Estimation of Aqueous Solubility for a Diverse Set of Organic Compounds Based on Molecular Topology

Jarmo Huuskonen†

Division of Pharmaceutical Chemistry, Department of Pharmacy, POB 56, FIN-00014, University of Helsinki, Finland

Received October 21, 1999

An accurate and generally applicable method for estimating aqueous solubilities for a diverse set of 1297 organic compounds based on multilinear regression and artificial neural network modeling was developed. Molecular connectivity, shape, and atom-type electrotopological state (E-state) indices were used as structural parameters. The data set was divided into a training set of 884 compounds and a randomly chosen test set of 413 compounds. The structural parameters in a 30−12−1 artificial neural network included 24 atom-type E-state indices and six other topological indices, and for the test set, a predictive $r^2 = 0.92$ and $s = 0.60$ were achieved. With the same parameters the statistics in the multilinear regression were $r^2 = 0.88$ and $s = 0.71$, respectively.

INTRODUCTION

The aqueous solubility of drug compounds is one of the most important factors in determining its biological activity. In many cases drugs that show a good activity when administered parenterally maybe totally inactive when given orally. In such cases poor oral activity is often due to the fact that a sufficient amount of drug to desired response is not reached in the site of action. Hence an insufficient aqueous solubility is likely to hamper bioavailability of the drugs. In recent years high-throughput screening, where collections of thousands of compounds are screened with the intention of finding relevant biological activity, has proven valuable in finding new lead compounds.[1] It has been noticed that the synthesis of combinatorial libraries tends to result in compounds with higher molecular weights and higher lipophilicity, and presumably lower aqueous solubility, than with conventional synthetic strategies. For this reason computational screens have been suggested and used to select sublibraries with relevant physicochemical properties to the range of known values, such as lipophilicity and solubility, of the orally active drugs.[2−5] Hence there is a strong interest in fast, reliable, and generally applicable structure-based methods for prediction of aqueous solubility of new drugs before a promising drug candidate has even been synthesized.

Several approaches have been developed for the prediction of aqueous solubility based on nonexperimental structural parameters. These can be divided in substructure (group contribution) approaches[6−8] and in approaches where parameters are calculated directly from molecular structure,[9−18] such as topological indices, molecular volume, molecular surface area, etc. These methods employ multilinear regression or neural network modeling and varying ways of structural parametrization. However, currently used methods were developed from relatively small training sets ($n = 200-300$). One problem with small training sets is that they might not be representative but compiled from structural analogues. The use of a small and limited set of compounds in the training sets leads to models of closed systems, and their general applicability is questionable. This is clearly demonstrated by the fact that only three of above-mentioned methods[6,7,17] have been applied to the test set designed by Yalkowsky.[19] This test set contains 21 drug molecules and environmentally interesting compounds, like pesticides, with complex chemical structures.

In our earlier studies we have shown that aqueous solubilies,[17] log *S*, and partition coefficients,[20] log *P*, for drug compounds can be estimated with a reasonable accuracy on the basis of parameters derived from molecular topology. In this study we propose a method for estimating log *S* values with the same parameters but for a much larger and diverse set of organic compounds.

DATA SETS

The applicability and accuracy of a log *S* estimation method are strongly affected by the size and quality of the training set used. Experimental aqueous solubility values for the compounds used in this study were obtained from the AQUASOL dATAbASE of the University of Arizona[21] and SCR's PHYSPROP Database.[22] A set of 1297 organic compounds was extracted from these databases and was divided into a training set of 884 compounds and a randomly chosen test set of 413 compounds. The aqueous solubility values in 20−25 °C expressed as log *S*, where *S* is solubility in mol/L, were used. The log *S* values of the training set ranged from −11.62 to +1.58 with a mean of −2.70 and standard deviation of 2.01. For the testing set, the smallest log *S* value was −10.41 and the largest +1.13. The mean and standard deviation were −2.77 and 2.07, respectively.

METHODS

Three different types of topological indices introduced by Kier and Hall[23−26] were used as structural parameters and

† Tel: 358 9 19159170. FAX: 358 9 19159556. E-mail: jarmo.huuskonen@helsinki.fi.

