Eur. J. Med. Chem. 35 (2000) 1081–1088

PII: S0223-5234(00)01186-7/FLA

Original article

Prediction of aqueous solubility for a diverse set of organic compounds based on atom-type electrotopological state indices

Jarmo Huuskonen[a]*, Jukka Rantanen[b], David Livingstone[c]

[a]Division of Pharmaceutical Chemistry, Department of Pharmacy, P.O. Box 56, University of Helsinki, FIN-00014 Helsinki, Finland
[b]Pharmaceutical Technology Division, Department of Pharmacy, P.O. Box 56, University of Helsinki, FIN-00014 Helsinki, Finland
[c]ChemQuest, Delamere House, 1 Royal Crescent, Sandown, Isle of Wight, PO3688LZ and Centre for Molecular Design, University of Portsmouth, Portsmouth, Hants PO12EG, UK

Received 10 January 2000; revised 17 July 2000; accepted 2 August 2000

Abstract – We describe robust methods for estimating the aqueous solubility of a set of 734 organic compounds from different structural classes based on multiple linear regression (MLR) and artificial neural networks (ANN) model. The structures were represented by atom-type electrotopological state (E-state) indices. The squared correlation coefficient and standard deviation for the MLR with 34 structural parameters were $r^2 = 0.94$ and $s = 0.58$ for the training set of 675 compounds. For the test set of 21 compounds, the equivalent statistics were $r^2_{\text{pred}} = 0.80$ and $s = 0.87$, respectively. Neural networks gave a significant improvement using the same set of parameters, and the standard deviations were $s = 0.52$ for the training set and $s = 0.75$ for the test set when an artificial neural network with five neurons in the hidden layer was used. The results clearly show that accurate models can be rapidly calculated for the estimation of aqueous solubility for a large and diverse set of organic compounds using easily calculated structural parameters. © 2000 Éditions scientifiques et médicales Elsevier SAS

aqueous solubility / electrotopological state indices / artificial neural networks

1. Introduction

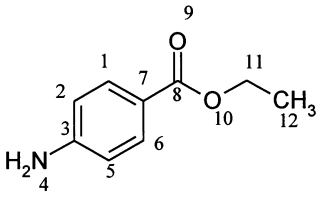
The aqueous solubility of drug compounds is one of the most important factors in determining its biological activity. In many cases, drugs that show a good activity when administered by the parenteral route may be totally inactive when given orally. In such cases, poor oral activity is often due to the fact that a sufficient amount of drug to achieve the desired response is not reached at the site of action. Hence an insufficient aqueous solubility is likely to hamper bioavailability of the drugs. In recent years, high throughput screening (HTS), 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 noted 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 sub-libraries with relevant physicochemical properties to the range of known values, such as lipophilicity and solubility, of the orally active drugs [2–5]. Although experimental HTS 'ranking' screens have been developed and used to evaluate solubility on a 96-well format, these methods require a sample of compound [6]. Hence there is much interest in fast, reliable and generally applicable structure-based methods for the prediction of aqueous solubility of new drugs before a promising drug candidate has even been synthesised.

* Correspondence and reprints.
E-mail address: jarmo.huuskonen@helsinki.fi (J. Huuskonen).

Several methods have been developed for the prediction of aqueous solubility, based on non-experimental structural parameters. These can be divided into two groups, substructure (group contribution) approaches [7–9], and approaches where parameters are calculated directly from molecular structure [10–19], such as topological indices, molecular volume, molecular surface area etc. These methods employ multiple linear regression or neural network modelling, and varying ways of structural parameterisation. 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 analogs. The use of small and limited sets of compounds in the training sets, leading to models of closed systems, and their general applicability are questionable. This is clearly demonstrated by the fact that only three of the above-mentioned methods [7, 8, 18] have been applied to the test set, designed by Yalkowsky and Banerjee [20]. 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 solubilities [18], log S, and partition coefficients [21], log P, for drug compounds can be estimated with reasonable accuracy based on parameters derived from molecular topology using neural network modelling. In this study, we describe a method for estimating log S values with the same parameters but for a much larger and diverse set of organic compounds.

Table I. The E-state indices calculated for benzocaine along with the atom-type E-state indices[a]



.

