane	3.56	3.58	2	10	85	propyne	1.04	1.06	3	13
16	2,3-dimethylbutane	3.58	3.60	2	10	86	1-butyne	1.28	1.34	3	13
17	2,4-dimethylpentane[c]	4.26	4.21	2	10	87	1-pentyne	1.64	1.74	2	13
18	2,2,4-trimethylpentane	4.74	4.69	2	10	88	1-hexyne	2.36	2.28	2	13
19	2,3,4-trimethylpentane	4.70	4.71	2	10	89	1-heptyne	3.01	2.94	3	13
20	2,2,5-trimethylhexane	5.38	5.26	2	10	90	1-octyne[d]	3.66	3.60	3	13
21	benzene	1.65	1.79	2	10	91	1-nonyne	4.24	4.41	3	13
22	toluene	2.21	2.19	2	10	92	1,6-heptadiyne	1.75	1.76	3	13
23	*o*-xylene[d]	2.70	2.71	2	10	93	1,8-nonadiyne	2.98	2.90	3	13
24	*m*-xylene[c]	2.82	2.74	2	10	94	1-chlorobutane[d]	2.03	2.20	2	12
25	*p*-xylene	2.76	2.78	2	10	95	1-chloroheptane	4.00	4.15	2	12
26	ethylbenzene[d]	2.78	2.67	2	10	96	1-bromobutane	2.20	2.40	2	12
27	*n*-propylbenzene	3.36	3.25	2	12	97	1-bromopentane	3.08	3.07	2	12
28	*n*-butylbenzene[c]	3.99	3.85	2	12	98	1-bromohexane[d]	3.81	3.78	2	12
29	*n*-pentylbenzene	4.59	4.54	2	12	99	1-bromooctane	5.06	4.98	2	12
30	*n*-hexylbenzene	5.20	5.20	2	12	100	bromochloromethane	0.89	1.05	2	12
31	1-ethyl-2-methylbenzene	3.21	3.23	2	12	101	1-bromo-3-chloropropane	1.85	1.82	2	12
32	iodobenzene	3.01	2.87	2	12	102	1-iodoheptane[d]	4.81	5.01	2	12
33	*o*-fluorobenzylchloride	2.54	2.52	2	12	103	trichloroethylene	1.98	2.12	2	12
34	*m*-fluorobenzylchloride	2.54	2.55	2	12	104	4-bromo-1-butene	2.25	1.79	2	12
35	*m*-cresol[e]	1.59		2	12	105	allyl bromide	1.50	1.27	2	12
36	nitrobenzene[e]	1.51		2	12	106	cyclopentane	2.65	2.48	3	13
37	chlorobenzene	2.58	2.41	2	11	107	cyclohexane	3.18	3.10	3	13
38	*o*-dichlorobenzene	3.20	3.18	2	11	108	cycloheptane	3.51	3.63	3	13
39	*m*-dichlorobenzene	3.07	3.13	2	11	109	cyclooctane	4.15	4.16	3	13
40	*p*-dichlorobenzene[d]	3.68	3.14	2	11	110	methylcyclopentane	3.30	3.14	3	13
41	1,2,3-trichlorobenzene	4.17	3.94	2	11	111	methylcyclohexane	3.85	3.77	3	13
42	1,2,4-trichlorobenzene	3.60	3.89	2	11	112	1-cis-2-dimethylcyclohexane[c]	4.27	4.36	3	13
43	1,3,5-trichlorobenzene[c]	4.64	3.88	2	11	113	cyclopentene	2.10	2.01	3	13
44	1,2,3,4-tetrachlorobenzene	4.25	4.71	2	11	114	cyclohexene	2.59	2.67	3	13
45	1,2,3,5-tetrachlorobenzene	4.87	4.65	2	11	115	cycloheptene	3.16	3.25	3	13
46	1,2,4,5-tetrachlorobenzene	4.96	4.66	2	11	116	1-methylcyclohexene	3.27	3.31	3	13
47	pentachlorobenzene	5.48	5.51	2	11	117	1,4-cyclohexadiene	2.06	2.24	3	13
48	hexachlorobenzene[d]	6.78	6.44	2	11	118	4-vinylcyclohexene	3.34	3.38	3	13
49	biphenyl	4.36	3.96	2	11	119	cycloheptatriene	2.17	2.42	3	13
50	2-chlorobiphenyl	4.57	4.65	2	11	120	2-butanone	−0.28	0.17	2	12
51	2,5-dichlorobiphenyl	5.06	5.47	2	11	121	3-pentanone	0.28	0.63	2	12
52	2,6-dichlorophenyl	5.21	5.36	2	11	122	2-heptanone[d]	1.45	1.32	2	12
53	2,4,5-trichlorobiphenyl	6.20	6.19	2	11	123	2-octanone[c]	2.05	1.93	2	12
54	2,4,6-trichlorobiphenyl	6.06	6.14	2	11	124	2-nonanone	2.58	2.50	2	12
55	2,3,4,5-tetrachlorobiphenyl	7.14	6.92	2	11	125	2-decanone	3.30	3.23	2	12
56	2,2′,4′,5-pcb	7.25	6.97	2	11	126	acetal[e]	0.12		2	12
57	2,3,4,5,6-pcb	7.77	7.69	2	11	127	2-furaldehyde	0.09	0.16	2	12
58	2,2′,4,5,5′-pcb	7.23	7.69	2	11	128	methylnonanoate	3.88	3.87	2	12
59	2,2′,3,3′,6,6′-pcb	7.78	8.33	2	11	129	methyldecanoate	4.69	4.70	2	12
60	2,2′,3,3′,4,4′-pcb	9.11	8.32	2	11	130	ethylacetate	0.14	0.34	2	12
61	2,2′,4,4′,6,6′-pcb	8.95	8.29	2	11	131	*n*-propylacetate	0.70	0.69	2	12
62	2,2′,3,3′,4,4′,6-pcb	8.26	8.94	2	11	132	*n*-butylacetate	1.24	1.14	2	12
63	2,2′,3,3′,5,5′,6,6′-pcb	9.04	9.46	2	11	133	ethylpropionate[c]	0.83	0.64	2	12
64	2,2′,3,3′,4,5,5′,6,6′-pcb	10.41	9.86	2	11	134	2-bromoethylacetate	0.67	0.47	2	12
65	decachlorobiphenyl[e]	10.83		2	11	135	1-butanol	0.70	0.69	2	12
66	ethene	2.33	2.51	3	13	136	1-pentanol[d]	0.88	0.70	2	12
67	propene	2.32	2.08	3	13	137	1-hexanol	1.38	1.17	2	12
68	1-butene	2.40	2.31	3	13	138	1-heptanol	1.95	1.72	2	12
69	2-methylpropene	2.33	2.35	3	13	139	1-nonanol	3.13	3.21	2	12
70	1-pentene	2.68	2.71	3	13	140	1,2,3-trimethylbenzene	3.26	3.30	2	12