10.1021/ci9901338 CCC: $19.00
Published on Web 02/19/2000

were calculated using the Molconn-Z (Hall Associated Consulting, Quincy, MA) software with structure input for each analyzed compound using the SMILES line notation code. Simple and valence molecular connectivity indices up to third-order path ($^{1-3}\chi$ and $^{1-3}\chi^{v}$), shape indices ($^{1-3}\kappa$, $^{1-3}\kappa_{\alpha}$), flexibility index (ϕ), the number of hydrogen-bonding donors (HBD) and acceptors (HBA), and 39 atom-type electrotopological state (E-state) indices were calculated. Cross-correlation analysis showed that pairwise correlations were $r^2 < 0.80$; hence, all these 55 parameters contain useful information and could be used in regression analysis.

The multilinear regression (MLR) analysis was performed with SPSS software (v.8.0, SPSS Inc., Chicago, IL) running on a Pentium PC. The quality criteria on the fit in MLR analysis were squared correlation coefficient, r^2, standard deviation, s, and Fischer significant value, F, when all parameters in the model were significant at the 95% confidential level.

The artificial neural network simulations were carried out using NeuDesk software (v 2.20, Neural Computational sciences, U.K.). A three-layered, fully connected neural network was trained by the standard back-propagation learning algorithm with a logistic $f(x) = 1/(1 + e^{-x})$ activation function for both hidden and output nodes. The same set of parameters as in the MLR equation was tested in artificial neural networks (ANNs) with one output neuron, log S.

Before the training was started, the input and output values were scaled between 0.1 and 0.9, and adjustable weights between neurons were given random values of between −0.5 and 0.5. The learning rate and momentum parameter were set at 0.1 and 0.9, respectively. The training end point was determined on the basis of the average training error (E), which is the mean-square error between the target and actual output. The optimal training end point was searched for overtraining the network. It has been accepted that the ratio, ρ, of the number of input parameters to the number of weights should be greater than 2.0, although cross-validation allows for the use of smaller values.[27,28] Hence networks with 8, 10, 12, and 14 neurons in the hidden layer were studied. The network architecture and the training end point giving the highest coefficient of determination, r^2_{pred}, and the lowest standard error s for the predictions of the test set were then used. To avoid chance effects, the predictions were repeated 10 times with different random starting weights in the network, and the averaged log S values were calculated.

RESULTS AND DISCUSSION

In this study the aqueous solubility values of a diverse set of 1297 organic compounds were compiled from two highly evaluated databases. The data set was divided into a training set of 884 compounds for developing the MLR and ANN models and a randomly chosen test set of 413 compounds (test set 1) for evaluating the predictive ability of the models. Another test set of 21 compounds (test set 2) was also used and allowed comparison of the predictions with earlier results.

Myrdal et al.[29] pointed out that the experimental solubility values can differ by ∼1.0 log unit, especially for compounds with a very low log S value. Hence, for the training sets that are compiled from relatively complex chemical structures, standard deviation, s, will be not lower than ∼0.5 log unit.

Table 1. Structural Parameters in the Multilinear Regression Model

(A) Topological Indices

symbol	explanation	contribution	*t*-score
$^{1}\chi$	path 1 simple connectivity index	−0.438	7.877
$^{1}\chi^{v}$	path 1 valence connectivity index	+0.117	2.489
ϕ	flexibility index	−0.052	2.629
HBA	the number of H-bond acceptors	−0.475	6.259
Ar	aromaticity indicator	−0.439	4.690
Alif[a]	indicator for aliphatic hydrocarbons	−1.960	14.008

(B) Atom-type Electrotopological State Indices[b]

symbol	group	frequency[c]	contributon	*t*-score
SsCH3	$-CH_3$	797	−0.174	8.208
SssCH2	$-CH_2-$	699	−0.205	6.954
SdsCH	=CH−	188	−0.076	2.542
SaaCH	aCHa	771	−0.080	4.749
SdssC	=C<	579	0.115	2.743
SaasC	asCa	743	−0.078	2.486
SaaaC	aaCa	195	−0.319	10.793
SsNH2	$-NH_2$	197	0.117	6.779
SssNH	−NH−	223	0.301	9.257
SdsN	=N−	37	0.125	3.701
SaaN	aNa	141	0.173	8.094
SsssN	>N−	189	0.795	16.273
SddsN	−N≪	44	0.656	4.664
SssssNp	>N≪+	2	4.691	5.657
SsOH	−OH	400	0.087	9.645
SdO	=O	645	0.048	5.492
SssO	−O−	307	0.160	10.289
SsF	−F	43	−0.020	4.801
SsSH	−SH	11	−0.315	4.925
SdS	=S	35	−0.180	4.475
SdssS	>S=	3	−0.916	1.997
SsCl	−Cl	269	−0.135	13.504
SsBr	−Br	46	−0.336	10.711
SsI	−I	19	−0.619	9.561

[a] Indicator variable for compounds that contain only aliphatic C and H. [b] According to Kier and Hall.[25] [c] The number of compounds for which atom group is present.