Atom ID	Atom-type	Symbol	E-state index
1	aCHa	aaCH	1.646
2	aCHa	aaCH	1.673
3	aCa	AasC	0.642
4	$-NH_2$	sNH_2	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	$-CH_2-$	$ssCH_2$	0.392
12	$-CH_3$	sCH_3	1.773
		Atom-type E-state value	
	$SsCH_3$	1.773	
	$SssCH_2$	0.392	
	SaaCH	6.638	
	SdssC	−0.308	
	SaasC	1.175	
	$SsNH_2$	5.449	
	SdO	11.093	
	SssO	4.788	

[a] According to Hall and Kier [24].

2. Datasets

The applicability and accuracy of a log S estimation method is 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 database of Kühne et al. and allowed a comparison with other group contribution methods as well [8]. This database contains aqueous solubility values at 25°C expressed as log S, were S is solubility in moles per litre for a diverse set of 694 organic compounds. Although many different structural classes are presented, there are only a few commonly used pharmaceuticals. Hence, a random selection of 38 drug compounds with accurately measured log S values was made from the literature and used as a validation set for MLR and ANN models. In addition, the test set of 21 compounds designed by Yalkowsky and Dannelfelser was used to compare the results of this study with the earlier proposed methods [22]. There were 19 compounds from this test set in the training set used by Kühne et al. [8], and, when these compounds were excluded, the training set of 675 was used to build the MLR and ANN models.

3. Methods

The atom-type electrotopological state indices introduced by Kier and Hall [23, 24] and Hall and Story [25] were used as structural parameters in a manner similar to group additive schemes. Each atom in the

Table II. The atom-type E-state indices and indicator variables used in multilinear regression and neural network models.

(A) Atom-type E-state indices

No	Symbol	Atom-type	Frequency	Contribution	*t*-score
1	$SsCH_3$	$-CH_3$	376	−0.275	18.289
2	$SdCH_2$	$=CH_2$	21	−0.157	4.062
3	$SssCH_2$	$-CH_2-$	304	−0.366	24.188
4	StCH	≡CH	5	−1.228	2.303
5	SdsCH	=CH–	62	−0.186	4.962
6	SaaCH	aCHa	378	−0.237	29.373
7	SsssCH	–CH<	136	−0.152	2.867
8	StsC	≡C–	14	2.266	2.214
9	SdssC	=C<	177	0.166	2.440
10	SaasC	asCa	369	−0.216	7.135
11	SaaaC	aaCa	120	−0.261	11.322
12	SssssC	>C<	70	−0.358	12.170
13	$SsNH_3^+$	$-NH_3+$	12	0.335	2.157
14	$SsNH_2$	$-NH_2$	57	0.030	2.103
15	SdsNH	=NH–	3	−0.183	3.336
16	SssNH	–NH–	46	0.058	1.684
17	StN	≡N	9	−0.625	2.762
18	SaaN	aNa	25	−0.172	7.387
19	SsssN	>N–	29	0.399	5.103
20	SddsN	–N≪	43	0.499	3.253
21	SsOH	–OH	165	0.022	3.853
22	SdO	=O	230	−0.063	13.567
23	SaaO	aOa	8	−0.260	10.244
24	SsF	–F	10	−0.107	16.799
25	SdsssP	–>P=	22	−0.282	4.108
26	SsSH	–SH	5	−0.355	5.048
27	SdS	=S	22	−0.410	10.938
28	SssS	–S–	15	−0.244	3.223
29	SsCl	–Cl	195	−0.209	60.308
30	SsBr	–Br	35	−0.413	22.655
31	SsI	–I	9	−0.945	13.142

(B) Indicator variables

32	Aliphatic hydrocarbons	32	−1.368	7.86
33	Aromatic hydrocarbons	66	−0.530	4.59
34	Pyridines	10	2.518	11.46

molecular graph is presented by an E-state value that encodes the intrinsic electronic state of the atom perturbed by the electronic influence of all atoms in the molecule within the context of topological character of the molecule. Thus, the E-state for a given atom (or atom type) varies from molecule to molecule and depends on detailed structure of the molecule. The atom-type E-state indices were calculated using the Molconn-Z (Hall Associated Consulting, Quincy, MA) software and structure input for each analysed compound was the SMILES line notation code. An example for the calculation of the atom-type E-state indices for benzocaine molecule is given in *table I*. Totally, 37 atom-type E-state indices were calculated. Cross-correlation analysis showed that all pair-wise correlations were $r^2 < 0.50$, hence all these 37 parameters could be used in multiple linear regression analysis. The chemical diversity of the training set can be seen from *table II* which lists the atom-type E-state indices and the frequency of use to derive MLR and ANN models. The logarithm values for partition coefficients, log *P*, were calculated using the KOWWIN program (Version 1.62; Syracuse Research Corporation Inc.).

