Prediction of Partition Coefficient Based on Atom-Type Electrotopological State Indices

JARMO J. HUUSKONEN,[†] ALESSANDRO E. P. VILLA,[‡] AND IGOR V. TETKO*,[‡,§]

Contribution from *Division of Pharmaceutical Chemistry, Department of Pharmacy, POB 56, FIN-00014 University of Helsinki, Finland, Laboratoire de Neuro-Heuristique, Institut de Physiologie, Université de Lausanne, Rue du Bugnon 7, Lausanne, CH-1005, Switzerland, and Biomedical Department, Institute of Bioorganic & Petroleum Chemistry, Murmanskaya 1, Kiev-660, 253660, Ukraine.*

Received June 29, 1998. Accepted for publication October 28, 1998.

Abstract □ The aim of this study was to determine the efficacy of atom-type electrotopological state indices for estimation of the octanol–water partition coefficient (log *P*) values in a set of 345 drug compounds or related complex chemical structures. Multilinear regression analysis and artificial neural networks were used to construct models based on molecular weights and atom-type electrotopological state indices. Both multilinear regression and artificial neural networks provide reliable log *P* estimations. For the same set of parameters, application of neural networks provided better prediction ability for training and test sets. The present study indicates that atom-type electrotopological state indices offer valuable parameters for fast evaluation of octanol–water partition coefficients that can be applied to screen large databases of chemical compounds, such as combinatorial libraries.

Introduction

The logarithm of the partition coefficient between octanol and water, log *P*, is extensively used to describe lipophilic or hydrophobic properties of chemical compounds. It has been shown that log *P* is a useful parameter to correlate transport properties of drug molecules, interactions between drugs and receptors, and changes in the structure of drugs with various biochemical or toxic effects of these compounds.[1] The measurement of log P throughout the synthesis of the compound and its subsequent experimental determination is time-consuming and expensive. Hence, there is a strong interest in the structure-based prediction of log P for rational development of new drugs before potential drug compounds have been synthesized.

Several approaches for computing log P on the basis of chemical structure have been proposed. Among others there are two essentially empirical methods for the estimation of log P: Rekker's *f* constant method and Leo and Hansch's fragment approach.[3] Both methods divide a compound into basic fragments and calculate its log P by the summation of the hydrophobic contributions of each fragment. However, the difficulties of these methods is how to fragment a molecule, especially large drug molecules, into basic fragments. New fragment methods (atomic fragments) were developed to overcome this problem.[4−6] These methods are conceptually simple and are able to give fast and accurate estimations for diverse organic compounds. However, correction factors are usually needed for complex structures to compensate for the interactions between functional groups.

Recently, Kier and Hall[7,8] introduced electrotopological state (E-state) indices for molecular structure description in which both electronic and topological characteristics are combined together. The E-state can be used in a group contribution manner and has been found to be useful in structure−property relationship studies. Using only these indices and neural network modeling, Hall and Story[9] were able to predict the boiling points and critical temperatures for a set of heterogeneous organic compounds.

In our recent studies we suggested methods for estimations of aqueous solubility, log S, of structurally related[10] and diverse sets[11] of drug compounds based on molecular topology and neural network modeling. The present study shows that the same indices can be successfully used to estimate log P coefficients, another important solubility property of drug compounds for drug design studies.

Experimental Section

Three hundred twenty-six drugs or related compounds from different structural classes were randomly selected from the Hansch−Leo compilation.[12] The partition coefficients of these compounds were represented as logarithm values, log P, and were in the range −2.11−5.9, corresponding to urea and thioridazine, respectively. This data set was divided into a training set of 300 compounds and a test set of 26 compounds (selected at random). An additional test set of 19 compounds[13] was included in the present study to compare our approach with currently available ones.

Structural parameters were calculated by Molconn-Z software (Hall Associated Consulting, Quincy, MA). Molecular weights and 32 atom-type E-state indices were calculated for each analyzed compound by SMILES line notation code. These 33 parameters were analyzed using multilinear regression (MLR) analysis and artificial neural networks (ANNs). MLR analysis was done with the SPSS package (v. 5.1, SPSS Inc., Chicago, IL) running on a Pentium PC. The ANNs employed in this study were fully connected feed-forward back-propagation networks with one hidden layer and bias neurons. ANN training was accomplished using the SuperSAB algorithm.[14] The logistic $f(x) = 1/(1+e^{-x})$ activation function was used both for hidden and output nodes. All calculated parameters and the set of parameters optimized by MLR were used for neural network training. 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 P values.