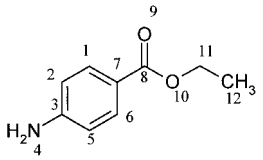
Stepwise and backward methods were employed in the regression analysis, and the following equation containing 30 parameters was calculated for the training set

$$\log S = \sum(a_i S_i) - 1.350 \qquad (1)$$

$$n = 884, \quad r^2 = 0.89, \quad s = 0.67, \quad F = 227.31, \quad r^2_{\text{cv}} = 0.88, \quad s_{\text{cv}} = 0.71$$

In this equation, n is the number of compounds used in the fit, F is the overall F-statistics for the addition of each successive term, r^2_{cv} is squared correlation coefficient of prediction in leave-one-out cross-validation, and a_i and S_i are the regression coefficients and the corresponding structural parameters. The regression coefficients in the equation are indicated in Table 1 with the t-scores of the significant parameters, and an example calculation of log S values by regression coefficients is given in Table 2. In the leave-one-out prediction of the MLR model, standard deviation of prediction, $s_{\text{cv}} = 0.71$, is only 0.04 unit higher than for the fitting model, $s = 0.67$. Such a small increase indicates a robustness of the model. Multilinear regression was also able to predict the log S values for 413 compounds in the test set with a coefficient of determination of $r^2_{\text{pred}} = 0.88$ and a standard deviation of prediction $s = 0.71$, which are in a good agreement with the results for the training set.

Table 2. E-State Indices Calculated for Benzocaine along with the Atom-type E-State Indices[a] and an Example of Calculating log *S* Value[b] by Regression Coefficients



atom ID	atom-type	symbol	E-state index
1	aCHa	aaCH	1.646
2	aCHa	aaCH	1.673
3	aCa	aasC	0.642
4	$-NH_2$	sNH2	5.449
5	aCHa	aaCH	1.673
6	aCHa	aaCH	1.646
7	aCa	aasC	0.533
8	=C <	dssC	−0.308
9	=O	dO	11.093
10	−O−	ssO	4.788
11	$-CH_2-$	ssCH2	0.392
12	$-CH_3$	sCH3	1.773

atom-type	atom-type E-state value
$SsCH_3$	1.773
SssCH2	0.392
SaaCH	6.638
SdssC	−0.308
SaasC	0.533
SsNH2	5.449
SdO	11.093
SssO	4.788

[a] According to Hall and Kier.[24] [b] $\log S = -0.438^1\chi + 0.117^1\chi^v - 0.052\phi - 0.475\text{HBA} - 0.438\text{Ar} - 1.96\text{Alif} - 0.174\text{SsCH}_3 - 0.205\text{SssCH2} - 0.08\text{SaaCH} + 0.115\text{SdssC} - 0.078\text{SaasC} + 0.117\text{SsNH}_2 + 0.048\text{SdO} + 0.160\text{SssO} - 1.35 = -1.85$ (estimated), −2.32 (experimental), −0.47 (error).

Table 3. Comparison of Predictive Ability of Multilinear Regression and Neural Network Models Using the Same Set of Parameters

	training set			test set 1			test set 2		
model	r^2	*s*	*n*	r^2_{pred}	*s*	*n*	r^2_{pred}	*s*	*n*
MLR[a]	0.89	0.67	884	0.88	0.71	413	0.83	0.88	21
ANN[a]	0.94	0.47	884	0.92	0.60	413	0.91	0.63	21
ANN[b]	0.90	0.46	160	0.86	0.53	51	0.68	1.25	21

[a] This study. [b] Our previous study.[17]

It was possible that there were some nonlinear dependencies between MLR optimized parameters and log *S* values. Hence, an application of nonlinear methods of data analysis could provide a better modeling of data. The back-propagation artificial neural networks were used to detect the presence of nonlinear dependencies in the analyzed data set as described in the next section.

