Estimation of Aqueous Solubility of Chemical Compounds Using E-State Indices

Igor V. Tetko,*,†,‡ Vsevolod Yu. Tanchuk,‡ Tamara N. Kasheva,‡ and Alessandro E. P. Villa†

Laboratoire de Neuro-Heuristique, Institut de Physiologie, Rue du Bugnon 7, Lausanne, CH-1005, Switzerland, and Biomedical Department, Institute of Bioorganic & Petroleum Chemistry, Murmanskaya 1, Kiev-660, 253660, Ukraine

Received July 3, 2000

The molecular weight and electrotopological E-state indices were used to estimate by Artificial Neural Networks aqueous solubility for a diverse set of 1291 organic compounds. The neural network with 33-4-1 neurons provided highly predictive results with $r^2 = 0.91$ and $RMS = 0.62$. The used parameters included several combinations of E-state indices with similar properties. The calculated results were similar to those published for these data by Huuskonen (2000). However, in the current study only E-state indices were used without need of additional indices (the molecular connectivity, shape, flexibility and indicator indices) also considered in the previous study. In addition, the present neural network contained three times less hidden neurons. Smaller neural networks and use of one homogeneous set of parameters provides a more robust model for prediction of aqueous solubility of chemical compounds. Limitations of the developed method for prediction of large compounds are discussed. The developed approach is available online at http://www.lnh.unil.ch/~itetko/logp.

INTRODUCTION

The solubility of liquids and solids in water is a very important molecular property that affects their biological activity. This parameter influences the uptake, distribution, transport, and eventually bioavailability of the drugs in the site of their action. Thus, there is a great interest in developing new methods for prediction of aqueous solubility.[1−12]

Recently, a new method for prediction of aqueous solubility of chemical compounds based on molecular topology and artificial neural network was proposed.[13] This method was applied to a set of 1297 organic compounds extracted from the AQUASOL database of the University of Arizona[14] and the PHYSPROP database.[15] This set was divided into a training set of 884 compounds and a randomly chosen validation set of 413 compounds. The method was tested using a set of 21 compounds previously analyzed by several authors.[3−5]

The initial input parameters used by the author included three different types of 55 topological indices introduced by Kier and Hall.[16,17] The author reported low redundancy of these indices that had pairwise correlations $r^2 < 0.80$. The indices included 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 E-state indices. These indices were analyzed by multiple linear regression (MLR) that calculated a final equation with 30 significant parameters. The selected parameters included 24 E-state and six other topological indices including indicator variables for aliphatic hydrocarbons and aromaticity. Artificial Neural Networks (ANNs) applied to analyze a set of 30 selected indices improved the calculated statistical parameters. These results indicated the presence of significant nonlinear dependencies amid input variables and target values that was revealed by ANNs but not by MLR, which is based on assumptions of linear dependencies amid parameters and activity.[18]

The atom-type E-state indices encode information about both the topological environment and the electronic interactions of an atom due to all other atoms in the molecule. Other topological indices used in the previous study[13] encoded similar information, and it is possible that some redundant information was presented in the input variables. In the present study we demonstrate that the results presented in ref 13 can be reproduced exclusively with E-state indices, i.e., using only one set of homogeneous parameters.

DATA SETS

The same data set of molecules from ref 13 was used in our study. The aqueous solubility was expressed as logS, where S is the solubility in mol/L.

After examination of the molecules, it was found that compounds saccharin and karbutilate were indicated twice under numbers **284**, **280** and **522**, **1126,**[19] respectively (numeration of molecules here and later in article is according to Appendix 1 of ref 13). The second repetition of the molecule was excluded.[35] Two compounds, minoxidil (**349**) and cyhexatin (**814**), were the only NO-oxide and organometal compound (Sn), while another two compounds, betaine (**26**) and cephaloridine (**216**), were the only two inner salts in the analyzed data set. All these compounds were in the training set, and thus no validation was performed for such classes of compounds. In addition, each of these molecules produced an E-state index (i.e., SssssSn for cyhexatin) that was repeated in the analyzed set only one time. Thus, these

* Corresponding author phone: ++41-21-692.5534; fax: ++41-21-692.5505; e-mail: itetko@eliot.unil.ch.
† Institut de Physiologie.
‡ Institute of Bioorganic & Petroleum Chemistry.

