Prediction of Aqueous Solubility of Organic Compounds

Todd M. Nelson and Peter C. Jurs*

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

Received August 25, 1993®

A set of mathematical models is developed for predicting the aqueous solubility of organic compounds from their structures. The structures are represented by topological, geometrical, and electronic descriptors. The solubilities are given as $\log(1/S)$, where S is in moles per liter. Successful nine-variable regression models are reported for three sets of compounds—hydrocarbons, halohydrocarbons, alcohols and ethers—with standard errors of 0.17 log units, and a fourth model is reported for the combined set of all compounds with a standard error of 0.37 log units.

INTRODUCTION

Aqueous solubility is a particularly useful parameter in many applications. It has many uses in the pharmaceutical, environmental, and other chemical sciences. It is key in understanding drug transport and environmental impact, and it also has uses in the development of analytical methods. Accordingly, experimental aqueous solubility values are valuable in the aforementioned areas. However, experimental values are not always available, making the ability to predict aqueous solubility very useful. There are several methods that have been proposed for performing this function.[1] Melting point, log P (P is the partition coefficient of a solute between 1–octanol and water), and other experimental parameters have been used as predictors of aqueous solubility. There are also various group contribution methods, as well as the approach of quantitative structure–activity relationships (QSAR).[2]

Although group contribution methods can yield accurate estimated solubility values in many cases, these methods have some shortcomings. The groups that are included must be defined in advance, so values for a new compound containing any new group cannot be estimated. They are entirely topological and do not take into account geometric information. They do not take into account the nature of the bonding of groups, that is, their proximity, as in substitutional isomers. Also, the major problems with estimations of solubility based upon other experimental data are the availability and the quality of the those experimental parameters. QSARs based on parameters that can be derived directly from the molecular structure avoid many of these problems, especially since many of the molecular structure descriptors developed for QSAR are mathematical relationships and are known exactly.

The molecular properties that affect aqueous solubility the most are the size, shape, and polarity of the molecule. These features of a compound can be represented numerically in many ways such as molecular weight, surface area, dipole moment, and other descriptors. The primary objective of this study is to find suitable descriptors to represent the molecules being investigated and to develop mathematical models for aqueous solubility.

METHODOLOGY

The Automated Data Analysis and Pattern recognition Toolkit (ADAPT) software system was used to develop the QSARs for this study.[3] ADAPT consists of a series of interactive programs that allow structure entry, molecular modeling, descriptor generation, descriptor analysis, model generation, and model validation. ADAPT and the methodology that is used in conjunction with it have proven useful in many different types of studies including structure-retention studies, structure–property studies, ^{13}C NMR spectral simulation, and pattern recognition studies, as well as structure–activity studies.

To start a study once the data have been obtained, the molecular structures and the dependent variable are entered and stored. Structures can be entered either by using a graphics terminal or by uploading from preexisting mol files generated off-line. These two-dimensional structures, that are stored as connection tables, are converted to three-dimensional structures with molecular mechanics modeling and then stored with their Cartesian coordinates. These three-dimensional models can be improved upon, if necessary, by more rigorous, semiempirical molecular orbital calculation programs.

Descriptor generation follows. Three types of descriptors are available: topological, geometric, and electronic. The topological descriptors include various fragment and atom counts, substrate counts, and molecular connectivity indices.[4] The geometric descriptors include molecular volume, surface area,[5] cross-sectional areas, and moments of inertia. The electronic descriptors include Del Re sigma charges,[6] partial atomic charges,[7] and extended Hückel calculations.[8,9] In addition there is a subclass of descriptors known as charged partial surface area descriptors (CPSA) which involve a combination of partial atomic charges with solvent accessible surface area.[10]

Following descriptor calculation, objective feature selection is used to reduce the number of descriptors to a manageable number. There are a number of methods for reducing the number of descriptors. The first is to eliminate those descriptors which have no information or are redundant, such as descriptors with large numbers of identical values. It is also necessary to eliminate descriptors that are correlated highly with one another. One also needs to look for and eliminate multicollinearities. This can be done by performing multiple linear regression of each descriptor against all of the others. Or it can be done using vector space descriptor analysis which makes use of Gram–Schmidt orthogonalization to find those descriptors which contain the most useful information.[11]

From the remaining descriptors, models are generated using multiple linear regression analysis.[12] Regression analysis can be done using methods such as leaps-and-bounds regression,

® Abstract published in *Advance ACS Abstracts,* March 1, 1994.

forward stepwise regression, and progressive deletion, or it can be done interactively.[13] For this study, leaps-and-bounds regression and interactive regression were used to do the majority of the regression analysis.

Finally, models must be validated. This can be accomplished through a combination of the following methods: outlier detection, jackknifing, variance decomposition, or employing an external prediction set. For this study some rather intensive outlier detection and an internal validation method were used to determine the validity of the models. Outlier detection involved using data diagnostics generation,[13,14] robust regression analysis,[15] and duplexing. The internal validation consisted of using a leave-*n*-out method, which is a more generalized version of jackknifing.[16] In addition, calculated versus observed plots and residual plots can be useful qualitative guides in judging the validity of a model.

DATA SET

The data for this study were obtained from three sources: a paper by Bodor and Huang,[17] a paper by Suzuki,[18] and the Solubility Data Series.[19] The two papers are studies that also attempt to develop accurate models for predicting aqueous solubility. The Bodor and Huang paper presents a QSAR based on a diverse set of compounds ranging from alkanes to steroids. The Suzuki paper presents an algorithm for predicting aqueous solubility based on two pathways, one using log *P* and the other using a group contribution method for a set of compounds including aliphatic and aromatic hydrocarbons, ethers, alcohols, halogenated hydrocarbons, amines, and thiols. The Solubility Data Series is a database of critiqued solubility data that presents values for many different kinds of solubility for many thousands of compounds, inorganic and organic, in many different solvents.

The compounds that were used in this study are hydrocarbons, both aliphatic and aromatic, halogenated hydrocarbons, and ethers and alcohols. Table 1 shows the compounds and their corresponding experimental aqueous solubilities as $\log(1/S)$, where S is the solubility in moles per liter. The 123 hydrocarbons are listed in order of molecular weight, followed by the 80 halogenated hydrocarbons, and followed by the 97 ethers and alcohols. The data presented for those compounds which are miscible in water at 25 °C are actually S_g/K_{gw}, wherer S_g is the solute molar concentration of the compound in its own saturated vapor and K_{gw} is the gas–water partition coefficient.[20] The experimental values of $\log(1/S)$ for those 300 compounds range from −1.26 for methanol to 7.67 for dodecane. The structures of some of the compounds in this data set are shown in Figure 1 to demonstrate the diversity of the data set.