[a] *S* is the molarity of the solute. [b] See ref 9. [c] Members of the cross-validation set. [d] Members of the prediction set. [e] Compounds that were found to be outliers. [f] The solubilities were predicted using a 9:3:1 computational neural network.

Table 2. List of All the Compounds That Were Found in More Than One of the References

no.	compd name	−log(*S*) used	ref 10	ref 11	ref 12	ref 13
5	*n*-pentane	3.18	3.18		3.25	3.27
8	*n*-hexane	3.84	3.84		3.84	3.95
9	*n*-heptane	4.47	4.47		4.45	4.53
10	*n*-octane	5.12	5.12		5.01	5.23
11	2-methylpentane	3.74	3.74			3.80
13	3-methylpentane	3.68	3.68			3.82
15	2,2-dimethylbutane	3.56	3.56			3.67
17	2,4-dimethylpentane	4.26	4.26			4.39
18	2,2,4-trimethylpentane	4.74	4.74			4.67
20	2,2,5-trimethylhexane	5.38	5.38			5.04
21	benzene	1.65	1.65			1.64
22	toluene	2.21	2.21		2.20	2.52
23	*o*-xylene	2.70	2.70		2.68	2.78
24	*m*-xylene	2.82	2.82		2.82	
25	*p*-xylene	2.76	2.76			
26	ethylbenzene	2.78	2.78		2.75	2.84
37	chlorobenzene	2.58		2.58	2.58	
75	1-hexene	3.08			3.08	3.23
78	1-octene	4.44			4.44	4.62
87	1-pentyne	1.64			1.81	1.64
88	1-hexyne	2.36			2.08	2.36

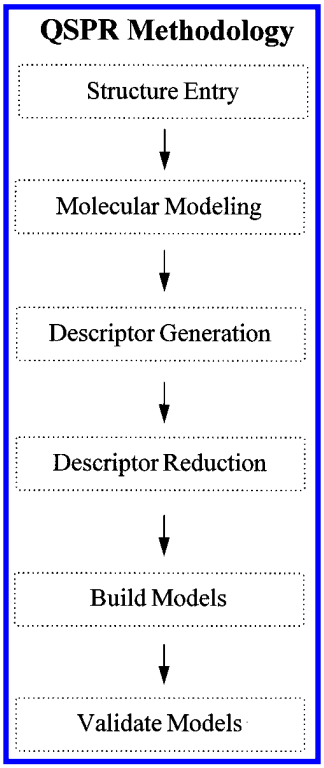

Figure 1. Flow diagram of the ADAPT methodology used to build QSPR models to predict the aqueous solubility.