10.1021/ci000392t CCC: $20.00
Published on Web 09/19/2001

Table 1. Observed and Predicted Aqueous Solubility for the Test Set of 21 Compounds

no.	compound	logS exp	ANN4	Huuskonen[13]	Klopman[4]	Kühne[5]
1	2,2′,4,5,5′-PCB	−7.89	−7.57	−7.21	−7.90	−7.47
2	benzocaine	−2.32	−1.63	−1.79	−1.71	n/a[b]
3	acetylsalicylic acid	−1.61	−1.81	−1.69	−1.52	−1.93
4	theophylline	−1.37	−0.69	−1.71	−1.07	0.54
5	antipyrine	0.39	−0.89	−1.29	−2.76[a]	−1.90
6	atrazine	−3.55	−3.70	−3.51	−3.05	−3.95
7	phenobarbital	−2.34	−2.89	−2.97	−2.08	−2.41
8	diuron	−3.76	−3.01	−2.86	−2.85	−3.38
9	nitrofurantoin	−3.38	−3.09	−3.42	−2.19	−2.62
10	phenytoin	−3.99	−3.52	−3.40	−3.47	−5.25
11	diazepam	−3.76	−4.37	−4.05	−6.54[a]	−4.51
12	testosterone	−4.07	−4.13	−3.98	−5.17	−4.62
13	lindane	−4.60	−4.91	−4.71	−4.88	−5.08
14	parathion	−4.29	−4.31	−4.13	−3.94	−4.59
15	diazinon	−3.76	−3.43	−4.01	−5.29	−4.98
16	phenolphthalein	−2.90	−4.31	−3.99	−4.48	−4.61
17	malathion	−3.36	−3.73	−3.24	−2.94	−3.48
18	chlorpyriphos	−5.67	−5.31	−5.61	−5.77	−3.75
19	prostaglandin E2	−2.47	−3.52	−3.29	−4.21	n/a[b]
20	*p,p′*-DDT	−8.08	−7.59	−7.67	−8.00	−7.75
21	chlordane	−6.86	−7.23	−7.29	−7.55	−6.51
		r^2	0.91	0.92	0.70 (0.84)[c]	0.75
		s	0.64	0.63	1.24 (0.86)[c]	1.08
		n	21	21	21 (19)[c]	19

[a] Outliers in Klopman's model. [b] Predicted values are not indicated. [c] Values with and without (in parentheses) outliers.

four molecules were also excluded leaving 879 and 412 molecules in the training and in the validation sets, respectively.

For the test set of 21 compounds (Table 1) the experimental values provided by Yalkowsky[3] were used, except 2,2′,4,5,5′-PCB (−7.89 instead of −6.77) and chlordane (−6.86 instead of −5.35) as it was corrected by Klopman[4] and used by Huuskonen.[13] The list of molecules and their experimental and calculated logS values are available at http://www.lnh.unil.ch/~itetko/logp/aqueous.html.

METHODS

Molecular weights and 38 atom-type E-state indices (Table 2) were calculated using a program developed in-house (checked against the Molconn-Z software − Hall Associated Consulting, Quincy, MA) with structure input for each analyzed compound using the SMILES line notation code.[20]

The analysis of data was performed using multiple linear regression (MLR) analysis (SPSS v. 5.0) and neural networks with one hidden layer and bias neurons. The number of neurons in the hidden layer was optimized as indicated in the Results section. One single output node was used to code log *S* values. All calculations reported in this study were performed using an ensemble of 200 neural networks.

Two methods of neural network training were used, that were mainly different in the data partition scheme. One method was similar to that used in ref 13. All molecules from the first set of 879 molecules were used as the learning set, while the performance of ANNs was measured on the validation set of 412 molecules. The training was terminated when the neural network calculated minimum error for the validation set. Thus, this method used exactly the same training and validation sets for all neural networks in the ensemble. To get better statistical results 200 instead of 10 neural networks, as it was used in ref 13, were trained, and their results were averaged.