In the majority of the cases, the aqueous solubility values reported for the compounds forming the three data sets agreed fairly well. In fact, it was only the hydrocarbon compounds that showed any substantial degree of disagreement. The halogenated and oxygen-containing compounds all showed a surprising degree of agreement, which suggests that the data for these subsets all came from the same original source. However, for the hydrocarbon subset, there were a substantial number of discrepancies between the three sources. Since only the Solubility Data Series documented the original sources of the data well, it is difficult to know whether or not the data which agree among the three sources is from the same original source.

Prior to model development, it was necessary to reconcile the worst of the discrepancies in the aqueous solubility values. The experimental errors for this data set range between 0.03 and 0.10 log units.[19] Most of the discrepancies between the sources are about 0.06 log units or less, which is approximately the magnitude of the experimental error. Thus, these discrepancies were ignored for the most part. However, the larger errors need to be reconciled. Since these large discrepancies were only present in the hydrocarbon subset, the following procedure was applied to that set of compounds.

The general principles used to decide which value to use for the dependent variable were as follows. If two out of three sources agreed, then the value corresponding to the two which were in agreement was used. If there were only two sources, and those did not agree, or all three sources disagreed, the value from the Solubility Data Series was used, since the values in that database have been critically evaluated. If there was no value from the Solubility Data Series, then chemical intuition had to be used to make a choice between the disagreeing values. An example that was actually encountered involves the solubility of decanol, which should be greater than the solubility of decane, due to the presence of the oxygen which would create polar interactions with the water. Though the difference is not great, one would not expect the solubility of decanol to be less than that of decane.

The data set was analyzed from two different perspectives. First, the overall data set was divided into three subsets that were analyzed individually: the hydrocarbons (123 compounds), the halogenated hydrocarbons (80 compounds), and the ethers and alcohols (97 compounds). Second, the 300 compound data set was analyzed as a whole.

STRUCTURE ENTRY

The structures for this study were entered by sketching their structures. The structures were stored as connection tables of atom types, bond types, and atom connections. The initial three-dimensional structures were generated using a simple classical mechanics modeling routine. They were further refined using MOPAC,[21] since many of the halogenated compounds could not be modeled successfully using Allinger's MM2.[22] In addition, it was determined that many of the structures could not be properly modeled by classical methods. This was the situation for many of the halogenated and oxygen-containing compounds.

DESCRIPTOR GENERATION

Topological, geometric, and electronic descriptors were calculated for each of the compounds in this study. These descriptors are numerical representations of various structural aspects of the compounds.

The topological descriptors that were calculated included shape indices, weighted and unweighted path indices, and fragment counts. These descriptors are calculated from the two-dimensional connection table and are independent of the three-dimensional geometry. Various connectivity indices and path counts were important in the development of the final models in this study.

The geometric descriptors that were calculated included surface area and volume and the various moments of inertia and geometric moments. These descriptors are dependent on the three-dimensional models that are generated. Several of the moment of inertia and geometric moment descriptors were consistently present in the final models. In addition, many of the charged partial surface area (CPSA) descriptors were found to be important.

DESCRIPTOR ANALYSIS

A total of 157 descriptors were calculated. The majority of these descriptors were eliminated using objective feature

Table 1. Compound Names, Observed Aqueous Solubilities, and Estimated Aqueous Solubilities