minima and converge to near global conditions. These techniques used the rms error from multiple linear regression to evaluate the fitness of a subset of descriptors. In each routine, a guided evaluation of thousands of descriptor subsets led to a subset with an rms error close to the error of the global minimum. A separate optimization was performed for each different subset size. If there was no significant improvement in the rms error from a subset size of n to $n + 1$ then the subset size of n was chosen.

The next step of the ADAPT method was to build the mathematical models that would link the aqueous solubility to the small subset of structural descriptors found to be important. The mathematical models were built using either multiple linear regression or computational neural networks. A subset of descriptors was found by the optimization routines based on linear regression, and once a good linear model was found the same descriptors were submitted to a computational neural network.

When used for QSPR development, a computational neural network can be thought of as a nonlinear modeling technique. The computational neural network used in this study is shown in Figure 2, which depicts a fully connected, feed-forward, three-layer neural network.[24] The weights and biases were adjusted using a quasi-Newton algorithm.[25,26] The probability of overtraining the weights and biases was reduced by keeping the number of observations (compounds) to connections (adjustable parameters) below 2.0 and by employing a cross-validation set to monitor training.[27] A more detailed description of the mechanics of the neural network can be found in previous group studies,[25,26] therefore a detailed description will not be given here.

The mathematical models that were developed to relate the aqueous solubility to the structure of the compound had to be validated. The multiple linear regression models were validated by inspecting the residuals, rms errors, and *F*-values. Both the multiple linear regression and computational neural network models were validated by looking at the rms error of the prediction set. The randomly selected prediction set was not used during the descriptor reduction, descriptor selection, or model development. It therefore resembled a true prediction of future unknown compounds. A low rms error of the prediction set was indicative of a high quality model.

After the models were built and before they were validated, it was possible to detect and reject outliers to improve the fit and predictive ability of the model. Outliers were detected using traditional regression diagnostics.[28] Standard statistical values were calculated for each compound including residuals, standardized residuals, studentized residuals, leverage (the weight a point has on the regression equation), and DFITS (the measure of the difference in the estimated value of the *i*th dependent variable when the regression coefficients are recalculated without the *i*th value). If these values exceeded a cutoff value, the compound failed the statistical test. If four of these tests failed for a single compound, that compound was considered a possible outlier.

RESULTS AND DISCUSSION

Numerous multiple linear regression models were created using the genetic algorithm and simulated annealing optimization routines. Both techniques discovered small subsets of descriptors that produced very similar models. The quality was measured by looking at the *F*-value, the rms error of the training set, and the correlation coefficient, *R*. Several models were investigated ranging from four descriptors to ten descriptors. The best model found was a nine descriptor model that had a low rms error and a relatively large *F*-value. The best ten descriptor subset did not significantly improve the rms error of the training set; therefore the nine descriptor model was considered the best model.

The model contained four topological, one electronic, one geometric, and three combination descriptors. The four topological descriptors were an oxygen atom count, a carbon atom count, a weighted path count known as the molecular ID,[29] and the molecular ID divided by the number of atoms