The same set of the structural parameters as in the regression equation was used as inputs in neural network modeling. Several assays were made to find the optimal training end point and network architecture. The best performance of the network was achieved with 12 neurons in the hidden layer with the value of $\rho = 2.30$. The optimal training end point, $E = 0.032$, required ≈2300 training epochs when an ANN architecture of 30−12−1 was used. The neural network was able to estimate, with a reasonable degree of accuracy, most of the aqueous solubilities of the

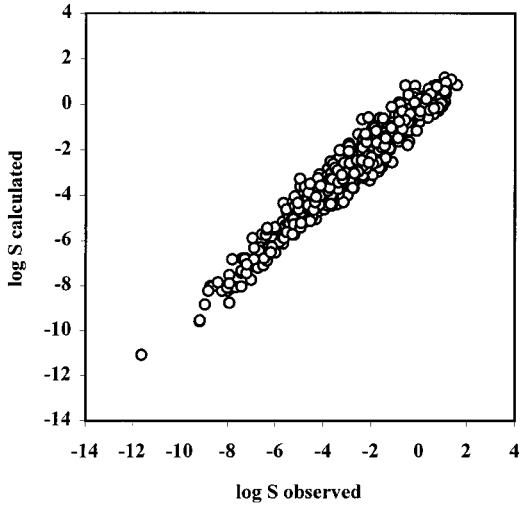


Figure 1. Correlation of calculated log *S* vs observed log *S* values for the training set by neural network.

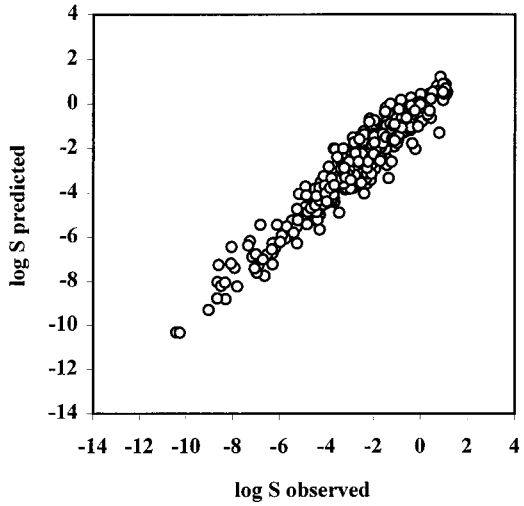


Figure 2. Correlation of predicted log *S* vs observed log *S* values for the test set 1 by neural network.

training set, $r^2 = 0.94$, $s = 0.47$, and $n = 884$ and the test set $r^2_{pred} = 0.92$, $s = 0.60$, and $n = 413$, respectively.

Statistics for the estimated aqueous solubilities of the organic compounds in the training set and test sets are presented in Table 3. The calculated and experimental aqueous solubilities of the training set and test sets are plotted in Figures 1−3. The list of all compounds and experimental and estimated log *S* values is available as Supporting Information.

The general applicability for the prediction ability of aqueous solubility was tested by the test set designed by Yalkowsky.[19] This test set is compiled of 21 commonly used compounds of pharmaceutical and environmental interest. The results of the predictions for this test set are presented in Table 4. The present multilinear regression and neural network models gave standard deviations $s = 0.88$ and 0.63. In our previous study[17] the results by neural network were $s = 1.25$ for all 21 compounds and $s = 0.55$ for a subset of 13 pharmaceuticals. Hence a significant improvement was achieved, and the predictions were better than those made by Klopman[6] and Kühne.[7] An interesting point of view is that Kühne used melting points in their group contribution approach and got a better fit for the training set of 694

Table 4. Observed and Predicted Aqueous Solubilities for the Test Set 2

no.	compound	log S_{obs}		ANN	MLR	Klopman[6]	Kuhne[7]
1	2,2',4,5,5'-PCB	−7.89		−7.21	−7.40	−7.90	−7.47
2	benzocaine[b]	−2.32		−1.79	−1.85	−1.71	na
3	aspirin	−1.72		−1.69	−1.74	−1.52	−1.93
4	theophylline	−1.39		−1.71	−0.78	−1.07	0.54
5	antipyrine[a]	0.39		−1.29	−1.20		−1.90
6	atrazine	−3.85		−3.51	−2.18	−3.05	−3.95
7	phenobarbital	−2.32		−2.97	−2.88	−2.08	−2.41
8	diuron	−3.80		−2.86	−3.20	−2.85	−3.38
9	nitrofurantoin	−3.38		−3.42	−3.03	−2.19	−2.62
10	phenytoin	−3.90		−3.40	−3.48	−3.47	−5.25
11	diazepam[a]	−3.76		−4.05	−4.26		−4.51
12	testosterone	−4.09		−3.98	−4.17	−5.17	−4.62
13	lindane	−4.64		−4.71	−5.34	−4.88	−5.08
14	parathion	−4.66		−4.13	−3.98	−3.94	−4.59
15	diazinon	−3.64		−4.01	−4.10	−5.29	−4.98
16	phenolphthalein	−2.90		−3.99	−4.05	−4.48	−4.61
17	malathion	−3.37		−3.24	−3.63	−2.94	−3.48
18	chlorpyriphos	−5.49		−5.61	−5.46	−5.77	−3.75
19	prostaglandin E2[b]	−2.47		−3.29	−4.35	−4.21	na
20	4,4′-DDT	−8.08		−7.67	−7.82	−8.00	−7.75
21	chlordane	−6.86		−7.29	−8.35	−7.55	−6.51
			r^2_{pred}	0.91	0.83	0.82	0.75
			s	0.63	0.88	0.86	1.06
			n	21	21	19	19