Multiple linear regression (MLR) analysis was performed with the SPSS software (Version 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.D.), s, and Fischer significance value, F, when all the parameters in the model were significant at the 95% confidence level. Six atom-type E-state indices were found to be non-significant in the regression modelling ($P > 0.05$) and were thus omitted from the final equations.

The artificial neural network simulations were carried out using the NeuDesk software (Version 2.20; Neural Computational Sciences, UK). 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, both for hidden and output nodes. The same set of

Table III. Predicted and experimental aqueous solubility values for the 38 pharmaceuticals in the validation set.

No	Compound	log S_{exp}	MLR1	MLR2	MLR3	ANN
1	Aminopyrine	−0.36	−1.08	−1.40	−1.58	−1.14
2	Ephedrine	−0.47	−0.90	−2.02	−1.90	−1.57
3	Caffeine	−0.88	−1.82	−0.99	−1.55	−1.13
4	Paracetamol	−0.99	−1.31	−1.29	−1.43	−1.28
5	Metronidazole	−1.22	−0.94	−1.21	−1.39	−1.37
6	DL-conine	−1.50	−2.11	−2.03	−1.77	−2.56
7	Procaine	−1.78	−2.06	−2.37	−2.19	−2.35
8	Methylparaben	−1.83	−2.55	−1.30	−1.46	−1.19
9	Cyclobarbital	−2.02	−3.39	−2.52	−2.23	−2.97
10	Atropine	−2.20	−2.47	−3.11	−3.02	−3.11
11	Cocaine	−2.25	−2.58	−3.33	−3.38	−3.27
12	Phenacetin	−2.35	−2.41	−2.26	−2.55	−2.35
13	Barbital	−2.40	−1.83	−1.51	−1.43	−1.84
14	Tripelennamine	−2.64	−2.25	−2.02	−1.70	−1.88
15	Pipedemic acid	−2.98	−0.08	−3.11	−2.81	−3.47
16	Chlorfenac	−3.08	−4.35	−3.97	−4.06	−3.97
17	Prednisolone	−3.18	−3.09	−2.58	−2.19	−2.50
18	Nalidixic acid	−3.37	−3.24	−3.13	−3.37	−3.27
19	Doxepin	−3.40	−3.52	−4.99	−4.83	−4.17
20	Lorazepam	−3.60	−3.44	−5.10	−5.06	−4.15
21	Dibucaine	−3.70	−4.15	−3.36	−2.91	−3.07
22	Ibuprofen	−3.76	−4.41	−3.90	−4.12	−3.70
23	Nitrazepam	−3.80	−4.01	−4.25	−4.53	−4.06
24	Mebendazole	−3.88	−4.85	−4.48	−4.79	−4.41
25	Oxazepam	−3.95	−3.69	−4.32	−4.57	−3.42
26	Prasterone	−4.06	−3.83	−4.36	−4.50	−4.21
27	Triazolam	−4.08	−5.52	−4.66	−5.55	−3.90
28	Hydrastine	−4.11	−2.60	−3.66	−4.10	−3.46
29	Imipramine	−4.19	−5.65	−5.27	−5.66	−4.82
30	Naproxen	−4.20	−4.01	−3.94	−4.15	−3.69
31	Warfarin	−4.26	−3.30	−4.85	−4.89	−4.12
32	Indomethacine	−4.62	−5.17	−5.64	−5.32	−5.18
33	Norethindrone	−4.63	−4.35	−3.94	−3.39	−3.94
34	Sulindac	−5.00	−5.47	−6.00	−6.15	−4.71
35	Estradiol	−5.03	−5.47	−4.55	−4.88	−4.59
36	Chlorpromazine	−5.10	−6.58	−5.50	−6.15	−5.28
37	Thioridazine	−5.82	−6.66	−6.71	−6.62	−6.47
38	Fluotrimazole	−8.40	−6.50	−8.30	−8.34	−7.34
		r^2_{pred}	0.70	0.84	0.82	0.85
		s	0.92	0.67	0.71	0.62

Table IV. Comparison of the multiple linear regression and artificial neural network models to estimate aqueous solubility.