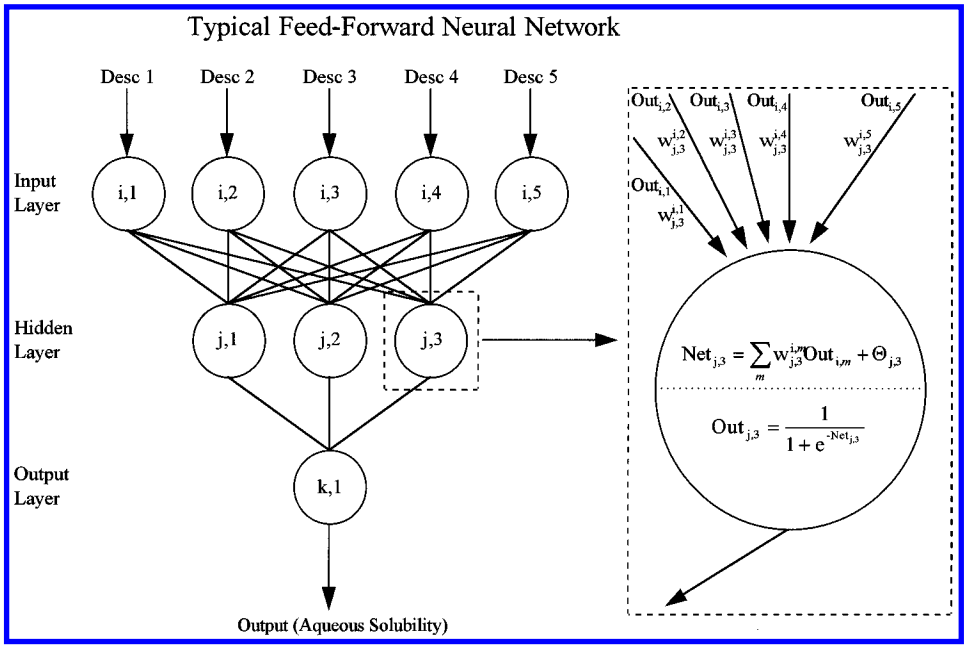

Figure 2. A typical computational neural network architecture.

Table 3. The Best Nine Descriptor Model Found by the Optimization Techniques[a]

descriptor	coefficient	error estimate	explanation
constant	9.33	0.79	Y-intercept
NO	−2.92	0.20	number of oxygens
NC	−0.775	0.047	number of carbons
WTPT 1	0.464	0.016	molecular ID (combines connectivity indexes and path counts)
WTPT 2	−4.93	0.44	molecular ID/ the number of atoms
QSUM	2.13	0.17	sum of the absolute value of the atomic charges
SAAA 1	0.0959	0.011	sum of the surface area of acceptor atoms
SAAA 2	−0.103	9.6×10^{-3}	SAAA 1 divided by the number of acceptor atoms
FNSA 3	33.1	2.6	negative charged partial surface area divided by the total surface area
GEOH	7.33×10^{-4}	1.1×10^{-4}	first divided by the third moment of inertia (length/thickness)

[a] rms = 0.277, R = 0.990, N = 123, F = 607.1.

in the compound. The molecular ID is a descriptor that was developed by Randić that combines features of connectivity indexes and path counts. Each contiguous path in the molecule can be assigned a weight based on the number of atoms adjacent to the atoms in the path. The molecular ID is the summation of all paths found in the compound. The electronic descriptor was the sum of the absolute value of the partial atomic charges of each atom in the compound.[21] The absolute value of the partial atomic charges has proved to be an important descriptor in other QSPR studies as well.[30−32] The geometric descriptor was the length divided by the thickness of the molecule. The three descriptors were a CPSA[20] descriptor and two variations of CPSA descriptors created to account for hydrogen bonding affects. The first combination descriptor found, FNSA 3, is the negative charged partial surface area of the molecule divided by the surface area of the entire molecule. The second combination descriptor found, SAAA 1, is the summation of the surface area of atoms that are capable of accepting hydrogen bonding interactions. The final combination descriptor found, SAAA 2, is the value of SAAA 1 divided by the number of acceptor atoms.

The nine descriptors were used to build a multiple linear regression model for the entire training set (127 compounds), and the rms error was 0.321 log units. The standard statistical diagnostic tools previously mentioned were used to detect possible outliers. Four compounds were detected and rejected using this method, and the training set rms error improved to 0.277 log units. The four outliers found were *m*-cresol (35), nitrobenzene (36), acetal (126), and decachlorobiphenyl (65). Upon inspection of the data set, it was obvious that the first three outliers were poorly represented. The only phenol in the entire data set was *m*-cresol, the only diether in the data set was acetal, and the only compound containing a nitro group was nitrobenzene. Therefore it was not surprising that descriptors were not found that properly encoded these functionalities. The polychlorinated biphenyl (PCB) was not under-represented. However, all the PCBs seemed to have larger errors associated with them than any other class of compounds in the data set.

The final linear model with the four outliers removed is shown in Table 3. The rms error of the training set was 0.277 log units, the correlation coefficient was 0.990, the number of compounds in the training set was 123, and the overall F value was 607.1. The rms error of the prediction set was 0.278 log units, which was similar to the training set error. This was indicative of a sound model that should be able to accurately predict future unknown compounds.