The avoidance of overfitting/overtraining has been shown to be an important factor for improvement of generalization ability and correct selection of variables in neural networks studies.[14,15] The Early Stopping over an Ensemble technique was used in the current study to accomplish this problem. A detailed description of this approach can be found elsewhere.[14,15] In brief, each analyzed ensemble was composed of $M = 100$ networks. The values calculated for analyzed cases were averaged over all M neural networks, and their means were used for computing statistical

* Address to whom all correspondence should be addressed. FAX: ++41-21-692.5505. E-mail: itetko@eliot.unil.ch.
[†] Division of Pharmaceutical Chemistry.
[‡] Laboratoire de Neuro-Heuristique.
[§] Biomedical Department.

10.1021/js980266s CCC: $18.00
Published on Web 01/05/1999

coefficients with targets. We used a subdivision of the initial training set into two equal learning/validation subsets. The first set was used to train the neural network, whereas the second one was used to monitor the training process measured by root-mean-square error. An early stopping point determined as a best fit of a network to the validation set was used to stop the neural network learning. It was shown that a neural network trained up to an early stopping point provided better prediction ability than a network trained to the error minimum for the learning set.[10,14] Thus, statistical parameters calculated at the early stopping point were used. The training was terminated by limiting the network run to 10 000 epochs (total number of epochs) or after 2000 epochs (local number of epochs) following the last improvement of root-mean-square error at the early stopping point.

The computer code for the ANN was programmed in ANSI C++. The calculations were performed at HP Workstation Cluster at the Swiss Center for Scientific Computing.

The quality of the models was tested in two ways. An analysis of predictive ability was done in terms of both predictive q^2 and actual prediction. Predictive q^2 in leave-one-out cross-validation was defined as

$$q^2 = (\mathrm{SSY} - \mathrm{PRESS})/\mathrm{SSY} \qquad (1)$$

Here, SSY is the sum of squares of the deviation between the observed log P values and their mean value and PRESS is the prediction error sum of squares obtained from leave-one-out (LOO) procedure. The standard deviation s_{LOO} was also considered. This coefficient was defined as

$$s_{\mathrm{LOO}} = [\mathrm{PRESS}/\mathrm{n}]^{1/2} \qquad (2)$$

where n was the number of compounds in the model. In addition, the two test sets already described were used to estimate the actual prediction of the models using square of correlation coefficient r^2 and standard deviation s.

Results and Discussion

A total of 32 atom-type E-state indices (Table 1) and molecular weights were used as parameters for analysis by MLR and ANNs. Stepwise and backward methods were employed in the regression analysis and the following 19 parameters yielded a satisfactory statistical model:

$$\log P = 0.228\ (\pm 0.021)\ \mathrm{SsCH_3} + 0.294\ (\pm 0.072)\ \mathrm{SdCH_2} + 0.275\ (\pm 0.023)\ \mathrm{SssCH_2} + 0.171\ (\pm 0.011)\ \mathrm{SaaCH} + 0.221\ (\pm 0.053)\ \mathrm{SsssCH} - 0.190\ (\pm 0.063)\ \mathrm{SdssC} + 0.215\ (\pm 0.044)\ \mathrm{SaasC} + 0.246\ (\pm 0.047)\ \mathrm{SaaaC} - 0.093\ (\pm 0.016)\ \mathrm{SsNH_2} - 0.198\ (\pm 0.028)\ \mathrm{SssNH} - 0.051\ (\pm 0.013)\ \mathrm{SaaN} - 0.360\ (\pm 0.039)\ \mathrm{SsssN} + 0.013\ (\pm 0.004)\ \mathrm{SdO} - 0.035\ (\pm 0.013)\ \mathrm{SssO} + 0.071\ (\pm 0.006)\ \mathrm{SsF} + 0.264\ (\pm 0.068)\ \mathrm{SdS} + 0.199\ (\pm 0.098)\ \mathrm{SssS} + 0.444\ (\pm 0.185)\ \mathrm{SaaS} + 0.187\ (\pm 0.025)\ \mathrm{SsCl} - 0.468\ (\pm 0.124).$$

$$(n = 300,\ r^2 = 0.87,\ s = 0.68,\ F = 87.5,\ q^2 = 0.83,\ s_{\mathrm{LOO}} = 0.71) \qquad (3)$$

where n is the number of compounds used in the fit, F is the overall F-statistics for the addition of each successive term, and values in parentheses are the 95% confidence limit of each coefficient. The correlation analysis for parameters in eq 3 showed that all pairwise correlations were $R < 0.5$, indicating a low multicollinearity as well.