Another method used a subdivision of the total set of 1291 compounds into learning (645) and validation (646) subsets. Thus in this method only 645 instead of 879 molecules were used to train ANNs. As in the first method, neural network training was terminated when it calculated minimum error for the validation set. The partition of data onto training and validation sets was done by chance for each neural network in the ensemble. Thus, the main difference of this method is that for all 200 neural networks training and validated sets were different. This method is known in the literature as the Early Stopping over Ensemble (ESE).[21,22] One of the features of ESE method is that it calculates leave-one-out (LOO) estimates for the initial training set (i.e., in our case for 1291 molecules) that can be used to estimate a predictive ability of neural network method for the training set as described in ref 21. ESE can also be used to calculate the validation results for any subset of the training set, e.g. for the validation set used in ref 13. For example, let us take the compound naphthacene (**1292**) and consider first 10 neural networks from the ensemble. Since the input data were partitioned on the learning and validation sets by chance for each neural network, this compound was in the learning sets of neural networks 2, 5, 6, and 7 and in the validation sets of neural networks 1, 3, 4, 8, 9, and 10. Thus, the networks 1, 3, 4, 8, 9, and 10 predict (more correctly calculate validated value) activity of naphthacene, since this compound was never used to train these neural networks. On the contrary, neural networks 2, 5, 6, and 7 calculate fitted values for naphthacene. In the same way the validated/fitted results can be calculated for any molecule in the initial training set. Since validated and learning sets are selected by chance, each molecule is on average the same number of times in training and in validating sets.

The performance of ANNs was estimated according square of correlation coefficient r^2 and root-mean-squared (RMS) error. The calculations using ANNs were performed on the HP Workstation Cluster at the Swiss Center for Scientific

Table 2. Calculated Atom-Type E-State Indices

no.	index	group	freq[a]	used as	MLR1[b]	MLR2[c]	MLR/ANN[d]
1	MW				0.005	0.009 (0.002)	
2	SsCH3	$-CH_3$	797		−0.32	−0.36 (0.03)	−0.174
3	SssCH2	$-CH_2-$	695		−0.41	−0.46 (0.03)	−0.205
4	SsssCH	>CH−	363		−0.08	−0.12 (0.03)	
5	SssssC	>C<	214		−0.22	−0.23 (0.05)	
6	SdCH2	$=CH_2$	45		−0.20	−0.23 (0.04)	
7	SdsCH	=CH−	191		−0.27	−0.30 (0.03)	−0.076
8	SdssC	=C<	588		0.099*		0.115
9	SaaCH	aCHa	764		−0.24	−0.27 (0.02)	−0.080
10	SaasC	aaC−	749		−0.26	−0.30 (0.03)	−0.078
11	SaaaC	aaCa	115		−0.38	−0.42 (0.03)	−0.319
12	SddC	=C=	1	SdssC	0.10*		
13	StCH	≡CH	10	StsC	−0.25		
14	StsC	−C≡	26		−0.27	−0.28 (0.04)	
15	SsNH2	$-NH_2$	166		−0.05	−0.08 (0.02)	0.117
16	SssNH	−NH−	230		0.04*		0.301
17	SsssN	>N−	166		0.53	0.48 (0.04)	0.795
18	SdNH	=NH	3	SdsN	−0.25		
19	SdsN	=N−	36		−0.12	−0.15 (0.03)	0.125
20	SaaN	aNa	138		−0.05	−0.08 (0.02)	0.173
21	SaaNH	aaNH	20	SaaN	−0.08*		
22	SaasN	aaN−	15		0.45	0.42 (0.16)	
23	SddsN	≫N−	44		0.71	0.61 (0.15)	0.656
24	StN	≡N	15		−0.07*	−0.09 (0.03)	
25	SsOH	−OH	397		−0.04	−0.07 (0.01)	0.087
26	SssO	−O−	308		−0.03	−0.05 (0.03)	0.160
27	SdO	=O	643		−0.07	−0.09 (0.01)	0.048
28	SaaO	aOa	16		−0.11	−0.14 (0.05)	
29	SsSH	−SH	6	SssS	−0.30		−0.315
30	SssS	−S−	52		−0.36	−0.41 (0.07)	
31	SdS	=S	39		−0.45	−0.49 (0.05)	−0.180
32	SaaS	aSa	20		−0.51	−0.67 (0.16)	
33	SdssS	>S=	3	SaaS	−1.53		−0.916
34	SddssS	≫S<	60		−0.06	−0.08 (0.04)	
35	SdsssP	=P	25		−0.31	−0.28 (0.13)	
36	SsF	−F	43		−0.12	−0.14 (0.01)	−0.020
37	SsCl	−Cl	269		−0.24	−0.28 (0.02)	−0.135
38	SsBr	−Br	46		−0.50	−0.63 (0.06)	−0.336
39	SsI	−I	20		−0.92	−1.23 (0.14)	−0.619
40[e]	SssssNp	>N<+	2				4.691
	constant				1.25	1.31 (0.09)	−1.350