The descriptors that were picked for this model seem to be reasonable selections for the prediction of the aqueous solubility. In fact, it might be possible to categorize the descriptors by the type of intermolecular interactions that the descriptors were representing. For instance, the number of oxygens (NO), sum of the absolute value of the atomic

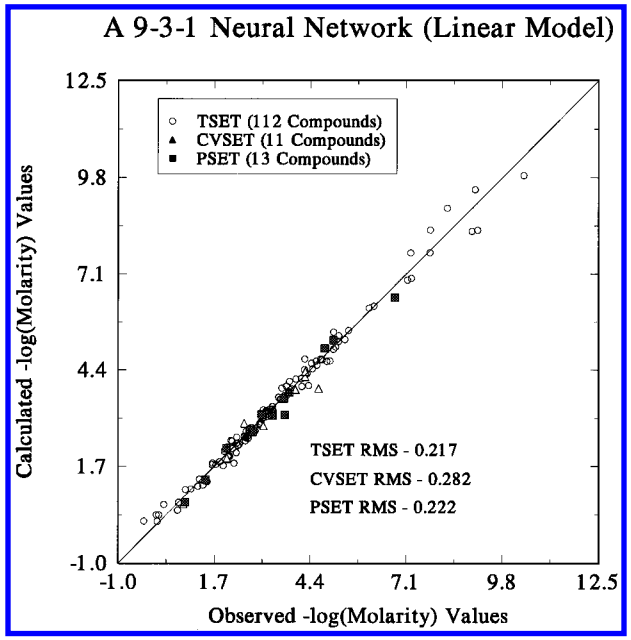

Figure 3. A plot of the calculated versus observed solubility values of the training set, cross-validation set, and prediction set compounds. The aqueous solubilities were calculated using a 9:3:1 computational neural network.

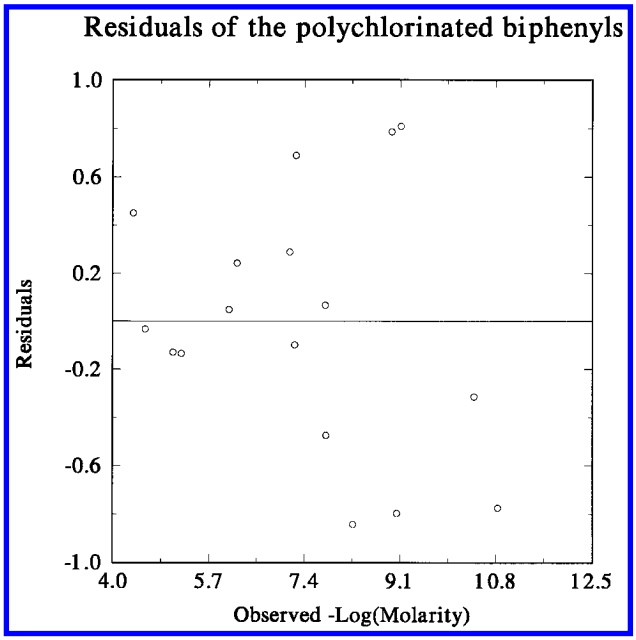

Figure 4. A residual plot of all the polychlorinated biphenyls in the data set. The residuals are the observed aqueous solubility minus the aqueous solubility calculated using the multiple linear regression model.

charges (QSUM), and the negative charged partial surface area divided by the total surface area (FNSA 3) could have been encoding dipole interactions. The number of carbons (NC), the connectivity and path count descriptors (WTPT 1 and 2), and the length divided by the thickness (GEOH) could have been responsible for encoding London dispersion forces. The combination descriptors (SAAA 1 and 2) were undoubtedly encoding the hydrogen bonding interactions between the molecules and solvent.

In an attempt to improve the overall rms error of the model, the set of nine descriptors was fed to a computational neural network. Several three layer computational neural network architectures with varying numbers of hidden layer neurons were investigated, and the computational neural network that had the fewest adjustable parameters and a reasonably low rms error was used. A computational neural network with the nine descriptors from the multiple linear regression model as the input layer, three neurons in the hidden layer, and the calculated aqueous solubility as the output gave rms errors of 0.217, 0.282, and 0.222 for the training set, cross-validation set, and prediction set, respectively. The calculated aqueous solubilities for all the compounds using this neural network model are listed in Table 1. Figure 3 is a plot of the calculated vs observed values of the aqueous solubilities, which were calculated using the computational neural network.

As was previously stated, the PCBs seemed to be problematic in the linear model. The rms error of the PCBs alone in the linear model was 0.510 log units, which was significantly higher than the overall errors in the model. There are several possible reasons that could account for this phenomenon. One possibility is that the PCBs are the less soluble compounds in the data set and lower soluble compounds usually have larger relative errors associated with them, because the measured values are closer to the lower limit of instrument sensitivity.[9] Another possibility is that PCBs have aberrant solution properties,[33] which could cause larger experimental errors. Still another possibility is that since there was no correction for crystallinity in this study, the calculated solubilities could be erroneous because the PDBs are solid compounds. When the PCB residuals in Figure 4 were inspected, the first two possibilities were supported. The residuals would not be randomly distributed if the problem were due to the absence of a correction for crystallinity. The residuals appeared to be greater for the less soluble PCBs (larger−log(molarity)value), which supported the idea that the larger rms error was due to the fact that the PCBs have a lower aqueous solubility than the other compounds in the data set.