For neural network studies, a preliminary analysis using all available parameters was done to determine the optimal number of neurons in the hidden layer chosen in the set 2, 3, 5, 7, 10, 15, and 30. The performance of neural networks was evaluated by LOO statistical coefficients calculated at early stopping point for the training data set. We found that q^2 increased (i.e., $q^2 = 0.824 \pm 0.002$, 0.825 ± 0.002, 0.829 ± 0.003) when the number of neurons in the hidden layer was changed from 2 to 5. However, further increase in the number of hidden neurons from 7 to 30 did not influence the prediction ability of neural networks (i.e., $q^2 = 0.829 \pm 0.002$, 0.828 ± 0.003, 0.829 ± 0.002, 0.830 ± 0.003). Thus, we fixed the number of neurons in the hidden layer equal to 5.

Table 1—The Atom-Type E-State Indices[a] Used in Multilinear Regression and Neural Network Models

			index value				
no.	symbol[b]	atom type[c]	min	max	training set	test set	MLR[d]
1	$\mathrm{SsCH_3}$	—$\mathrm{CH_3}$	0	12.9	158[e]	33	
2	$\mathrm{SdCH_2}$	=$\mathrm{CH_2}$	0	4.0	7	2	
3	$\mathrm{SssCH_2}$	—$\mathrm{CH_2}$—	–1.0	14.7	170	36	
4	StCH	≡CH	0	5.7	4	0	X
5	SdsCH	=CH—	0	6.0	60	5	X
6	SaaCH	..CH..	0	26.0	256	43	
7	SsssCH	—CH<	–4.7	3.1	107	18	
8	StC	≡C—	0	2.8	5	2	X
9	SdssC	>C=	–3.1	3.1	153	27	
10	SaasC	..C..	–3.8	6.8	225	42	
11	SaaaC	..C..	0	8.1	51	3	
12	SssssC	>C<	–6.8	0.5	70	11	X
13	$\mathrm{SsNH_2}$	—$\mathrm{NH_2}$	0	17.2	72	11	
14	SssNH	—NH—	0	8.1	100	22	
15	StN	≡N	0	10.2	2	2	X
16	SdsN	=N—	0	8.1	17	6	X
17	SaaN	..N..	0	19.8	77	7	
18	SsssN	>N—	0	7.6	95	24	
19	SddsN	—N≪	–0.7	0	6	0	X
20	SsOH	—OH	0	43.3	118	16	X
21	SdO	=O	0	47.6	167	26	
22	SssO	—O—	0	21.9	63	16	
23	SaaO	..O..	0	5.7	9	0	X
24	SsF	—F	0	39.9	26	3	
25	SsSH	—SH	0	4.1	3	0	X
26	SdS	=S	0	10.7	6	1	
27	SssS	—S—	0	3.7	24	8	
28	SaaS	..S..	0	1.7	7	0	
29	SdssS	=S<	–1.0	0	1	0	X
30	SddssS	>S≪	–8.9	0	29	3	X
31	SsCl	—Cl	0	12.2	23	9	
32	SsBr	—Br	0	3.4	1	0	X

[a] According to Hall and Kier.[8] [b] S states for the sum of the E-state values for a certain atom type or group; the sum for the hydroxyl groups is SsOH, for the ether or ester oxygen it is SssO, and for the keto oxygen it is SdO. [c] The formula of the atom type or group; the bond types between the heavy atoms are s = single (—), d = double (=), and a = aromatic (..). [d] The parameters that were eliminated in MLR regression are marked by X. [e] The number of compounds with the index.

Table 2—Comparison of the Predictive Ability of MLR and ANN Models

			training set		test set 1		test set 2	
model	params	#[a]	q^2	s_{LOO}	r^2	s	r^2	s
MLR	regressed	19	0.83	0.71	0.87	0.71	0.83	0.68
ANN1	regressed	19	0.84	0.69	0.90	0.62	0.84	0.62
ANN2	all	33	0.83	0.70	0.91 (0.93)[b]	0.60 (0.50)	0.87 (0.91)	0.57 (0.46)

[a] The number of input parameters in the model. [b] The results after excludance of loratidine and flufenamic acid are shown in the parentheses.