Model	Training set			Validation set			Test set		
	r^2	s	n	r^2_{pred}	s	n	r^2_{pred}	s	n
MLR1[a]	0.91	0.73	675	0.70	0.92	38	0.73	1.00	21
MLR2[b]	0.94	0.58	675	0.84	0.67	38	0.80	0.87	21
MLR3[c]	0.95	0.54	675	0.82	0.71	38	0.79	0.90	21
ANN[d]	0.96	0.51	675	0.85	0.62	38	0.84	0.75	21

[a] Values estimated with calculated log P values and melting points (Eq. (1)).
[b] Values estimated with 34 structural parameters without melting points (Eq. (2)).
[c] Values estimated with 34 structural parameters and melting points (Eq. (3)).
[d] Values estimated with 34 structural parameters from Eq. (2) in a 34-5-1 artificial neural networks.

structural parameters as in the MLR equation was tested in ANNs with one output neurone, log S.

Before the training was started, the input and output values were scaled between 0.1 and 0.9, and the adjustable weights between neurones were given random values between -0.5 and 0.5. The learning rate and the momentum parameters were set at 0.1 and 0.9, respectively. The optimal training endpoint and network architecture was determined on the basis of the validation set of 38 drug compounds. The network architecture and the training endpoint giving the highest coefficient of determination, r^2_{pred}, and the lowest standard error (S.E.), s, for the predictions of the validation set was then used. The predictions were repeated ten times with different random starting weights in the network and the averaged log S values were calculated.

4. Results and discussion

In this study, the aqueous solubility values of a diverse set of 734 organic compounds were compiled from two highly evaluated databases. The data set was divided into a training set of 675 compounds for developing the MLR and ANN models, and a randomly chosen validation set of 38 drug compounds for evaluating the predictive ability of the models. A test set of 21 compounds was also used and allowed comparison of the predictions with earlier results.

Stepwise and backward methods were employed in the regression analysis and the following equations were calculated for the training set

$$\log S = -1.01 \log P - 0.01\ \mathrm{mp} + 0.50$$

$$n = 675,\ r^2 = 0.91,\ s = 0.73,\ F = 3298.0,\ r^2_{cv} = 0.91,\ s_{cv} = 0.74 \quad (1)$$

$$\log S = \sum(a_i S_i) + 1.52$$

$$n = 675,\ r^2 = 0.94,\ s = 0.58,\ F = 303.9,\ r^2_{cv} = 0.93,\ s_{cv} = 0.63 \quad (2)$$

$$\log S = 0.005\ \mathrm{mp} + \sum(a_i S_i) + 1.39$$

$$n = 675,\ r^2 = 0.95,\ s = 0.54,\ F = 383.8,\ r^2_{cv} = 0.94,\ s_{cv} = 0.58 \quad (3)$$

In these equations, log P is the calculated octanol/water partition coefficient, m.p. is the melting point, n is the number of compounds used in the fit, F is the F-statistic, r^2_{cv} is the 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 Eq. (2) are shown in *table II* with the t-scores of the significant parameters. In the leave-one-out prediction of this MLR model without the melting points, the S.D. of prediction, $s_{cv} = 0.63$, is only 0.05 U higher than for the fitting model, $s = 0.58$. Such a small increase indicates a robustness of the model. The multiple linear regression was also able to predict the log S values for 38 drug compounds in the validation set with a coefficient of determination of $r^2_{pred} = 0.84$ and a S.D. of prediction, $s = 0.67$, which are in a good agreement with the results for the training set. As can be noted from the equations above, melting point correction for solid compounds offered only a little improvement in the statistics for the training set and no improvement at all for the validation set of 38 drug compounds (*table III*).

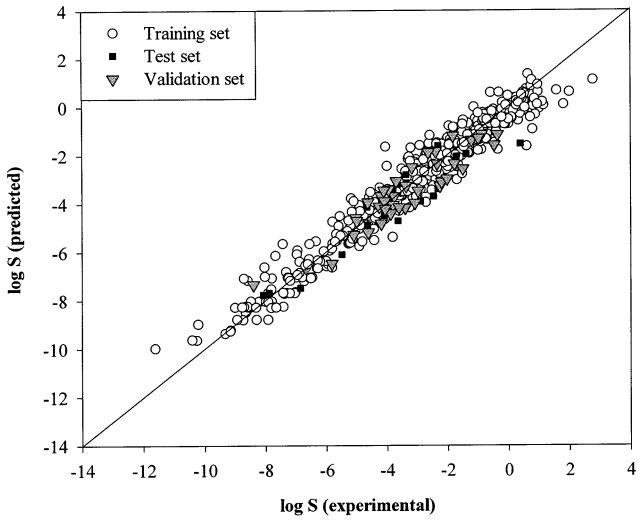


Figure 1. The predicted vs. experimental aqueous solubilities for training set ($n = 675$), validation set ($n = 38$) and test set ($n = 21$) by artificial neural networks.