Regardless of the reason, the fact remained that the PCBs had a significantly larger error associated with them than the other compounds in the data set. Therefore the PCBs may have influenced the choice of descriptors and the model may not be able to predict future unknown compounds with a great deal of accuracy. Another QSPR model was developed without the PCBs to see if this would improve the rms error of the model and to lend insight on how influential the PCBs were with regard to model development.

The standard ADAPT methodology was repeated except this time the training set of 110 compounds did not include the PCBs. The pool of 144 descriptors was reduced to 51 using the standard descriptor reduction techniques described in the Experimental Section. Several linear regression models were developed using the simulated annealing and genetic algorithm routines, and both found very similar results. A nine descriptor model was found that contained two topological, two electronic, one geometric, and four combination descriptors. The same three outliers (compounds 35, 36, and 126) that were removed in the previous model were detected and rejected using standard statistical procedures. The best model found, including the coefficients, rms error, correlation coefficient, and F value, is shown in Table 4. An explanation of the descriptors used in the model is also given in Table 4. Five of the nine descriptors were also found in the multiple linear regression model that included the PCBs, which supports the idea that although

Table 4. The Best Nine Descriptor Model Found by the Optimization Techniques for the Data Set Minus the PCBs[a]

descriptor	coefficient	error estimate	explanation
constant	−1.60	0.22	Y-intercept
NO[b]	−2.03	0.10	number of oxygens
WTPT 3	0.453	0.019	sum of all path weights starting from heteroatoms
QSUM[b]	3.14	0.010	sum of the absolute value of the atomic charges
DPOL	−0.198	0.045	dipole moment
SAAA 1[b]	0.0183	3.8×10^{-3}	sum of the surface area of acceptor atoms
CHAA 2	8.21	0.46	sum of the charge on acceptor atoms divided by the total molecular surface area
FNSA 3[b]	39.2	2.3	negative charged partial surface area divided by the total surface area
RNCG	5.39	0.45	charge of most negative atom divided by the sum of the negative charges
GEOH[b]	9.11×10^{-4}	9.1×10^{-5}	first divided by the third moment of inertia (length/thickness)

[a] rms = 0.201, $R = 0.987$, $N = 107$, $F = 407.6$. [b] These descriptors were also found in the linear model that included the PCBs.

the PCBs had a large rms error associated with them, they did not seem to greatly influence the model development. The remaining four descriptors that were not common to the previous model were WTPT 3, DPOL, CHAA 2, and RNCG.

The topological descriptor, WTPT 3, used in the model is a connectivity and path weight descriptor that is similar to the information present in WTPT 1 and 2 which were found in the previous model. The electronic descriptor, DPOL,[21] used in the model is the electrical dipole moment of the compound. CHAA 2 is a descriptor that was developed to encode hydrogen bonding and may be encoding similar information as SAAA 2 found in the previous model. The combination descriptor, RNCG, used in the QSPR model is the most negative atom divided by the sum of the negative charges. Although a different set of descriptors was found for the data set that did not contain the PCBs, the information contained in the descriptors was similar to the previous model. However, there was about a 25% improvement in the rms error of the entire model using the new descriptors. The rms errors were 0.201 and 0.197 log units for the training set and prediction set, respectively. The nine descriptors were fed to a 9:3:1 computational neural network, and the rms errors were 0.145, 0.151, and 0.166 for the training set, cross-validation set, and prediction set, respectively. Figure 5 is a plot of the observed solubility values versus the solubility values calculated by this computational neural network.

In order to determine how influential the PCBs were with regard to model development, the nine descriptors from the previous model (Table 3) were used to regress the training set that did not contain the PCBs. If this resulted in a bad model, it could be concluded that the PCBs greatly influenced descriptor selection and that the model could not be used to predict future compounds accurately. The rms errors from the multiple linear regression model were 0.226 and 0.242 log units for the training set and prediction set, respectively, which was very close to the rms errors of 0.201 and 0.197 from the previous model. This implied that most of the improvement in the rms error was a product of removing the less soluble PCBs with aberrant solution properties and not a product of finding a unique set of descriptors. Therefore it was concluded that although the rms error of the first QSAR was higher, it is a good sound model that will accurately predict future compounds.