[a] The number of compounds for which index is present. [b] MLR1 corresponds to the regression analysis of 879 compounds using 39 indices. The nonsignificant indices are marked with star. [c] MLR2 corresponds to the regression using 33 indices from the combined set. The significant indices and the confidence intervals (in parentheses) at the level of significance $p < 0.05$ are provided. [d] MLR/ANN corresponds to results from ref 13. In addition to the E-state indices six additional topological indices were also used in ref 13. [e] SssssNp index is absent since the inner salts, betaine (**26**) and cephaloridine (**216**), were excluded from the training set.

Computing and at the computer server of the University of Lausanne. The freely available online Internet versions of neural network, solubility, and lipophilicity programs are described in ref 23.

RESULTS

There were 38 different atom type E-state indices and molecular weight calculated for the analyzed molecules. A regression analysis applied to 879 molecules from the training set indicated that five indices, marked by a star in Table 2, were not significant in the regression. The statistical coefficients calculated by MLR (MLR1) are shown in Table 3.

Several indices, SddC, StCH, SdNH, SsSH, and SdssS, had a frequency (i.e., the number of compounds for which index was calculated) in the data set of not more than 10 per index. Some molecules with rare indices were in the training set, while other in the validation set. It was difficult to expect that a good generalization, especially with ANN, could be expected for molecules containing these indices. However, MLR analysis indicated that many of such indices were significant (Table 2). Many of the rare indices had coefficients similar to indices related to them in properties but that had higher frequency in the data set. For example, SsSH and SssS had −0.30 and −0.36, while SdNH and SdsN had −0.25 and −0.12, respectively. The rare indices were eliminated, but their values were added to the values of the alias indices indicated in the "used as" column of Table 2. This procedure decreased the number of available indices to 34. Two indices, SaaN and SaaNH, corresponding to the aromatic nitrogen in a ring with and without hydrogen, also had quite similar regression coefficients in MLR. The regression coefficient of SaaNH was of the same sign and approximately two times larger than that of SaaN. Nevertheless, the term corresponding to the SaaNH index was not significant and this index was eliminated by MLR, probably due to its limited frequency, 20, in the analyzed data set. These indices were joined together thus decreasing the total number to 33 indices. This set will be further referred to in the article as the "combined" set.

Table 3. Statistical Parameters Calculated for Analyzed Data Sets[a]

(1) Results Using Fixed Size Training and Validation Sets

			training set (879)		validation set (412)		test set (21)	
model	params	*n*	r^2	*RMS*	r^2	*RMS*	r^2	*RMS*
			Results from Ref 13					
MLR	regressed	30	0.89	0.67	0.88	0.71	0.83	0.88
ANN	regressed	30	0.94	0.47	0.92	0.60	0.91	0.63
			Current Study Results					
MLR1	regressed1	34	0.86	0.75	0.85	0.81	0.76	1.00
MLR2	regressed2	31	0.86	0.75	0.85	0.81	0.77	0.99
ANN1	combined	33	0.93	0.53	0.90	0.66	0.89	0.67
ANN2	regressed1	34	0.93	0.55	0.89	0.68	0.87	0.73
ANN3	regressed2	31	0.92	0.56	0.89	0.68	0.88	0.71