			estd aqueous solubilities					estd aqueous solubilities	
no.	name	obsd value	for given subset	for overall data set	no.	name	obsd value	for given subset	for overall data set
				Hydrocarbon Subset					
1[a]	methane	2.820	1.345	−1.597	63	*n*-heptane	4.620	4.509	4.011
2[a]	ethyne	−0.290	0.228	−5.851	64[a]	2-methylhexane	4.600	4.393	3.974
3[a]	ethene	2.330	1.144	0.510	65	3-methylhexane	4.580	4.361	3.984
4[a]	ethane	2.730	1.812	0.998	66	2,2-dimethylpentane	4.360	4.278	3.998
5	propyne	0.410	0.575	0.734	67	2,3-dimethylpentane	4.280	4.245	3.959
6[a]	propene	2.030	1.755	1.330	68	2,4-dimethylpentane	4.390	4.222	4.040
7[a]	cyclopropane	1.070	1.525	2.009	69	3,3-dimethylpentane	4.230	4.120	3.947
8	propane	2.820	2.272	1.997	70	4-vinylcyclohexene	3.340	3.417	3.186
9[a]	butadiyne	0.440	1.037	−5.392	71	1-octyne	3.610	3.633	3.757
10[a]	1-buten-3-yne	1.090	0.079	1.366	72	2,2-dimethyl-3-hexyne	3.030	3.153	3.986
11[a]	1-butyne	0.880	1.096	1.663	73	1-octene	4.620	4.720	3.968
12[a]	1,3-butadiene	1.870	1.332	1.731	74	cyclooctane	4.150	4.155	3.966
13[a]	1-butene	2.400	2.330	1.983	75	1,2-dimethylcyclohexane	4.270	4.295	4.072
14	*cis*-2-butene	1.930	2.053	2.295	76	1,4-dimethylcyclohexane	4.470	4.553	4.109
15	*trans*-2-butene	2.040	2.079	2.312	77	1,1,3-trimethylcyclopentane	4.480	4.654	4.467
16	2-methylpropene	1.990	2.334	1.969	78	propylcyclopentane	4.740	4.596	4.161
17	*n*-butane	2.910	2.919	2.615	79	*n*-octane	5.220	5.275	4.527
18	isobutane	3.040	2.788	2.616	80	3-methylheptane	5.160	4.899	4.418
19[a]	1,2-cyclopentadiene	1.990	0.969	2.139	81	2,2,4-trimethylpentane	4.710	4.688	4.591
20	1-pentyne	1.640	1.737	2.229	82	2,3,4-trimethylpentane	4.800	4.605	4.472
21	1,4-pentadiene	2.080	1.980	2.189	83	1,8-nonadiyne	2.980	2.901	3.696
22	cyclopentene	2.100	2.010	2.404	84	1-nonyne	4.260	4.262	4.305
23	2-methyl-1,3-butadiene	2.030	1.817	2.220	85	2,2,5-trimethyl-3-hexyne	3.510	3.732	4.412
24	1-pentene	2.670	2.930	2.501	86	2,2,5-trimethylhexane	5.050	5.124	4.687
25	2-pentene	2.540	2.366	2.760	87[a]	2,2,5,5-tetramethyl-3-hexyne	3.690	4.640	4.810
26	cyclopentane	2.650	2.796	3.036	88[d]	benzene	1.650	1.862	2.473
27	2-methyl-1-butene	2.730	2.641	2.453	89	toluene	2.290	2.164	2.895
28	2-methyl-2-butene	2.560	2.421	2.794	90	styrene	2.620	2.441	2.616
29	3-methyl-1-butene	2.730	2.807	2.469	91	ethylbenzene	2.800	2.954	3.163
30	*n*-pentane	3.270	3.374	3.100	92	*o*-xylene	2.790	2.695	3.349
31	neopentane	3.340	3.078	3.012	93	*m*-xylene	2.830	2.942	3.446
32	2-methylbutane	3.180	3.230	3.046	94	*p*-xylene	2.770	3.036	3.416
33[a]	1,4-cyclohexadiene	1.930	2.003	2.624	95	indan	3.080	2.986	3.607
34	1-hexyne	2.360	2.334	2.749	96	propylbenzene	3.340	3.566	3.562
35[a]	3-hexyne	1.990	2.352	3.124	97	1,2,3-trimethylbenzene	3.260	3.192	3.859
36	1,5-hexadiene	2.690	2.832	2.674	98	i-propylbenzene	3.380	3.375	3.556
37[a]	cyclohexene	2.580	2.765	2.858	99	1,2,4-trimethylbenzene	3.320	3.620	3.906
38	2,3-dimethyl-1,3-butadiene	2.400	2.190	2.664	100[a]	1,3,5-trimethylbenzene	3.390	4.050	4.048
39	1-hexene	3.230	3.528	2.977	101	*n*-butylbenzene	3.960	4.203	4.034
40	2-hexene	3.100	3.040	3.198	102	*sec*-butylbenzene	3.980	3.942	3.925
41	cyclohexane	3.180	3.362	3.401	103	*tert*-butylbenzene	3.620	3.743	4.065
42	methylcyclopentane	3.300	3.483	3.291	104[a]	*p*-cymeme	3.760	4.279	4.108
43	2-methyl-1-pentene	3.030	3.289	2.986	105	1-methylnaphthalene	3.710	3.360	3.783
44	4-methyl-1-pentene	3.240	3.301	2.899	106	*tert*-amylbenzene	4.150	4.290	4.450
45	*n*-hexane	3.960	4.085	3.560	107	1-ethylnaphthalene	4.190	4.192	4.025
46	2-methylpentane	3.790	3.794	3.557	108	2-ethylnaphthalene	4.290	4.177	4.116
47	3-methylpentane	3.830	3.754	3.528	109	1,3-dimethylnaphthalene	4.290	4.216	4.364
48	2,2-dimethylbutane	3.670	3.557	3.424	110	1,4-dimethylnaphthalene	4.140	4.406	4.322
49	2,3-dimethylbutane	3.610	3.714	3.418	111	1,4,5-trimethylnaphthalene	4.920	4.833	4.802
50[a]	2,5-norbornadiene	1.030	1.994	2.822	112[a]	1,1,3-trimethylcyclohexane	4.850	5.022	4.789
51[a]	cycloheptatriene	1.160	1.384	2.728	113[a]	1,1,4-trimethylcyclohexane	5.220	4.881	4.561
52	1,6-heptadiyne	1.750	1.515	2.615	114[a]	4-methyloctane	6.050	5.551	4.959
53	1-heptyne	3.010	3.011	3.230	115	*n*-nonane	5.880	5.736	5.085
54[a]	2-heptyne	2.640	2.749	3.481	116	*p*-mentha-1,8-diene	3.990	4.205	4.171
55[a]	2-methyl-3-hexyne	2.590	3.072	3.625	117[a]	decalin	5.190	5.163	4.875
56	1,6-heptadiene	3.340	3.306	3.106	118	pentylcyclopentane	6.080	5.623	5.039
57	cycloheptene	3.180	3.045	3.202	119[a]	1-decene	4.390	5.932	5.137
58	1-methylcyclohexene	3.270	3.166	3.411	120[a]	decane	6.980	6.440	5.725
59[a]	1-heptene	3.550	4.144	3.460	121[a]	2-methyldecalin	6.570	5.884	5.226
60	2-heptene	3.820	3.738	3.727	122[a]	*n*-undecane	7.590	6.903	6.451
61	cycloheptane	3.510	3.731	3.704	123	*n*-dodecane	7.670	7.654	7.297
62	methylcyclohexane	3.790	3.912	3.644					
				Halogenated Hydrocarbon Subset					
124	tetrachloromethane	2.280	1.850	2.080	134	1,1-dichlorotetrafluoroethane	2.930	2.729	2.353
125	bromoform	1.900	1.877	2.155	135	1,2-dichlorotetrafluoroethane	2.740	2.854	2.653
126	chloroform	0.920	1.048	1.711	136[b]	tetrafluoroethene	1.600	0.934	1.932
127	dibromomethane	1.180	1.060	1.200	137	1,1,2,2-tetrachlorodifluoroethane	3.190	3.317	2.645
128	bromochloromethane	1.160	0.838	0.784	138	1,1,2-trichlorotrifluoroethane	3.040	3.062	2.409
129	dichloromethane	0.630	0.731	0.844	139	trichloroethene	1.950	1.915	1.657
130	diiodomethane	2.340	2.023	1.764	140	pentachloroethane	2.610	2.780	2.721
131	iodomethane	1.000	0.830	−0.077	141	2-bromo-2-chloro-1,1,1-trifluoroethane	1.700	2.087	2.126
132	tetrachloroethene	2.530	2.580	2.657	142	1,1,2,2-tetrabromoethane	2.730	2.766	2.996
133	chloropentafluoroethane	2.790	2.481	2.459	143	*cis*-1,2-dichloroethene	1.100	1.272	0.885

Table 1 (Continued)