Neural networks represents essentially nonlinear methods of data analysis. However, the use of ANNs for the same set of parameters provided a prediction ability similar to that of MLRs for compounds in the training set. In LOO cross-validation procedures, ANNs gave $q^2 = 0.84$ and $s_{\mathrm{LOO}} = 0.69$ for the same set of parameters as in regression analysis, and $q^2 = 0.83$ and $s_{\mathrm{LOO}} = 0.70$ for all calculated parameters. The prediction ability of the MLR model given by the PRESS statistics, $s_{\mathrm{LOO}} = 0.71$, is only 0.03 log units

Table 3—Experimental and Estimated Log *P* Values for the Test Sets

A. test set 1

no.	compound	log P_{exp}	predicted	
			MLR	ANN2
1	acyclovir	–1.56	–1.70	–1.52
2	adrenalin	–1.37	0.10	–0.66
3	pyridoxine	–0.77	0.08	–0.59
4	isoniazid	–0.70	–0.14	–0.39
5	metaraminol	–0.27	0.46	–0.01
6	theophylline	–0.02	–0.22	–0.40
7	atenolol	0.16	1.13	0.83
8	sulpride	0.57	0.93	1.21
9	mescaline	0.78	1.25	1.18
10	primidone	0.91	0.97	0.82
11	carbutamide	1.01	0.85	0.76
12	ampicillin	1.35	0.91	1.12
13	clonidine	1.59	2.17	1.45
14	nalorphine	1.86	1.90	1.96
15	phenoxymethylpenicillin	2.09	1.40	1.67
16	hydrocortisoneacetate	2.19	2.01	2.05
17	lorazepam	2.39	3.81	3.28
18	dibenzepin	2.50	2.81	2.81
19	phenazine	2.84	2.71	2.38
20	ketoprofen	3.12	4.29	3.80
21	chlorpheniramine	3.38	4.61	4.19
22	disulfiram	3.88	4.06	3.18
23	fenethazine	4.20	2.84	3.39
24	methoxypromazine	4.90	4.15	4.26
25	trifluorperazine	5.03	4.95	4.57
26	loratidine	5.20	4.54	3.38

B. test set 2[a]

no.	compound	log P_{exp}	ANN2	XLOGP[b]	Moriguchi[b]	Rekker[b]	CLOGP[b]
1	chlorthiazide	–0.24	0.39	–0.58	–0.36	–0.68	–1.24
2	cimetidine	0.40	0.19	0.20	0.82	0.63	0.21
3	procainamide	0.88	1.32	1.27	1.72	1.11	1.11
4	trimethoprim	0.91	0.52	0.72	1.26	–0.07	0.66
5	chloramphenicol	1.14	1.29	1.46	1.23	0.32	0.69
6	phenobarbital	1.47	1.86	1.77	0.78	1.23	1.37
7	atropine	1.83	2.43	2.29	2.21	1.88	1.32
8	lidocaine	2.26	2.65	2.47	2.52	2.30	1.36
9	phenytoin	2.47	2.63	2.23	1.80	2.76	2.09
10	diltiazem	2.70	3.36	3.14	2.67	4.53	3.55
11	propranolol	2.98	3.22	2.98	2.53	3.46	2.75
12	diazepam	2.99	3.18	2.98	3.36	3.18	3.32
13	diphenhydramine	3.27	4.11	3.74	3.26	3.41	2.93
14	tetracaine	3.73	2.70	2.73	2.64	3.55	3.65
15	verapamil	3.79	4.33	5.29	3.23	6.15	3.53
16	haloperidol	4.30	4.41	4.35	4.01	3.57	3.52
17	imipramine	4.80	4.47	4.26	3.88	4.43	4.41
18	chlorpromazine	5.19	4.85	4.91	3.77	5.10	5.20
19	flufenamic acid	5.25	3.81	4.45	3.86	5.81	5.58

[a] This set was originally proposed by Moriguchi et al.[13] [b] The results calculated by XLOGP, CLOGP, and Moriguchi's and Rekker's methods are from ref 4.

higher than for the fitting model. Such a small increase indicates a robustness of the model.