(2) Results Using ESE Method

			training set (879				validation set (412)		test set (21)	
model	params	*n*	r^2	*RMS*	r^2_{LOO}	RMS_{LOO}	r^2	*RMS*	r^2	*RMS*
ANN4	combined	33	0.95	0.47	0.91	0.62	0.92	0.60	0.90	0.64
ANN5	regressed1	34	0.94	0.50	0.89	0.66	0.90	0.64	0.89	0.67
ANN6	regressed2	31	0.94	0.49	0.90	0.64	0.90	0.64	0.90	0.66

[a] MLR: multiple linear regression; ANN: artificial neural networks; r^2: square of correlation coefficient; RMS: root mean squared error; LOO: leave-one-out results calculated for the training set. ANN4 is available at http://www.lnh.unil.ch/~itetko/logp as ALOGPS 2.0 program.[23]

The regression analysis applied to the combined set of indices eliminated two nonsignificant indices, SdssC and SssNH, that were also nonsignificant in the regression analysis using all indices. The regression (Table 2) and statistical coefficients calculated by MLR (MLR2) were similar to those calculated when using all indices.

A preliminary study by neural networks was aimed to determine the optimal number of hidden neurons. The ESE method was applied using 33 indices. It was found that the LOO error calculated for the training set decreased (i.e., $RMS_{LOO} = 0.80 \pm 0.01, 0.66 \pm 0.02, 0.63 \pm 0.02, 0.62 \pm 0.02$) when the number of neurons in the hidden layer was changed from 1 to 4. However, further increase in the number of hidden neurons from 4 to 5, 7, and 10 did not change the prediction ability of the neural networks (i.e., $RMS_{LOO} = 0.62 \pm 0.01, 0.62 \pm 0.01, 0.63 \pm 0.01$). Thus, the number of neurons in the hidden layer for all analysis was chosen to be 4.

The neural networks trained by the ESE method were applied to the combined set of 33 indices and to the two sets of indices that were found to be significant by MLR. The best results were calculated when using the combined set of parameters, while a lower prediction ability of ANNs was calculated to both sets that were significant in MLR.

The redundancy of the combined set of parameters for ANNs regression was analyzed using pruning methods described elsewhere.[18,24] The application of these methods indicated that, in fact, all parameters were relevant, and no one can be deleted without decreasing the predictive performance of ANNs.

The ANNs were also trained using fixed training/test sets, as it was done in ref 13. The results calculated by this approach for the validation and the test sets were on average worse than those calculated using the ESE method for all analyzed data sets. This result looks unexpected since the training set for ESE method was about 1.3 times smaller than that used with the fixed training/test partition scheme. However, since the learning and validation sets in ESE were selected randomly, all molecules from the initial set contributed to the training of neural network ensemble thus improving the generalization ability of the ensemble prediction. A high prediction performance of ESE was recently explained in the general framework of bias/variance problem of neural networks.[25] It was shown that the "early stopping" and selection by chance of input/validation sets reduced bias, while the averaging decreased the variance of neural network prediction. A combination of both methods, as implemented by the ESE technique, provided an improved performance compared to other approaches. This analysis suggests that the use of fixed training/test sets to train ANNs decreases the variance of neural network results and provides a larger bias in neural network predictions compared to ESE.

The development of modern drug design approaches and especially the use of combinatorial libraries tend to generate compounds with larger number of atoms in molecules. Figure 1A, however, gives a rather pessimistic result concerning the applicability of the proposed method applied to compounds with a large number of atoms. Indeed, the prediction error of this method increases approximately linearly as

$$y = 0.16x + 0.39 \quad (1)$$

where x is a number of non-hydrogen atoms. Thus the error RMS = 0.4 calculated for compounds with less than 10 non-hydrogen atoms is approximately doubled for compounds with $n = 25 - 30$ atoms. Prediction error of MLR calculated for the test set of 412 molecules had a very similar dependency from the number of non-hydrogen atoms:

$$y = 0.17x + 0.57 \quad (2)$$

The test set of 21 compounds used in this and in the previous studies was composed of molecules with average number of 19 non-hydrogen atoms. Benzocaine and lindane had a minimum number of 12 and prostaglandin E2 had a maximum number of 25 such atoms. A prediction performance for this set estimated using eqs 1 and 2 was RMS =