			estd aqueous solubilities					estd aqueous solubilities	
no.	name	obsd value	for given subset	for overall data set	no.	name	obsd value	for given subset	for overall data set
				Halogenated Hydrocarbon Subset					
144	*trans*-1,2-dichloroethene	1.190	1.361	1.518	174	1-chloro-2-methylpropane	2.000	2.044	1.983
145	1,1,2,2-tetrachloroethane	1.760	1.852	1.888	175	iodobutane	2.960	2.923	3.040
146	2-chloro-1,1,1-trifluoroethane	1.150	1.109	1.553	176	1-bromo-3-methylbutane	2.890	2.588	3.011
147	1,1,1-trichloroethane	2.010	1.926	1.787	177	1-chloropentane	2.730	2.813	2.589
148	1,1,2-trichloroethane	1.460	1.567	1.836	178	2-chloropentane	2.630	2.676	2.622
149	1-chloro-1,1-difluoroethane	1.200	1.224	0.931	179	3-chloropentane	2.630	2.595	2.494
150	1,2-dibromoethane	1.640	1.513	2.227	180	1,2,4-trichlorobenzene	3.720	3.604	3.327
151	1,1-dichloroethane	1.290	1.482	1.454	181	1,2-dibromobenzene	3.500	3.481	3.316
152	1,2-dichloroethane	1.040	1.221	1.801	182	1,3-dibromobenzene	3.380	3.635	4.049
153	1-bromo-2-chloroethane	1.320	1.319	1.752	183	2-bromochlorobenzene	3.190	3.199	2.777
154	1-chloro-2-fluoroethane	0.510	0.736	0.983	184	3-bromochlorobenzene	3.210	3.137	3.278
155	1,1-difluoroethane	0.570	0.517	0.884	185	1,2-dichlorobenzene	3.010	2.952	2.520
156	bromoethane	1.060	1.102	0.978	186	1,3-dichlorobenzene	3.080	2.841	2.903
157	iodoethane	1.280	1.534	1.446	187	2-chloroiodobenzene	3.540	3.536	2.975
158	3-chloropropene	1.600	1.144	1.346	188	3-chloroiodobenzene	3.550	3.616	3.539
159	1,2-dibromopropane	2.140	2.132	2.285	189	1,2-difluorobenzene	2.000	1.977	1.579
160	1.3-dibromopropane	2.080	2.190	2.684	190	1,3-difluorobenzene	2.000	1.891	1.897
161	1,2-dichloropropane	1.610	1.770	2.279	191	1,4-difluorobenzene	1.970	1.932	2.277
162	1,3-dichloropropane	1.610	1.705	1.787	192	bromobenzene	2.550	2.598	2.845
163	1-bromopropane	1.730	1.700	1.814	193	chlorobenzene	2.360	2.335	2.433
164	2-bromopropane	1.630	1.664	1.718	194	fluorobenzene	1.790	1.838	1.624
165	1-chloropropane	1.530	1.555	1.455	195	iodobenzene	2.770	2.999	2.715
166	2-chloropropane	1.360	1.665	1.437	196[d]	2-chlorophenol	1.050	0.918	–0.041
167	1-iodopropane	2.290	2.167	2.257	197	α-chlorotoluene	2.430	2.485	2.360
168	2-iodopropane	2.090	2.115	2.199	198	α,α,α-trifluorotoluene	2.510	2.681	2.107
169	hexachloro-1,3-butadiene	4.910	4.901	4.053	199	1-bromo-2-ethylbenzene	3.670	3.790	3.478
170	1,1-dichlorobutane	2.400	2.477	2.539	200	1-bromo-2-propylbenzene	4.190	4.328	3.816
171	1-bromobutane	2.370	2.353	2.503	201	2,4-dichlorobiphenyl	5.200	5.432	4.692
172	1-bromo-2-methylpropane	2.430	2.137	2.386	202	2,5-dichlorobiphenyl	5.590	5.394	4.910
173	1-chlorobutane	2.140	2.155	2.111	203[d]	3-chlorobiphenyl	5.160	4.746	4.202
				Oxygen-Containing Compound Subset					
204	divinyl ether	0.960	1.001	1.395	246	2-methyl-1-pentanol	1.110	0.974	0.736
205	tetrahydrofuran	–0.620	–0.910	–0.549	247	4-methyl-1-pentanol	1.140	1.056	0.831
206	diethyl ether	0.060	0.194	0.037	248	2-ethyl-1-butanol	1.010	0.802	0.543
207	methyl *n*-propyl ether	0.370	0.186	0.095	249	2,2-dimethyl-1-butanol	0.910	0.646	0.523
208	methyl isopropyl ether	0.030	0.034	–0.010	250	1-hexanol	1.210	1.199	0.887
209	cyclopropyl vinyl ether	1.100	0.740	0.841	251	2-hexanol	0.870	0.947	0.692
210	cyclopropyl ethyl ether	0.640	0.521	0.269	252	3-hexanol	0.800	0.922	0.865
211	tetrahydropyran	–0.050	–0.078	0.246	253	2-methyl-2-pentanol	0.490	0.552	0.734
212	2-methyltetrahydrofuran	–0.310	–0.031	0.277	254	2-methyl-3-pentanol	0.700	0.730	0.774
213	3-methyltetrahydrofuran	–0.090	–0.064	0.373	255	3-methyl-2-pentanol	0.710	0.753	0.465
214	methyl *n*-butyl ether	0.990	0.958	0.889	256	3-methyl-3-pentanol	0.360	0.486	0.676
215	methyl isobutyl ether	0.900	0.766	0.785	257	4-methyl-2-pentanol	0.790	0.889	0.731
216	methyl *sec*-butyl ether	0.730	0.731	0.785	258	2,3-dimethyl-2-butanol	0.370	0.430	0.508
217	methyl *tert*-butyl ether	0.210	0.433	0.608	259	2,3-dimethyl-1-butanol	0.370	0.849	0.604
218	ethyl *n*-propyl ether	0.670	0.899	0.850	260	3,3-dimethyl-1-butanol	0.500	0.779	0.714
219	ethyl isopropyl ether	0.550	0.764	0.787	261	3,3-dimethyl-2-butanol	0.610	0.516	0.471
220[c]	diallyl ether	0.020	1.012	1.243	262	*m*-cresol	0.660	0.649	0.673
221	*n*-propyl ether	1.320	1.437	1.523	263[c]	benzyl alcohol	0.450	1.724	–0.050
222	*n*-propyl isopropyl ether	1.340	1.378	1.500	264	1-heptanol	1.810	1.687	1.533
223[c]	isopropyl ether	1.700	1.374	1.475	265	2-methyl-2-hexanol	1.070	1.179	1.381
224[d]	anisole	2.880	2.873	1.382	266	3-methyl-3-hexanol	0.980	1.005	1.383
225	*n*-butyl ether	2.770	2.524	2.979	267	3-ethyl-3-pentanol	0.830	0.972	1.423
226[c]	methanol	–1.260	–2.406	–3.234	268	2,2-dimethyl-3-pentanol	1.150	1.054	1.353
227	ethanol	–1.100	–1.453	–2.162	269	2,3-dimethyl-2-pentanol	0.870	1.015	1.326
228	propanol	–0.620	–0.705	–1.290	270	2,3-dimethyl-3-pentanol	0.840	0.928	1.231
229	1-butanol	–0.030	–0.001	–0.498	271	2,4-dimethyl-2-pentanol	0.930	1.136	1.493
230	2-butanol	–0.470	–0.184	–0.391	272	2,4-dimethyl-3-pentanol	1.220	1.229	1.379
231	2-methyl-1-propanol	–0.100	–0.187	–0.667	273	2-heptanol	1.550	1.524	1.463
232	1-penten-3-ol	–0.020	0.119	0.141	274	3-heptanol	1.440	1.480	1.498
233	3-penten-2-ol	–0.060	–0.097	–0.040	275	4-heptanol	1.400	1.462	1.525
234	4-penten-1-ol	0.150	0.284	–0.080	276	2-methyl-3-hexanol	1.320	1.313	1.374
235	1-pentanol	0.590	0.597	0.200	277	5-methyl-2-hexanol	1.380	1.430	1.331
236	2-pentanol	0.280	0.380	0.140	278	2,2-dimethylpentanol	1.520	1.070	1.157
237	3-pentanol	0.210	0.356	0.196	279	2,4-dimethylpentanol	1.600	1.412	1.338
238	2-methyl-1-butanol	0.460	0.428	0.031	280	4,4-dimethylpentanol	1.550	1.359	1.379
239	2-methyl-2-butanol	–0.150	0.044	0.178	281	2,3,3-trimethyl-2-butanol	0.710	0.550	1.046
240	3-methyl-1-butanol	0.510	0.494	0.160	282	1-octanol	2.350	2.409	2.310
241	3-methyl-2-butanol	0.180	0.207	–0.025	283	2-octanol	2.090	2.096	2.169
242	cyclohexanol	0.420	0.504	0.284	284	2-ethyl-1-hexanol	2.110	1.870	1.788
243	1-hexen-3-ol	0.590	0.572	0.614	285	2-methyl-2-heptanol	1.720	1.669	1.985
244	2-hexen-4-ol	0.400	0.325	0.453	286	3-methyl-3-heptanol	1.600	1.573	1.867
245	2-methyl-4-penten-3-ol	0.500	0.467	0.498	287	2,2,3-trimethyl-3-pentanol	1.270	1.217	1.980