The generalization ability of ANNs for the test sets was higher than that of MLR (Table 2). As in the case of the training set, the best predictions were calculated using all parameters. The results calculated using ANNs for the test set 2 ($n = 19$, $r^2 = 0.87$, $s = 0.57$) are comparable with those found using other known methods, such as CLOGP ($r^2 = 0.94$, $s = 0.44$), XLOGP ($r^2 = 0.89$, $s = 0.54$), and Moriguchi's method ($r^2 = 0.87$, $s = 0.63$), and are better than that of the Rekker's method ($r^2 = 0.84$, $s = 0.79$) (see Table 3).

The analysis of residuals showed that there were several compounds with a large calculation error in the training set. The compounds with a residual >1.4 log units, that is two times the standard deviation, are shown in Table 4. We found that two compounds in the analyzed test sets, loratidine (Test set 1) and flufenamic acid (Test set 2), were also characterized by high ANN prediction errors of 1.82 and 1.44 log units, respectively (Table 3). The elimination of these compounds from the test sets significantly improves prediction ability of ANNs (Table 2).

The low prediction ANN ability for loratidine and flufenamic acid and some other compounds from the training set can be explained by an analysis of the residuals. These two compounds have experimental log P values near to the largest value (5.90) in the training set. Let us note that for log P > 5.0, there were four compounds with large residual errors in the training set. However there were in total only seven compounds with such log P values in our training dataset. Thus, >50% of compounds with log P > 5.0 were poorly predicted. This result suggests that the

Table 4—Compounds with Large Prediction Errors in the Training Set

no.	compound[a]	log P_{exp}	MLR log P_{calc}	MLR resid	ANN2 log P_{calc}	ANN2 resid
4	methotrexate	–1.85	–0.84	–1.01	0.09	–1.94
6	penicillamine[a]	–1.78	–0.07	–1.71	0.14	–1.93
10	riboflavine	–1.46	0.00	–1.46	–0.24	–1.22
11	α-methylnoradrenalin	–1.43	0.11	–1.54	–0.36	–1.07
15	thiourea	–1.08	0.51	–1.59	0.62	–1.70
22	phenformin	–0.83	0.45	–1.28	0.58	–1.41
56	cephalotin	0.00	1.60	–1.60	1.41	–1.41
64	ranitidine	0.27	1.16	–0.89	1.77	–1.50
102	triamteren	0.98	–0.66	1.64	–0.72	1.70
119	minoxidil	1.24	–0.39	1.63	–0.46	1.69
144	2,4-dihydroxybencoic acid	1.63	0.29	1.34	–0.39	2.01
150	timolol	1.83	0.36	1.47	–0.08	1.91
174	clobazam	2.12	3.32	–1.20	3.52	–1.40
183	ketamine	2.18	3.20	–1.02	3.64	–1.46
190	salicylic acid	2.26	0.81	1.45	0.54	1.72
207	thiophenol	2.52	1.43	1.09	0.64	1.88
222	LSD	2.95	1.46	1.49	1.42	1.52
228	papaverine	2.95	4.11	–1.16	1.61	1.45
251	piroxicamine	3.06	1.79	1.27	1.75	1.68
261	dextromoramide	3.61	4.68	–1.07	2.40	1.50
293	DES	5.07	3.65	1.42	3.02	2.04
294	mefenic acid	5.12	3.22	1.90	3.19	1.93
294	tolfenamic acid	5.17	3.56	1.61	3.64	1.53
300	thioridazine	5.90	4.97	0.93	4.50	1.40

[a] Compounds with absolute value of residuals >1.4 log units (two times the standard deviation of the prediction error) for MLR and ANNs are shown.

number of analyzed molecules with high log P values did not provide a representative training data set for correct application of the analyzed methods and the training set should be extended by including more compounds with high log P values.

The atom-type E-state indices are used in a manner similar to group additive schemes. Each atom in the 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 the detailed structure of the molecule. An analysis of residuals (Table 4) provided some hints of which structural features makes difficulties for the proposed method. There were four carboxylic acids (2,4-dihydroxybencoic acid, salicylic acid, mefenic acid, and tolfenamic acid) and hydroxyl-containing compounds (riboflavine, α-methylnoradrenalin**,** pyridoxamine, and DES), which all have a large calculation error. Also, both of the compounds with a large calculation error in the test sets contain a carbonyl group; that is, loratidine contains the ester carbonyl and flufenamic acid has a carboxylic acid group.