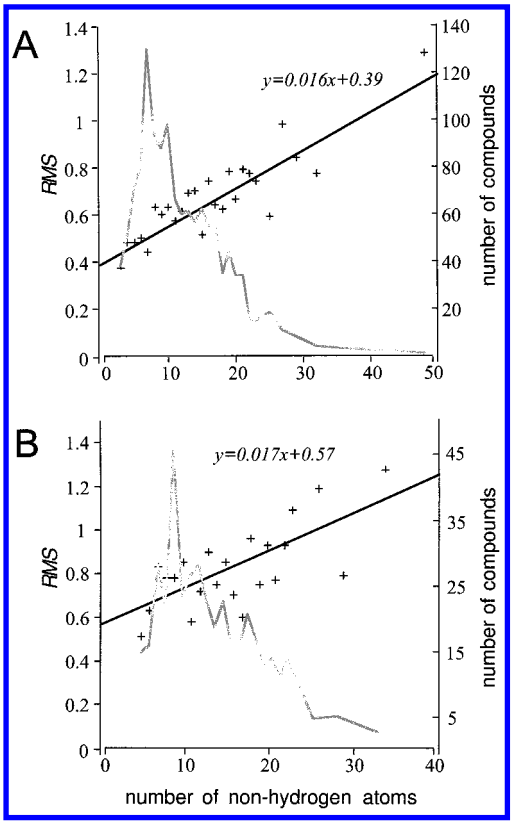

Figure 1. Root-mean-squared (RMS) error (crosses) and number of compounds in the analyzed data set (gray line) as a function of the number of non-hydrogen atoms in compound. (A) Neural networks leave-one-out results calculated with ANN4 model for the total set of 1291 molecules. (B) Multiple-linear regression results calculated for the test set of 412 molecules.

0.67 and RMS = 0.88, i.e., it was quite similar to the values RMS = 0.64 and RMS = 0.99 calculated by ANN and MLR, respectively.

DISCUSSION

The E-state indices, molecular weight, and ANN were used to calculate and to predict the activity of a large set of 1291 compounds. The prediction ability of the developed method measured on the validation set of 412 compounds and the test set of 21 compounds was similar to that reported for these sets by Huuskonen.[13] However, that study used the molecular connectivity, shape, flexibility, and indicator indices in addition to E-state indices. The new results demonstrate that use of only one set of homogeneous parameters can be sufficient to model the aqueous solubility of organic compounds.

Another important difference of the current study is that the neural networks had 3 times fewer hidden neurons than in ref 13. In the analysis using ESE with 1291 molecules each neural network contained only 141 neural network weights (i.e., adjustable parameters in neural network regression) that is more than nine molecules per weight. This ratio is about two times larger than the number of samples (5) per adjustable parameter considered as sufficient even for the linear models.[26] This indicates a high robustness of the calculated result. The prediction ability of the ANNs for the test set of 21 compounds was similar to that reported by Huuskonen,[13] and it was better than predictions made by Klopman[4] and by Kühne.[5]

The results calculated with ANNs were improved compared to the linear regression analysis. This indicated a presence of significant nonlinear dependencies amid input variables and aqueous solubility of analyzed compounds. The presence of nonlinear dependencies makes the parameters selected by MLR not optimal for ANNs. Nevertheless, MLR regression coefficients provide valuable information for analysis of indices with a small frequency. Such analysis made it possible to develop a combined set of 33 indices necessary to obtain the improved performance of ANNs. The developed set was optimal, and no parameters could be deleted from it using ANN pruning methods, without decreasing its prediction performance.