Table 1 (Continued)

			estd aqueous solubilities					estd aqueous solubilities	
no.	name	obsd value	for given subset	for overall data set	no.	name	obsd value	for given subset	for overall data set
				Oxygen-Containing Compound Subset					
288	1-nonanol	3.010	3.057	3.081	295	2,6-dimethyl-4-heptanol	2.160	2.386	2.600
289	2-nonanol	2.740	2.596	2.713	296	3,5-dimethyl-4-heptanol	2.510	2.460	2.485
290	3-nonanol	2.660	2.411	2.680	297	α-terpineol	1.890	2.058	2.363
291	4-nonanol	2.590	2.507	2.761	298	1-decanol	3.630	3.616	3.920
292	5-nonanol	2.490	2.282	2.489	299[c]	2-undecanol	2.940	2.941	4.435
293	7-methyloctanol	2.490	2.629	2.805	300	1-dodecanol	4.670	4.854	5.425
294	2,2-diethylpentanol	2.420	2.194	2.353					

[a] Outlier for development of hydrocarbon model. [b] Outlier for development of halogenated hydrocarbon model. [c] Outlier for development of alcohol and ether model. [d] Outlier for development of model for all compounds.

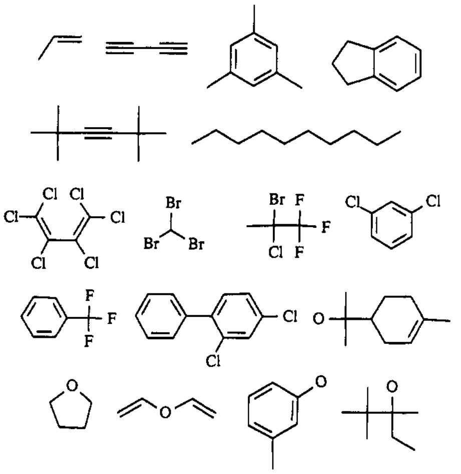

Figure 1. Selection of the compounds forming the aqueous solubility study set.

selection methods. After pairwise correlations were eliminated, vector space descriptor analysis was performed, and multiple linear correlations were eliminated, about 45 descriptors remained for the regression analysis in each of the data sets.

REGRESSION ANALYSIS AND OUTLIER DETECTION

Regression analysis was performed on the three individual subsets and on the whole data set. Initially, several models with varying numbers of descriptors were generated for the whole data set to determine what the optimal number of descriptors to use would be. With 300 compounds, a large number of descriptors could be evaluated, but after these exploratory studies, it was decided that nine descriptors would be used in the models that were generated for this data set. There are two principal reasons for this decision. The first is that using more descriptors does not substantially improve the correlation coefficient or the standard error of the model. The second reason is that when a principal component analysis was done on a model of fifteen variables, the first nine principal components represented 99.9% of the useful information contained in those descriptors.[23]

Models were then generated for the individual compound subsets. However, it soon became apparent that extensive outlier analysis would have to be done. This was especially true of the hydrocarbon subset, apparently because these data were taken from so many different sources. Outlier analysis was based on three different methods: a set of six data diagnostics,[14] robust regression analysis (RRA),[15] and duplexing. The data diagnostics include the use of residuals, leverage values, and other such standard diagnostics. Robust regression analysis uses the least median of squares criterion for detecting the presence of outliers. It is especially useful since it seeks outliers in a set of data all at once rather than individually as with the other methods. Duplexing is a method where the data set is divided randomly into two halves many times, and one half is a training set and the other half is a prediction set. Those compounds that cannot be predicted consistently are labeled as outliers. The results of the duplexing were consistent with the results from the data diagnostics and robust regression analysis.

The problems encountered with the aqueous solubility data for the hydrocarbon compounds were extensive. The data came from many different sources and were not internally consistent. In an effort to find a subset of the hydrocarbon set that was internally consistent, a model was generated from those compounds that were in near perfect agreement between the three different sources. There were 59 out of the 123 hydrocarbons that were in suitable agreement: 20–26, 28–30, 32, 34, 36, 38, 39, 41–48, 52, 53, 56–58, 60–63, 66, 68, 71, 73–75, 79, 81, 83, 84, 86, 88, 89, 91–94, 96, 98, 99, 101, 103, 105–107, 109, 110. These 59 compounds were used to generate a preliminary model, and the remaining 64 compounds were used as a prediction set. Those compounds from among the 64 prediction compounds whose solubuility was predicted within two standard deviations of the standard error of the model were then included in an expanded data set. By this method, 36 more compounds were added to the group: 1,4, 5, 8, 14–18, 27, 31, 40, 49, 65, 67, 69, 70, 72, 76–78, 80, 82, 85, 90, 95, 97, 100, 102, 108, 111, 115, 116, 118, 121, 123. The data set then totaled 95 compounds out of the original 123. A model was generated from these 95 compounds. However, outlier analysis from the data diagnostics, robust regression analysis, and duplexing showed that four of these compounds were outliers. These compounds were methane, ethane, 1,3,5–trimethylbenzene, and 2–methyldecalin: 1, 4, 100, 121. Methane and ethane are probably outliers due to some deficiency in the model, since their solubilities are known to a high degree of accuracy. The reason that the other two compounds are outliers is not clear. It could be inaccurate solubility data or it could be a problem with the model.