CONCLUSION

In this study it was shown that accurate QSPR models can be developed to predict the aqueous solubility of a compound. The models were based solely on information derived from the structure of the molecule. No experimental values, such as the melting point, were used in the models. An approach such as this could have certain advantages in fields like drug design where many different combinations of compounds are investigated. The aqueous solubility could be estimated before the molecule was even synthesized.

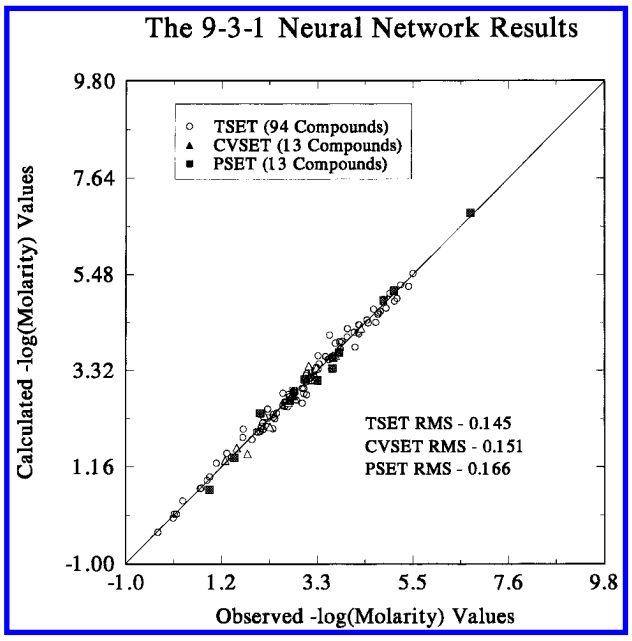

Figure 5. A plot of the calculated versus observed solubility values of the training set, cross-validation set, and prediction set compounds for the data set that did not contain the PCBs. The aqueous solubilities were calculated using a 9:3:1 computational neural network.

Two models were developed in this study, one for all the compounds in the data set and one for all the compounds minus the PCBs. There was a significant improvement in the overall rms error of the model that did not contain the PCBs. The improvement was attributed mostly to the removal of the PCBs which should have a larger relative error associated with them due to their low solubility and strange solution properties. The model that was developed for the entire data set should be able to accurately predict future unknown compounds that are similar in structure to the compounds found in the training set. The model could be improved by supplementing the data set with more compounds, especially from the groups that were underrepresented and removed as outliers (diethers, phenols, and nitrogen containing compounds).

REFERENCES AND NOTES

(1) Bodor, N.; Huang, M. A new method for the estimation of the aqueous solubility of organic compounds. *J. Pharm. Sci*. **1992**, *81*, 954−960.

(2) Patil, G. S. Correlation of aqueous solubility and octanol−water partition coefficient based on molecular structure. *Chemosphere* **1991**, *22*, 723−738.

(3) Patil, G. S. Prediction of aqueous solubility and octanol-water partition coefficient for pesticides based on their molecular structure. *J. Hazard. Mater*. **1994**, *36*, 35−43.

(4) Yalkowsky, S. H. Estimation of the aqueous solubility of complex organic compounds. *Chemosphere* **1993**, *26*, 1239−1261.

(5) 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.

(6) Nelson, T. M.; Jurs, P. C. Prediction of aqueous solubility of organic compounds. *J. Chem. Inf. Comput Sci*. **1994**, *34*, 601−609.

(7) Lyman, W.; Reehl, W.; Rosenblatt, D. *Handbook of Chemical Property Estimation Methods*; American Chemical Society: Washington, DC, 1990; Chapter 2.

(8) Isnard, P.; Lamber, S. Aqueous solubility and n-octanol/water partition coefficient correlations. *Chemosphere* **1989**, *18*, 1837−1853.

(9) Heller, S. R.; Bigwood, D. W.; May, W. E. Expert systems for evaluating physicochemical property values. 1. Aqueous solubility. *J. Chem. Inf. Comput. Sci*. **1994**, *34*, 627−636.

(10) Polak, J.; Lu, B. Mutual solubilities of hydrocarbons and water at 0 and 25 °C. *Can. J. Chem*. **1973**, *51*, 4018−4023.

(11) Miller, M. M.; Ghodband, S.; Wasik, S. P.; Tewari, Y. B.; Martire, D. E. Aqueous solubilities, octanol/water partition coefficients, and entropies of melting of chlorinated benzenes and biphenyls. *J. Chem. Eng. Data* **1984**, *29*, 184−190.