The prediction performance of our method almost linearly decreases with a number of non-hydrogen atoms in the analyzed molecule (Figure 1). The slope of the decrease of the error with a number of such atoms is practically the same for ANN and MLR. A similar behavior[27] was also observed for several methods, such as CLOGP,[28,29] KOWWIN,[30] XLOGP,[31] and our own method,[27] used to predict *n*-octanol/water partition coefficients. Thus, this dependency is rather general and should be taken into account when comparing/applying different approaches that were developed/tested using different sets of molecules. Equations 1 and 2 are too simplified to be used for a careful estimation of the prediction error for the test compounds, but still they can provide useful hints.

The decrease in the prediction power of the method with an increase in the number of non-hydrogen atoms is likely to be due to an increase of intermolecular interactions in complex structures, such as intramolecular hydrogen bonds[11] and folding of molecules due to London dispersion forces.[29] As a result of these interactions a part of the molecule become unavailable for solvent, thus provoking a decrease of prediction ability of the method. This problem could be probably addressed with molecular dynamics models and conformational analysis aimed to detect atoms that are not accessible for solvent and consequently to be excluded from the analysis. This leaves considerable room for improvement of the developed method.

The quality of experimental data is another issue for improvement of the prediction ability of the program. Most analyzed compounds, 1212, were present in the PHYSPROP[15] database including 1051 compounds with experimental values for 20−25 °C, i.e., in the same range that was reported for all compounds in ref 13. The experimental values for other compounds were either not reported in the PHYSPROP or they were for a different range of temperature. However, even for these 1051 compounds the experimental data in the PHYSPROP and in ref 13 were quite different, e.g. experimental logS values for benzylamine (**248**) and phenothrin (**1255**) were −1.53, −5.24 in the PHYSPROP and 0.97, −7.56 in ref 13, respectively. Overall, there were 45 compounds with absolute difference more than 0.5 log units and RMS = 0.24 log units (n = 1051) between two experimental sets was calculated. The compounds with different experimental values were selected, as indicated in ref 13, from the AQUASOL[14] database. This database contains over 15 000 solubility records for over 7000 compounds[32] including up to tens of measured experimental values for some chemicals. Some research[33] pointed out that

the experimental solubility values could differ by ~1.0 log unit, especially for compounds with low logS values. This can be due to many factors, including differences in experimental protocols, purity of solute, accuracy of measurements, control of temperature, etc. All these factors are very difficult to be taken into account when developing a water solubility prediction program. In general, there is a need of reliable data to be available for developing and testing new methods. To this extent, the efforts of, Cyprotex,[34] a company that has started a great work to collect and to measure physical properties of a large number of diverse compounds in the same conditions, using the same approach and the same apparatus is challenging. Over the next couple of years, this company will generate a database on a very large number of compounds. As this database builds up, it can provide the data resource that QSAR specialists could use to develop new reliable methods.

ACKNOWLEDGMENT

This study was partially supported by INTAS-Ukraine 95-0060 and INTAS-OPEN 97-0168 grants. We thank Jarmo Huuskonen for providing us data of the analyzed molecules.