The final nine-descriptor model for the hydrocarbons was developed with 91 compounds. The 32 hydrocarbon outliers are labeled as such in Table 1. The final model for the hydrocarbons is shown in Table 2, and a scatter plot of the calculated versus observed solubilities is shown in Figure 2. The standard error is only 0.168 log units. The fitted values for the 91 hydrocarbons used to generate the model are listed in Table 1, and predicted values for the 32 outliers are also given. When all 123 hydrocarbons were used to generate a model using the same nine descriptors, a standard error of

Table 2. Descriptors, Coefficients, Standard Errors, and *T*-Values for the Model of the Hydrocarbon Data Set

descriptor	coefficient	std error	*T*-value
CONS	-2.927	0.299	-9.785
QNEG	-12.06	1.68	-7.166
DPOL	-3.823	0.515	-7.418
PPSA 1	0.01715	0.00042	40.68
WNSA 3	-0.2048	0.0329	-6.221
V5PC	0.2122	0.0460	4.613
MOMH 4	0.2801	0.0971	2.883
ALLP 1	2.447×10^{-3}	0.636×10^{-3}	3.848
NDB	-0.4315	0.0453	-9.531
NTB	-0.8138	0.0830	-9.802

$N = 91$
$R = 0.9891$
std error = 0.168
overall $F = 404.84$

Descriptions of Descriptor Names

QNEG	partial atomic charge on the most negatively charged atom
DPOL	dipole moment
PPSA 1	partial positive surface area
WNSA 3	weighted partial negative surface area
V5PC	valence corrected path-cluster-five molecular connectivity index
MOMH 4	first moment of inertial/second moment of inertia (including hydrogens)
ALLP 1	total number of paths in the molecule
NDB	number of Double Bonds
NTB	number of Triple Bonds

Table 3. Descriptors, Coefficients, Standard Errors, and *T*-Values for the Model of the Halogenated Hydrocarbon Data Set

descriptor	coefficient	std error	*T*-value
CONS	1.179	0.303	3.894
QNEG	4.343	0.743	5.845
VOL	5.895×10^{-3}	0.756×10^{-3}	7.800
FPSA 3	-10.66	3.89	-2.738
FNSA 3	10.89	1.48	7.359
RPCS	-0.02753	0.01036	-2.657
NBR	-0.08693	0.03130	-2.777
V6PC	0.3719	0.0793	4.688
S5C	0.4949	0.0561	8.825
MOMH 1	5.413×10^{-4}	1.267×10^{-4}	4.272

$N = 79$
$R = 0.9874$
std error - 0.180
overall $F = 297.62$

Descriptions of Descriptor Names

QNEG	partial atomic charge on the most negatively charged atom
VOL	molecular volume
FPSA 3	fractional partial positive surface area
FNSA 3	fractional partial negative surface area
RPCS	relative positive charged surface area
NBR	number of bromines
V6PC	valence corrected path-cluster-six molecular connectivity index
S5C	simple cluster five molecular connectivity index
MOMH 1	first moment of inertia (including hydrogens)

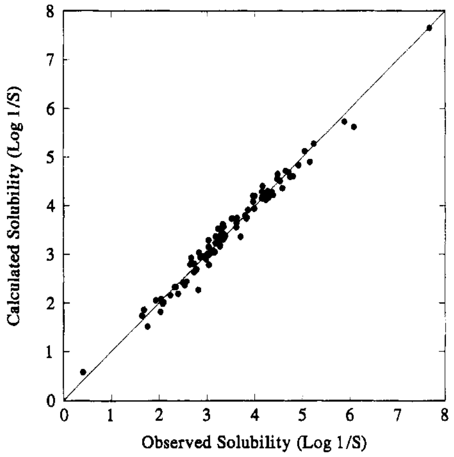

Figure 2. Scatter plot of the calculated and observed log (aqueous solubility) values for the hydrocarbon compounds.

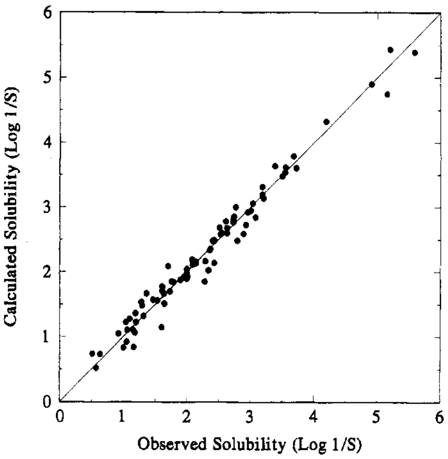

Figure 3. Scatter plot of the calculated and observed log(aqueous solubility) values for the halohydrocarbon compounds.

0.376 log units was obtained which indicates the size of the problem with the data for the 32 outlier hydrocarbons.

The problem of outliers was not so pronounced in the halogen-substituted hydrocarbon subset. There was only one compound out of the 80 that was classified as an outlier by the data diagnostics, robust regression analysis, and duplexing. This compound was tetrafluoroethene, 136, the only fully fluorinated compound in the data set. It is also possible that its aqueous solubility value is erroneous. The final nine-descriptor model for the halogenated hydrocarbons is shown in Table 3, and a scatter plot of the calculated versus observed solubilities is shown in Figure 3. The standard error is only 0.180 log units. The fitted values for the 79 halogenated hydrocarbons used to generate the model are listed in Table 1, and the predicted value for the one outlier is also given.

Five out of the 97 compounds in the alcohol and ether data set were labeled as outliers by the three outlier detection techniques. These compounds were diallyl ether, isopropyl ether, methanol, benzyl alcohol, and 2-undecanol: 220, 223, 226, 263, 299. It is difficult to propose a reason for why these compounds would be outliers, other than to say that the data used may be erroneous. The final nine-descriptor model for the alcohols and ethers is shown in Table 4, and a scatter plot of the calculated versus observed solubilities is shown in Figure 4. The standard error is only 0.167 log units. The fitted values for the 92 alcohols and ethers used to generate the model are listed in Table 1, and the predicted values for the five outliers are also given.

Next, models were generated for the entire data set based on the 262 compounds that were used in the final models for the three subsets. Then, outlier analysis was done on this set of compounds. Four compounds were determined to be outliers based on the results of data diagnostics and robust regression analysis. These compounds were benzene, *o*-chlorophenol, 3-chlorobiphenyl, and anisole: 88, 196, 203, 224. For benzene, there does not seem to be a good explanation for its being an outlier. However, the other three compounds are all unique compounds in the data set. *o*-Chlorophenol is the only compound in the data set that has more than one kind of functional group, anisole is the only aromatic ether, and 3-chlorobiphenyl is the only biphenyl that has one chlorine.

Table 4. Descriptors, Coefficients, Standard Errors, and *T*-Values for the Model of the Ethers and Alcohols Data Set

descriptor	coefficient	std error	*T*-value
CONS	4.965	0.293	16.96
DPOL	–0.7332	0.1618	–4.531
PNSA 1	9.374×10^{-3}	2.102×10^{-3}	4.459
RPCG	–0.6954	0.1860	–3.739
RNCG	–6.399	0.331	–19.34
MOMI 1	9.477×10^{-4}	0.594×10^{-4}	15.96
GEOH 6	–0.02152	0.00279	–7.720
N4C	–0.2181	0.0454	–4.800
N6PC	0.02910	0.00437	6.663
NDB	–0.7641	0.0666	–11.48

$N = 92$
$R = 0.9872$
std error = 0.167
overall F = 349.58

Descriptions of Descriptor Names

DPOL	dipole moment
PNSA 1	partial negative surface area
RPCG	relative positive charge
RNCG	relative negative charge
MOMI 1	first moment of inertia (without hydrogens)
GEOH 6	second geometric moment/third geometric moment (including hydrogens)
N4C	number of cluster-four paths
N6PC	number of path–cluster-six paths
NDB	number of double bonds

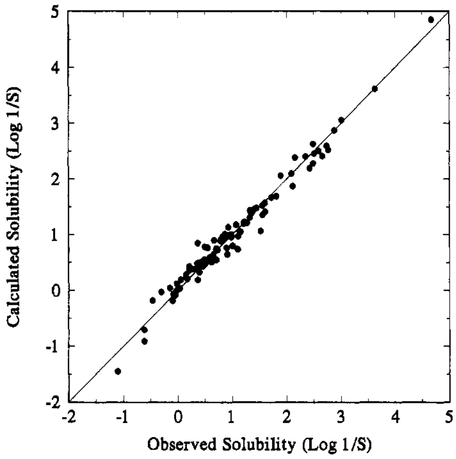

Figure 4. Scatter plot of the calculated and observed log(aqueous solubility) values for the alcohol and ester compounds.