It was possible that there were some non-linear dependencies between the MLR optimised parameters and log S values. Hence, the application of non-linear methods of data analysis might provide a better modelling of the data. Back-propagation artificial neural networks were used to detect the presence of non-linear dependencies in the analysed data set, as described in the next section.

The same set of 34 structural parameters as in regression Eq. (2) was used as the inputs in neural network modelling. Several assays were made to find the optimal training endpoint and network architecture. The best performance of the network was achieved with five neurones in the hidden layer. The optimal training endpoint required ca. 800 training epochs when the ANN architecture of 34-5-1 was used. Neural networks were able to estimate, with a reasonable degree of accuracy, most of the aqueous solubility values of the training set, $r^2 = 0.96(\pm 0.01)$, $s = 0.51(\pm 0.01)$ and $n = 675$ and the validation set $r^2_{\text{pred}} = 0.85(\pm 0.01)$, $s = 0.62(\pm 0.02)$ and $n = 38$, respectively.

Statistics for the estimated aqueous solubility values of the organic compounds in the training set, validation set and test set are presented in *table IV*, and the predicted and experimental aqueous solubility values in the validation set of 38 pharmaceuticals are

Table V. Predicted and experimental aqueous solubility values for the 21 compounds in the test set.

Number	Compound	log S_{exp}	MLR2	ANN	Klopman[a]	Kühne[b]
1	2,2′,4,5,5′-PCB	−6.77	−7.62	−7.63	−7.89	−7.47
2	Benzocain[b]	−2.32	−1.53	−1.61	−1.71	na
3	Aspirin	−1.61	−2.05	−2.02	−1.52	−1.93
4	Theophylline	−1.37	−1.67	−2.16	−1.07	0.54
5	Antipyrine[a]	0.39	−1.48	−1.62	(−2.76)	−1.90
6	Atrazine	−3.55	−3.61	−3.64	−3.50	−3.95
7	Phenobarbital	−2.34	−3.06	−3.16	−2.08	−2.49
8	Diuron	−3.76	−3.31	−3.32	−2.85	−3.38
9	Nitrofurantoin	−3.38	−3.12	−2.87	−2.19	−2.62
10	Phenytoin	−3.99	−4.11	−3.80	−3.47	−5.25
11	Diazepam[a]	−3.76	−4.63	−4.02	(−6.54)	−4.51
12	Testosterone	−4.07	−4.55	−3.94	−5.17	−4.62
13	Lindane	−4.60	−5.44	−4.73	−4.88	−4.52
14	Parathion	−4.29	−4.05	−3.97	−3.94	−4.59
15	Diazinon	−3.76	−5.24	−4.94	−5.29	−4.98
16	Phenolphthalein	−2.90	−4.84	−3.57	−4.48	−4.61
17	Malathion	−3.36	−3.45	−3.17	−2.94	−3.48
18	Chlorpyriphos	−5.67	−5.81	−6.15	−5.77	−3.75
19	Prostaglandin_E2[b]	−2.47	−4.53	−3.05	−4.21	na
20	*p*,*p*′-DDT	−8.08	−8.27	−7.70	−8.00	−7.75
21	a-chlordane	−5.35	−7.65	−7.27	−7.55	−6.51
		r^2_{pred}	0.80	0.84	0.79	0.72
		s	0.87	0.75	1.00	1.08

[a] Based on [7] with two outliers (antipyrine and diazepam).
[b] Based on [8] with two missing estimations.

presented in *table III*. The calculated and experimental aqueous solubility values of the training set, validation set and test set are plotted in *figure 1*.

The general applicability for the prediction ability of aqueous solubility was evaluated using the test set designed by Yalkowsky and Banerjee [20]. 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 V*. The present multiple linear regression and neural network models gave S.D.s, $s = 0.87$ and $s = 0.75(\pm 0.01)$. These results indicate that the use of neural networks gave clearly better predictions of log *S* values for this test set. In our previous study [18], 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 the models of Klopman et al. [7] and Kühne et al. [8]. An interesting point to note is that Kühne et al. used melting points in their group contribution approach and got a better fit for the training set of 694 compounds than the Klopman model using only group contributors for a training set of 483 compounds. However, Klopman's model made better predictions of the aqueous solubility values in the test set of 21 compounds ($s = 1.00$ and $n = 19$) after the exclusion of two outliers than Kühne's model ($s = 1.08$ and $n = 19$). Hence, we could also ask if the correction term for solid compound, melting point, really is necessary for group contribution approaches or other proposed methods as well. It might appear that the information encoded in melting point has been incorporated in this particular approach to modelling using group contribution descriptors, like atom-type E-state indices.