REFERENCES AND NOTES

(1) Nirmalakhandan, N. N.; Speece, R. E. Prediction of Aqueous Solubility of Organic Chemicals Based on Molecular Structure. *Environ. Sci. Technol.* **1988**, *22*, 328−338.
(2) 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.
(3) Yalkowsky, S. H.; Banerjee, S. Aqueous Solubility, Methods of Estimation for Organic Compounds: Marcel Dekker: New York, Basel and Hong Kong, 1992; pp 128−148.
(4) Klopman, G.; Wang, S.; Balthasar, D. M. Estimation of Aqueous Solubility of Organic Molecules by the Group Contribution Approach. Application to the Study of Biodegradation. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 474−482.
(5) 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.
(6) Nelson, T. M.; Jurs, P. C. Prediction of Aqueous Solubility of Organic Compounds. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 601−609.
(7) 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.
(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) 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.
(10) 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.
(11) Abraham, M. H.; Le, J. The Correlation and Prediction of the Solubility of Compounds in Water Using an Amended Solvation Energy Relationship. *J. Pharm. Sci.* **1999**, *88*, 868−880.
(12) Ran, Y.; Yalkowsky, S. H. Prediction of Drug Solubility by the General Solubility Equation (GSE). *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 354−357.
(13) Huuskonen, J. Estimation of Aqueous Solubility for a Diverse Set of Organic Compounds Based on Molecular Topology. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 773−777.
(14) Yalkowsky, S. H.; Dannelfelser, R. M. The ARIZONA Database of Aqueous Solubility; College of Pharmacy, University of Arizona: Tucson, AZ, 1990.
(15) Syracuse Research Corporation. Physical/Chemical Property Database (PHYSPROP); SRC Environmental Science Center: Syracuse, NY, 1994.
(16) Kier, L. B.; Hall, L. H. *Molecular Structure Description: The Electrotopological State*; Academic Press: 1999.
(17) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure−Activity Analysis*; John Wiley & Sons: Chichester, 1986.
(18) Tetko, I. V.; Villa, A. E. P.; Livingstone, D. J. Neural network studies. 2. Variable selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 794−803.
(19) This compound had a name Kasugamycin in ref *13* but provided CAS RN corresponded to Karbutilate.
(20) Weininger, D. SMILES 1. Introduction and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31.
(21) Tetko, I. V.; Livingstone, D. J.; Luik, A. I. Neural Network Studies. 1. Comparison of Overfitting and Overtraining. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 826−833.
(22) Tetko, I. V.; Villa, A. E. P. Efficient Partition of Learning Data Sets for Neural Network Training. *Neural Networks* **1997**, *10*, 1361−1374.
(23) Tetko, I. V.; Tanchuk, V. Yu.; Kasheva. T. N.; Villa. A. E. P. Internet Software for Calculation of Lipophilicity and Aqueous Solubility of Chemical Compounds. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 246−252.
(24) Kovalishyn, V. V.; Tetko, I. V.; Luik, A. I.; Kholodovych, V. V.; Villa, A. E. P.; Livingstone, D. J. Neural Network Studies. 3. Variable Selection in the Cascade-Correlation Learning Architecture. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 651−659.
(25) Tetko, I. V.; Villa, A. E. P. An Enhancement of Generalization Ability in Cascade Correlation Algorithm by Avoidance of Overfitting/Overtraining Problem. *Neural Processing Lett.* **1997**, *6*, 43−50.
(26) Andrea, T. A.; Kalayeh, H. Applications of Neural Networks in Quantitative Structure−Activity Relationships of Dihydrofolate Reductase Inhibitors. *J. Med. Chem.* **1990**, *33*, 2583−2590.
(27) Tetko, I. V.; Tanchuk, V. Yu.; Villa. A. E. P. Prediction of *n*-Octanol/Water Partition Coefficients from PHYSPROP Database Using Neural Networks and E-state Indices. *J. Chem. Inf. Comput. Sci.* **2001**, 1407−1421.
(28) Leo, A. Calculating logPoct from Structures. *Chem. Rev.* **1993**, *93*, 1281−1306.
(29) Leo, A. I.; Hansch, C. Role of Hydrophobic Effects in Mechanistic QSAR. *Persp. Drug Design.* **1999**, *17*, 1−25.
(30) Meylan, W. M.; Howard, P. H. Atom/Fragment Contribution Method for Estimating Octanol−Water Partition Coefficients. *J. Pharm. Sci.* **1995**, *84*, 83−92.
(31) Wang, R.; Gao, Y.; Lai, L. Calculating Partition Coefficient by Atom-Additive Methodol. *Persp. Drug Design* **2000**, *19*, 47−66.
(32) http://www.Pharmacy.Arizona.EDU/peopleprograms/aquasol.
(33) Myrdal, P. B.; Manka, A. M.; Yalkowsky, S. H. AQUADAC 3: Aqueous Functional Group Activity Coefficients: Application to the Estimation of Aqueous Solubility. *Chemosphere* **1995**, *30*, 1619−1637.
(34) Cyprotex, Lloyd Street North, Manchester, M15 6SH, UK.
(35) Dr. Ai-xia Yan (University of Erlangen, Germany) found that lindane was indicated twice under numbers **740** in the training and **13** in the test sets when the article was already in press.

CI000392T