[a] Outliers in Klopman's model. [b] Predicted values not given.

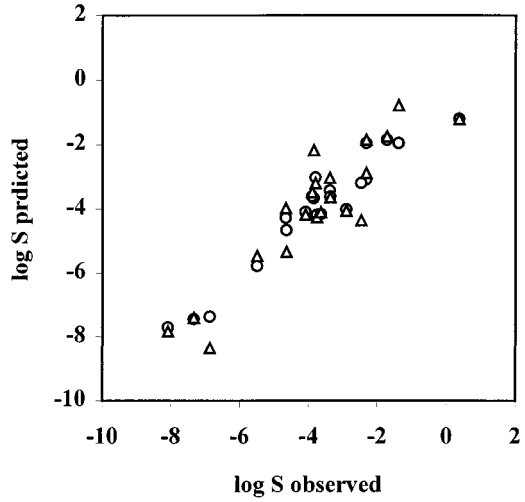


Figure 3. Correlation of predicted log S vs observed log S values for the test set 2 by multilinear regression (△) and neural network (○).

compounds than Klopman using only group contributors for a training set of 483 compounds. However, Klopman's model ($s = 0.86$ and $n = 19$) predicted better the solubilities in the test set of 21 compounds than Kühne's model ($s = 1.06$ and $n = 19$). Hence we could also ask if the correction term for solid compound, melting point, is really necessary for group contribution approaches.

An accurate and generally applicable method for estimating aqueous solubilities for a diverse set of 1297 organic compounds based on multilinear regression and artificial neural network modeling was developed. Topological indices cannot account for three-dimensional and conformational effects. Topological indices, however, are attractive because they can be calculated easily and rapidly and are error-free. The results of this study show that a practical solubility-predicting model can be constructed for a large and structurally diverse set of organic compounds with both multilinear regression and neural network modeling.

ACKNOWLEDGMENT

We thank William Howard from Syracuse Research Corporation for giving the PHYSOPROP database for our use and the Technology Development Center in Finland for financial support.

Supporting Information Available: Appendix I, giving the names of the compounds used in this study with their calculated and experimental aqueous solubility values (24 pages). This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Gillet, V. J.; Willet, P.; Bradshaw, J. Identification of Biological Active Profiles Using Substructural Analysis and Genetic Algorithm. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 165−179.

(2) Milne, G. W. A.; Wang, S.; Nicklaus, M. C. Molecular Modeling in the Discovery of Drug Leads. *J. Chem. Inf. Comput. Sci*. **1996**, *36*, 726−730.

(3) Ferguson, A. M.; Patterson, D. E.; Garr, C. D.; Underinger, T. L. Designing Chemical Libraries for Lead Discovery. *J. Biomol. Screen.* **1996**, *1*, 65−73.

(4) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3−25.

(5) Ghose, A. K.; Viswanadhan, V. N.; Wendoloski, J. J. A Knowledge-Based Approach in Designing Combinatorial or Medical Chemistry Libraries for Drug Discovery. 1. A Qualitative and Quantitative Characterization of Known Drug Databases. *J. Comb. Chem.* **1999**, *1*, 55−68.

(6) 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.

(7) Kühne, R.; Ebert, R.-U.; Kleint, F.; Scmidt, G.; Schüürmann, G. Group Contribution Methods to Estimate Water Solubility of Organic Chemicals. *Chemosphere* **1995**, *30*, 2061−2077.

(8) Lee, Y.-H.; Myrdal, P. B.; Yalkowsky, S. H. Aqueous Functional Group Activity Coefficients (AQUAFAC) 4: Application to Complex Organic Compounds. *Chemosphere* **1996**, *33*, 2129−2144.