5. Conclusions

An accurate and generally applicable method for estimating aqueous solubility for a diverse set of 734 organic compounds based on multiple linear regression and artificial neural network modelling has been developed. Topological indices cannot account for three-dimensional and conformational effects. Topological indices, however, are attractive because they can be calculated easily, rapidly and are error free. The results of this study show that a practical solubility–prediction model can be constructed for a large and structurally diverse set of organic compounds especially with neural network modelling.

Acknowledgements

We would like to thank William Meylan from Syracuse Research Corporation for allowing us to use the KOWWIN program and the Technology Development Centre in Finland for financial support. The reviewers are thanked for their constructive and penetrating comments.

References

[1] Gillet V.J., Willet P., Bradshaw J., J. Chem. Inf. Comput. Sci. 38 (1998) 165–179.
[2] Milne G.W.A., Wang S., Nicklaus M.C., J. Chem. Inf. Comput. Sci. 36 (1996) 726–730.
[3] Ferguson A.M., Patterson D.E., Garr C.D., Underinger T.L., J. Biomol. Screening 1 (1996) 65–73.
[4] Lipinski C.A., Lombardo F., Dominy B.W., Feeney P.J., Adv. Drug Deliv. Rev. 23 (1997) 3–25.
[5] Ghose A.K., Viswanadhan V.N., Wendoloski J.J., J. Comb. Chem. 1 (1999) 55–68.
[6] Quarterman C.P., Bonham N.M., Irwin A.K., Eur. Pharm. Rev. 3 (1998) 27–32.
[7] Klopman G., Wang S., Balthasar D.M., J. Chem. Inf. Comput. Sci. 32 (1992) 474–482.
[8] Kühne R., Ebert R.-U., Kleint F., Scmidt G., Schüürmann G., Chemosphere 30 (1995) 2061–2077.
[9] Lee Y.-H., Myrdal P.B., Yalkowsky S.H., Chemosphere 33 (1996) 2129–2144.
[10] Nirmalakhandan N.N., Speece R.E., Environ. Sci. Technol. 22 (1988) 328–338.
[11] Bodor N., Huang M.-J., J. Am. Chem. Soc. 113 (1991) 9480–9483.
[12] Bodor N., Huang M.-J., J. Pharm. Sci. 881 (1992) 954–960.
[13] Patil G.S., J. Hazard. Mater. 36 (1994) 35–43.
[14] Nelson T.M., Jurs P.C., J. Chem. Inf. Comput. Sci. 34 (1994) 601–609.
[15] Sutter J.M., Jurs P.C., J. Chem. Inf. Comput. Sci. 36 (1996) 100–107.
[16] Huuskonen J., Salo M., Taskinen J., J. Pharm. Sci. 86 (1997) 450–454.
[17] Huibers P.D.T., Katritzky A.R., J. Chem. Inf. Comput. Sci. 38 (1998) 283–292.
[18] Huuskonen J., Salo M., Taskinen J., J. Chem. Inf. Comput. Sci. 38 (1998) 450–456.
[19] Mitchell B.E., Jurs P.C., J. Chem. Inf. Comput. Sci. 38 (1998) 489–496.
[20] Yalkowsky S.H., Banerjee S., Aqueous solubility, in: Methods of Estimation for Organic Compounds, Marcel Dekker, New York, 1992.

[21] Huuskonen J.J., Villa A.E.P., Tetko I.V., J. Pharm. Sci. 88 (1999) 229–233.

[22] Yalkowsky S.H., Dannelfelser R.M., The ARIZONA dATAbASE of Aqueous Solubility, College of Pharmacy, University of Arizona, Tucson, AZ, 1990.

[23] Kier L.H., Hall L.H., Pharm. Res. 7 (1990) 801–807.

[24] Hall L.H., Kier L.B., J. Chem. Inf. Comput. Sci. 35 (1995) 1039–1045.

[25] Hall L.H., Story C.T., J. Chem. Inf. Comput. Sci. 36 (1996) 1004–1014.