The reason for low prediction results for the compounds just mentioned can be explained by the fact that all E-state values calculated for hydroxyl groups are used to calculate only one parameter, SsOH, and no division for different types (i.e., alcohol, phenols, and carboxylic acids) is made. Likewise, the parameter SdO accounts for carbonyl oxygen, making no distinction between neighbor atom type (i.e., carboxylic acids, amides, ketones, and esters). The effects of the atom group could be considered similarly to the group contribution approach.[5,6,16] In this approach, contribution values are calculated separately for hydroxyl groups in aliphatic and aromatic compounds and for carbonyl oxygen-containing compounds, the division is made between carboxylic acids, esters, aldehydes, ketones, and amides. In both cases, the contribution values varied considerably, depending on the type of the atom group. Nearly all compounds with a large calculation error in our training set contain different types of hydroxyl and carbonyl compounds. Thus, it might be possible that the atom-type E-state index values for hydroxyl (SsOH) and carbonyl (SdO) groups are not enough to discriminate them according to their binding environment. We suggest that the usefulness of E-state formalism could be improved by taking into account the binding environment of an atom type, especially in the case of hydroxyl and carbonyl groups, like in Meylan's atom/fragment contribution method[5] and in Klopman's group contribution approach.[6,16]

The atom-type E-state indices are similar to some extent to the group contribution variables (e.g., numbers of $-CH_2-$ groups instead of $SssCH_2$). It is possible to assume that the same results would be calculated if these variables were used instead of the atom-type E-state indices. The numbers of atom-types computed using Molconn-Z software and the molecular weights were fitted as input parameters for MLR and ANNs. The best MLR model calculated using stepwise and backward methods contained the following 15 indices:

$$\log P = 0.163\,(\pm 0.072)\,\mathrm{ISsCH_3} + 0.164\,(\pm 0.038)\,\mathrm{ISssCH_2} + 0.128\,(\pm 0.037)\,\mathrm{ISaaCH} + 0.260\,(\pm 0.062)\,\mathrm{ISaasC} + 0.471\,(\pm 0.090)\,\mathrm{ISaaaC} - 0.766\,(\pm 0.162)\,\mathrm{ISsNH_2} - 0.454\,(0.126)\,\mathrm{ISssNH} - 0.522\,(\pm 0.264)\,\mathrm{ISdsN} - 0.457\,(\pm 0.095)\,\mathrm{ISaaN} - 0.465\,(\pm 0.119)\,\mathrm{ISsssN} - 0.361\,(0.107)\,\mathrm{ISsOH} - 0.516\,(0.122)\,\mathrm{ISssO} + 0.254\,(\pm 0.111)\,\mathrm{ISsF} - 0.472\,(\pm 0.231)\,\mathrm{SddssS} + 0.0046\,(\pm 0.0012)\,\mathrm{MW} + 0.145\,(\pm 0.234)$$

$$(n = 300,\ r^2 = 0.55,\ s = 1.18,\ F = 23.4,\ q^2 = 0.49,\ s_{LOO} = 1.23) \quad (4)$$

where I refers to the number of groups corresponding to the appropriate atom-type E-state index and MW is the molecular weight (Table 2). The prediction ability of this equation was $r^2 = 0.59$, $s = 1.26$ and $r^2 = 0.13$, $s = 1.67$ for test sets 1 and 2, respectively. ANNs computed similar results; that is, $q^2 = 0.48$, $s_{\mathrm{LOO}} = 1.24$ for the training set and $r^2 = 0.59$, $s = 1.29$ and $r^2 = 0.22$, $s = 1.74$ for test sets 1 and 2, respectively. Thus, both MLR and ANN methods provided significantly worse results if the group contribution variables were used instead of atom-type E-state indices. To understand these results, recall that the atom-type E-state indices account both for electronic and topological characteristics of the molecular structure. It is also important that the range of the atom-type E-state indices (Table 1) is considerably larger than that calculated by counts of the number of corresponding groups. This fact also significantly contributes to the performance and high prediction ability of models based on the atom-type E-state indices.