Using the remaining 258 compounds, the final nine-variable model for this data set was developed. The model for the whole data set is shown in Table 5, and a scatter plot showing the calculated versus observed values is shown in Figure 5. The standard error for this combined model is 0.374, substantially larger than those of the three subsets. The fitted values for the 258 compounds used to generate the model are listed in the final column of Table 1, and the predicted values for the 42 outliers are also given.

The descriptors in the final models are not highly correlated with one another. Of the 144 pairwise correlations, r, between every pair of descriptors in the four models, only twelve have $|r| > 0.5$ and most values are in the zero to 0.3 range.

MODEL VALIDATION

Internal validation was used to validate the final models generated for this data set. Since jackknifing for this data set would be an extremely tedious process, the more general leave-n-out method was used to do the validation. For the each of the three subsets, a leave-4-out method was used, and for the entire data set a leave-10-out method was used. All of the residuals for each of the generated models fell within one standard deviation of the standard error cutoff. All of the coefficients of the regenerated models also fell within one standard deviation of the standard error of the coefficients. As a quantitative evaluation of the results of the internal validation method, the results of the comparison of the root mean square error of the residuals of the fit to the root mean square (RMS) error of the jackknifed residuals is presented in Table 6. It is be expected that the RMS error of the jackknifed residuals should be larger than the RMS error of the residuals of the fit, and one can see that for each of the four models presented here that the RMS error of the jackknifed residuals is only slightly larger than the RMS error of the residuals of the fit.

Table 5. Descriptors, Coefficients, Standard Errors, and *T*-Values for the Model of the Whole Data Set

descriptors	coefficient	std error	*T*-value
CONS	4.448	0.103	43.40
DPOL	–0.2782	0.0453	–6.143
DPSA 1	3.179×10^{-3}	0.292×10^{-3}	10.89
FPSA 3	–25.95	2.96	–8.777
RPCG	–0.7725	0.2013	–3.838
RNCG	–6.108	0.291	–20.99
RNCS	0.02948	0.00367	8.042
MOMH 1	9.641×10^{-4}	0.686×10^{-4}	14.06
GEOH 4	-6.375×10^{-3}	2.893×10^{-3}	–2.204
V5PC	0.3070	0.0573	5.356

$N = 258$
$R = 0.9678$
std error = 0.374
overall F = 407.83

Descriptions of Descriptor Names

DPOL	dipole moment
DPSA 1	difference of partial surface areas
FPSA 3	fractional partial positive surface area
RPCG	relative positive charge
RNCG	relative negative charge
RNCS	relative negative charged surface area
MOMH 1	first moment of inertia (including hydrogens)
GEOH 4	first geometric moment/second geometric moment (including hydrogens)
V5PC	valence corrected path–cluster-five molecular connectivity index

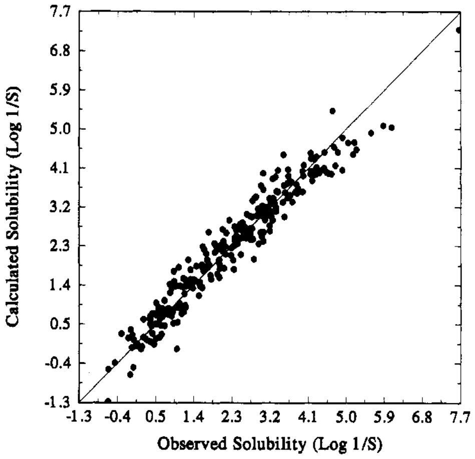

Figure 5. Scatter plot of the calculated and observed log(aqueous solubility) values for the entire set of compounds.

Table 6. Jackknifing Results for All Four Models

model	RMS residual	RMS jackknifed residuals
hydrocarbon	0.1582	0.1790
halogen	0.1681	0.1960
oxygen	0.1576	0.1892
all	0.3669	0.3921

Calculated versus observed plots and residual plots can also be used as evidence of the validity of a model. Residual plots were generated for each model, and they were seen to be completely random, which is what is desired. The calculated versus observed plots visually demonstrate the models' success at modeling solubility. From the preceding evidence, these four models can be considered successfully validated.

DISCUSSION

The standard errors of the models for the three subsets are approximately double the experimental errors, and the standard error of the model for the whole data set is greater than three times the experimental errors. This demonstrates that, even considering the errors that resulted from all the discrepancies in the data, there is a significant amount of work that can be done to improve prediction of aqueous solubilities. The main requirement for sound models is the quality of the data being used. These models do show that statistically sound models can be generated for aqueous solubility based on calculated structural descriptors alone.

The physical interpretation of the descriptors included in structure–property models is not always clear. This is especially true of descriptors that are mathematical constructs. However, aqueous solubility is a well understood phenomenon, and the following is an interpretation of the descriptors that appear in the model of the entire data set. The descriptors that appear in this model are as follows: the dipole moment, the difference in charged partial surface areas, the fractional partial positive surface area, the relative positive charge, the relative negative charge, the relative negative charged surface area, the radius of gyration, the first moment of inertia divided by the second moment of inertia, and the path-cluster-five valence-corrected molecular connectivity index.

There are many factors that influence a compound's aqueous solubility. Some of the most obvious are the polarity of the molecule, the size of the molecule, the shape of the molecule, steric effects, and the ability of the molecule to participate in hydrogen bonding. Since the polarity of a molecule is a very important factor in a compound's aqueous solubility, it is not surprising that dipole moment, DPOL, would be an important factor in determining aqueous solubility. In fact, dipole moment itself yields a one-variable equation with a correlation coefficient of 0.8 for this data set. In general, as dipole moment increases, the aqueous solubility increases. However, this is not always the case. For instance, butadiyne, which has a very small dipole moment, is not the least soluble. A compound such as 1,1-dibromopropane, which has a much larger dipole moment, is 50 times less soluble than butadiyne.