(12) Tewari, Y. B.; Miller, M. M.; Wasik, S. P.; Martire, D. E. Aqueous solubility and octanol/water partition coefficient of organic compounds at 25.0 °C. *J. Chem. Eng. Data* **1992**, *27*, 451−454.

(13) McAuliffe, C. Solubility in water of paraffin, cycloparaffin, olefin, acetylene, cycloolefin, and aromatic hydrocarbons. *J. Phys. Chem*. **1966**, *70*, 1267−1275.

(14) Yalkowsky, S. H.; Banerjee, S. *Aqueous Solubility. Methods of Estimation for Organic Compounds*; Marcel Dekker, Inc: New York, 1992; Chapter 5.

(15) May, W. E.; Wasik, S. P. Determination of the solubility behavior of some polycyclic aromatic hydrocarbons in water. *Anal. Chem*. **1978**, *50*, 997−1000.

(16) Stuper, A. J.; Brugger, W. E.; Jurs, P. C. *Computer-Assisted Studies of Chemical Structure and Biological Function*; Wiley Interscience: New York, 1979.

(17) Jurs, P. C.; Chou, J. T.; Yuan, M. In *Computer-Assisted Drug Design*; Olsen, E. C., Christoffersen, R. E., Eds.; The American Chemical Society: Washington, DC, 1979; pp 103−129.

(18) Stewart, J. P. P. Mopac 6.0, Quantum Chemistry Program Exchange; Indiana University, Bloomington, IN, Program 455.

(19) Stewart, J. P. P. Mopac: A semi-empirical molecular orbital program. *J. Comput.-Aided Mol. Des*. **1990**, *4*, 1.

(20) Stanton, D. T.; Jurs, P. C. Development and use of charged partial surface area structural descriptors in computer-assisted quantitative structure-property relationship studies. *Anal. Chem*. **1990**, *62*, 2323.

(21) Dixon, S. L. Ph.D. Thesis, The Pennsylvania State University, Aug 1994, Chapter 4.

(22) Luke, B. T. Evolutionary programming applied to the development of quantiative structure−activity relationships and quantitative structure−property relationships. *J. Chem. Inf. Comput. Sci*. **1994**, *34*, 1279−1287.

(23) Sutter, J. M.; Jurs, P. C. In *Data Handling in Science and Technology (Vol 15). Adaption of Simulated Annealing to Chemical Optimization Problems;* Kalivas, J. H., Ed.; Elsevier: Amsterdam, 1995; Chapter 5.

(24) Xu, L.; Ball, J. W.; Dixon, S. L.; Jurs, P. C. Quantitative structure-activity relationships for toxicity of phenols using regression analysis and computational neural networks. *Environmental Toxicol. Chem*. **1994**, *13*, 841−851.

(25) Wessel, M. D.; Jurs, P. C. Prediction of reduced ion mobility constants from structural information using multiple linear regression analysis and computational neural networks. *Anal. Chem*. **1994**, *66*, 2480−2487.

(26) Sutter, J. M.; Dixon, S. L.; Jurs, P. C. Automated descriptor selection for quantitative structure-activity relationships using generalized simulated annealing. *J. Chem. Inf. Comput. Sci*. **1995**, *35*, 77−84.

(27) Livingstone, D. J.; Manallack, P. T. Statistics using neural networks: chance effects. *J. Med. Chem*. **1993**, *36*, 1295−1297.

(28) Belsey, D. A.; Kuh, E.; Welsch, R. E. *Regression Diagnostics*; New York, 1980.

(29) Randić, M. On molecular identification numbers, *J. Chem. Inf. Comput. Sci*. **1984**, *24*, 164−175.

(30) Buydens, L.; Massart, D. Prediction of gas chromatographic retention indexes with topological, physicochemical, and quantum chemical parameters. *Anal. Chem*. **1983**, *55*, 738−744.

(31) Buydens, L.; Massart, D.; Geerlings, P. Relationship between gas chromatographic behavior and topological, physicochemical, and quantum chemically calculated charge parameters for neuroliptica. *J. Chrom. Sci*. **1985**, *23*, 304−307.

(32) Collantes, E.; Dunn III, W. Amino acid side chain descriptors for quantitative structure-activity relationship studies of peptide analogues. *J. Med. Chem*. **1995**, *38*, 2705−2713.

(33) Dunnivant, F.; Coates, J.; Elzerman, A. Experimentally determined Henry's law constants for 17 polychlorobiphenyls congeners. *Environ. Sci. Technol*. **1988**, *22*, 448−453.

CI9501507