The most important advantage of the present approach is that only 32 parameters and no corrections factors were used for coding each molecule, whereas other methods require hundreds of parameters.[2−6,17] We are well aware of the shortcoming of the present model. Topological indices cannot account for three-dimensional and conformational effects, which may play a major role for solubility properties of chemical compounds, as recently suggested by Palm and co-workers.[18] However, topological indices are attractive because they can be easily and rapidly calculated from the structures of analyzed compounds. This feature makes it possible to obtain fast estimations of the solubility properties of compounds belonging to large databases, such as virtual combinatorial libraries. Probably, these indices can also be used to improve the prediction ability of other methods that are based on calculation of theoretical descriptors derived from the molecular structures of compounds.[19,20]

The prediction of partition coefficients using atom-type E-state indices is accurate and provides reliable log P estimations that are comparable to those obtained by other methods. An advantage of the proposed approach is that the atom-type E-state indices can be quickly and easily estimated directly from the chemical structure of analyzed compounds. Moreover, the number of parameters is small. Thus, the present approach introduces a fast method for estimation of log P of chemical compounds.

References and Notes

1. Hansch, C.; Leo, A. *Substituent Constants for Correlation Analysis in Chemistry and Biology*, Wiley: New York, 1979.
2. Rekker, R. E. *Hydrophobic Fragment Constant*, Elsevier: New York, 1977.
3. Leo, A.; Jow, P.; Silipo, C.; Hansch, C. Correlation of Hydrophobic Constant (log P) from π and ×a6 Constants. *J. Med. Chem.* **1975**, *18*, 865−868.
4. Wang, R.; Fu, Y.; Lai, L. A. New Atom-Additive Method for Calculating Partition Coefficients. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 615−621.
5. Meylan, W. M.; Howard, P. H. Atom/Fragment Contribution Method for Estimating Octanol−Water Partition Coefficients. *J. Pharm. Sci.* **1995**, *83*, 83−92.
6. Klopman, G.; Li, J.-Y.; Wang, S.; Dimayuga, M. Computer Automated log P Calculations Based on an Extended Group Contribution Approach. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 752−781.
7. Kier, L. B.; Hall, L. H. An Electrotopological-State Index for Atoms in Molecules. *Pharm. Res.* **1990**, *7*, 801−807.
8. Hall, L. H.; Kier, L. B. Electrotopological State Indices for Atom Types: A Novel Combination of Electronic, Topological and Valence State Information. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1039−1045.
9. 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.
10. 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.
11. 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.
12. Hansch, C.; Leo, A.; Hoekman, D. *Exploring QSAR: Hydrophobic, Electronic, and Steric Constants*, American Chemistry Society: Washington, 1995; Vol. 2.
13. Moriguchi, I.; Hirono, S.; Nakagome, I.; Hirano, H. Comparison of Reliability of log P Values for Drugs Calculated by Several Methods. *Chem. Pharm. Bull.* **1994**, *42, 976−978*.
14. 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.
15. Tetko, I. V.; Villa, A. E. P. Efficient Partition of Learning Data Sets for Neural Network Training. *Neural Networks* **1997**, *10*, 1361−1374.
16. Klopman, G.; Wang. S.; Balthasar, D. M. Estimation of Aqueous Solubility of Organic Molecules by the Group Contribution Approach. Application to the Study of Biodegradiation. *J. Chem. Inf. Comput. Sci.* **1992,** *32*, 474−482.
17. Gombar, V. K.; Enslein, K. Assessment of N−Octanol/Water Partition Coefficient: When is the Assessment Reliable? *J. Chem. Inf. Comput. Sci.* **1996,** *36*, 1127−1134.
18. Palm, K.; Stenberg, P.; Luthman K.; Arthursson, P. Polar Molecular Surface Properties Predict the Intestinal Absorption of Drugs in Human. *Pharm. Res.* **1997**, *14*, 568−571.
19. Haeberlin, M.; Brinck, T. Prediction of Water-Octanol Partition Coefficients Using Theoretical Descriptors Derived from the Molecular Surface Area and the Electrostatic Potential. *J. Chem. Soc., Perkin Trans. 2* **1997**, 289−294.
20. Bodor, N.; Buchwald, P. Molecular Size Bazed Approach to Estimate Partition Properties for Organic Solutes. *J. Phys. Chem.* **1997**, *101*, 3404−3412.

Acknowledgments

This study was partially supported by the Technology Development Center in Finland (TEKES), INTAS-Ukraine grant 95-0060, and Swiss National Science Foundation grant FNRS 2150-045689.95. The authors thank the reviewers for their helpful remarks that improved the clarity and significance of our study.

JS980266S