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

(10) Bodor, N.; Huang, M.-J. Neural Network Studies. 1. Estimation of the Aqueous Solubilty of Organic Compounds. *J. Am. Chem. Soc.* **1991**, *113*, 9480−9483.
(11) Bodor, N.; Huang, M.-J. A New Method for the Estimation of the Aqueous Solubility of Organic Compounds. *J. Pharm. Sci.* **1992**, *81*, 954−960.
(12) Patil, G. S. Prediction of Aqueous Solubility and Octanol−Water Partition Coefficient for Pesticides Based on Their Molecular Structures. *J. Hazard. Mater*. **1994**, *36*, 35−43.
(13) Nelson, T. M.; Jurs, P. C. Prediction of Aqueous Solubility of Organic Compounds. *J. Chem. Inf. Comput. Sci*. **1994**, *34*, 601−609.
(14) Sutter, J. M.; Jurs, P. C. Prediction of Aqueous Solubility for a Diverse Set of Heteroatom-Containing Organic Compounds. *J. Chem. Inf. Comput. Sci*. **1996**, *36*, 100−107.
(15) Huuskonen, J.; Salo, M.; Taskinen, J. Neural Network Modeling for Estimation of the Aqueous Solubility of Structurally Related Drugs. *J. Pharm. Sci.* **1997**, *86*, 450−454.
(16) Huibers, P. D. T.; Katritzky, A. R. Correlation of the Aqueous Solubility of Hydrocarbons and Halogenated Hydrocarbons with Molecular Structure. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 283−292.
(17) Huuskonen, J.; Salo, M.; Taskinen, J. Aqueous Solubility Prediction of Drugs Based on Molecular Topology and Neural Network Modeling. *J. Chem. Inf. Comput. Sci*. **1998**, *38*, 450−456.
(18) Mitchell, B. E.; Jurs, P. C. Prediction of Aqueous Solubility of Organic Compounds from Molecular Structure. *J. Chem. Inf. Comput. Sci*. **1998**, *38*, 489−496.
(19) Yalkowsky, S. H.; Banerjee, S. In *Aqueous Solubility. Methods of Estimation for Organic Compounds*; Marcel Dekker: New York, 1992.
(20) Huuskonen, J. J.; Villa, A. E. P.; Tetko, I. V. Prediction of Partition Coefficients Based on Atom-Type Electrotopological State Indices. *J. Pharm. Sci.* **1999**, *88*, 229−233.
(21) Yalkowsky, S. H.; Dannelfelser, R. M. *The ARIZONA dATAbASE of Aqueous Solubility*; College of Pharmacy, University of Arizona: Tucson, AZ, 1990.
(22) Syracuse Research Corporation. *Physical/Chemical Property Database (PHYSOPROP)*; SRC Environmental Science Center: Syracuse, NY, 1994.
(23) Kier, L. B.; Hall, L. H. In *Molecular Connectivity in Structure−Activity Analysis;* Research Studies Press: Letchworth, 1986.
(24) Kier, L. B. Shape Indices of Orders One and Three from Molecular Graphs. *Quant. Struct.-Act. Relat*. **1986**, *5*, 1−7.
(25) Hall, L. H.; Kier, L. B. Electrotopological State Indices for Atoms Types: A Novel Combination of Electronic, Topological and Valence State Information. *J. Chem. Inf. Comput. Sci*. **1995**, *35*, 1039−1045.
(26) Hall, L. H.; Story, C. T. Boiling Point and Critical Temperature of a Heterogeneous Data Set: QSAR with Atom Type Electrotopological State Indices Using Artificial Neural Networks. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1004−1014.
(27) Manallack, D. T.; Livingstone, D. J. Artificial Neural Networks: Application and Chance Effects for QSAR Data Analysis. *Med. Chem. Rev*. **1992**, *2*, 181−190.
(28) Andrea, T. A.; Kalayeh, H. Application of Neural Notworks in Quantitative Structure Activity Relationships of Dihydrofolate Reductase Inhibitors. *J. Med. Chem.* **1991**, *34*, 2824−2836.
(29) Myrdal, P. B.; Manka, A. M.; Yalkowsky, S. H. AQUAFAC 3: Aqueous Functional Group Activity Coefficients: Application to the Estimation of Aqueous Solubility. *Chemosphere* **1995**, *30*, 1619−1637.

CI9901338