The difference in charged partial surface areas, DPSA 1, is a descriptor that calculates the difference between the total positive surface area and the total negative surface area. This descriptor is also a measure of the polarity of a molecule. But this descriptor also takes into account the size of nonpolar molecules. However, it is completely insensitive to the actual charges on the atoms, which means that large polar molecules will be more soluble than small polar molecules. This is generally not the case. Thus compounds such as dodecane and dodecanol, which have large positive values for this descriptor, are calculated to be much less soluble than 1,2,3-trichlorobenzene, which has a large negative value for this descriptor. Another descriptor is needed that encodes the size of the molecule.

The fractional partial positive surface area, FPSA 1, is the total positive surface area times the total positive charge divided by the total surface area. This descriptor is sensitive to both the size of the compound and the charge of the compound. The compounds that have small values for this descriptor are totally halogenated compounds which have very little positive surface area. On the other hand, compounds which have large values for this descriptor are small alcohols, which are very soluble in water. Thus the larger the value for this descriptor, the more soluble the compound.

The next three descriptors, the relatie positive charge, RPCG, the relative negative charge, RNCG, and the relative negative charged surface area, RNCS, all encode similar information in different ways. The relative positive charge is the charge on the most positive atom divided by the total positive charge; the relative negative charge is the charge on the most negative atom divided by the total negative charge; and the relative negative charged surface area is the charge on the most negative atom divided by the total negative charge times the surface area of the most negatve atom. These descriptors encode information about the size and charge of the molecule. These descriptors differentiate between molecules that are small and polar and tend to be more soluble, and those molecules which are large and nonpolar and tend to be less soluble.

The last three descriptors describe size and shape of the molecule. These three descriptors are the radius of gyration, MOMH 1, the first geometric moment divided by the second geometric moment, GEOH 4, and the valence-corrected path-cluster-five molecular connectivity index, V5PC. The radius of gyration has small values for compounds that are small and spherical or highly branched, and it has larger values for compounds that are large and linear. Thus, compounds that have large values for the radius of gyration are calculated to be less soluble. However, it should be noted that the coefficient of the radius of gyration is very small, so this descriptor has little influence on the calculation of the solubility of small molecules. The geometric moment descriptor is essentially a measure of length divided by width. Thus linear molecules have large values and cyclic and spherical molecules have small values. Generally, linear molecules are less soluble than cyclic and spherical molecules. For the connectivity index, any compound that has less than six non-hydrogen atoms or no branching has a value of zero for this descriptor, which means that this descriptor only influences the calculation for large, branched molecules. This descriptor will lead to lower solubilities for compounds that have larger values.

CONCLUSIONS

This study demonstrated the application of QSAR techniques to studying the relationship between chemical structure and aqueous solubility for a diverse group of organic compounds.

Regression analysis was used to build mathematical models which quantitatively describe the activity based on parameters derived from the structures of the compounds. These models can be used to predict the aqueous solubilities of structurally similar compounds for which the solubility is not known. In some cases, the descriptors can be interpreted in ways that help in understanding what structural features of a compound are important to the activity of interest.

These models also demonstrate that it is not necessary to use other experimentally derived quantities to predict aqueous solubility (such as the octanol–water partition coefficient or the melting temperature). That is, it is possible to use structural parameters alone to develop these quantitative structure–activity relationships.

ACKNOWLEDGMENT

This research was partially supported by a grant from the Exxon Corp.

REFERENCES AND NOTES

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

(2) Hansch, C.; Leo, A. J. *Substituent Constants for Correlation Analysis in Chemistry and Biology*; Wiley: New York, 1979.

(3) Stuper, A. J.; Brugger, W. E.; Jurs, P. C. *Computer Assisted Studies of Chemical Structure and Biological Function*; Wiley Interscience: New York, 1979; p 83.

(4) Kier, L. B.; Hall, L. H. *Molecular Connectivity In Chemistry and Drug Research*; Academic Press: New York, 1976.

(5) Pearlman, R. S. Molecular surface areas and volumes and their use in structure-activity relationships. In *Physical Chemical Properties of Drugs*; Yalkowsky, S. H., Sinkula, A. A., Valvani, S. C., Eds.; Dekker: New York, 1980; pp 321–345.

(6) Del Re, G. A Simple MO-LCAO Method for the Calculation of Charge Distributions in Saturated Organic Molecules. *J. Chem. Soc.* **1958**, 4031–4040.

(7) Dixon, S. L.; Jurs, P. C. Atomic Charge Calculations for Quantitative Structure-Property Relationships. *J. Comput. Chem.* **1992**, *13*, 492–504.

(8) Yates, K. *Hückel Molecular Orbital Theory*; Academic Press: New York, 1980.

(9) Lowe, J. P. *Quantum Chemistry*; Academic Press: New York, 1978.

(10) Stanton, D. T.; Jurs, P. C. Development and Use of Charged Partial Surface Area Structural Descriptors for Quantitative Structure–Property Relationship Studies. *Anal. Chem.* **1990**, *62*, 2323.

(11) Strang, G. *Linear Algebra and Its Applications*, 2nd ed.; Academic Press: New York, 1980.

(12) Draper, N. R.; Smith, H. *Applied Linear Regression Analysis*, 2nd ed.; Wiley Interscience: New York, 1981.

(13) Neter, J.; Wasserman, W.; Kutner, M. H. *Applied Linear Statistical Models*, 3rd ed.; Irwin: Boston, 1990.

(14) Belsey, D. A.; Kuh, E.; Welsch, R. E. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*; Wiley Interscience: New York, 1980.

(15) Rousseeuw, P. J.; Leroy, A. M. *Robust Regression and Outlier Detection*; Wiley: New York, 1987.

(16) Allen, P. M. Technical Report No. 23; Department of Statistics, University of Kentucky, Lexington, KY, 1971.

(17) Bodor, N.; Huang, M. J. A New Method for the Estimation of the Aqueous Solubility of Organic Compounds. *J. Pharm. Sci.* **1992**, *81*, 954.

(18) Suzuki, T. Development of an Automatic Estimation System for Both the Partition Coefficient and Aqueous Solubility. *J. Comput.-Aided Mol. Des.* **1991**, *5*, 149.

(19) Shaw, D. G. *Hydrocarbons with Water and Seawater*; Pergamom Press: Oxford, U.K., 1989; Vols. I and II.

(20) Kamlet, M. J.; et al. Linear Solvation Energy Relationships. 41. Important Differences between Aqueous Solubility Relationships for Aliphatic and Aromatic Solutes. *J. Phys. Chem.* **1987**, *91*, 1996.

(21) MOPAC, Version 5.0. QCPE Program No. 445; Quantum Chemistry Program Exchange, Indiana University: Bloomington, IN.

(22) Allinger, N. L.; Yul, Y. H. MM2/MMP2, 85-Force Field, QCPE Program No. 395; Quantum Chemistry Program Exchange, Indiana University: Bloomington, IN, 1985.

(23) Massant, D. L.; Vanderginste, BGM; Deming, S. N.; Michotte, Y.; Kaufman, L. *Chemometrics: A Textbook*; Elsevier: Amsterdam, 1988.