Aqueous Solubility Prediction of Drugs Based on Molecular Topology and Neural Network Modeling

Jarmo Huuskonen, Marja Salo, and Jyrki Taskinen*

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

Received November 21, 1997

A method for predicting the aqueous solubility of drug compounds was developed based on topological indices and artificial neural network (ANN) modeling. The aqueous solubility values for 211 drugs and related compounds representing acidic, neutral, and basic drugs of different structural classes were collected from the literature. The data set was divided into a training set ($n = 160$) and a randomly chosen test set ($n = 51$). Structural parameters used as inputs in a 23−5−1 artificial neural network included 14 atom-type electrotopological indices and nine other topological indices. For the test set, a predictive $r^2 = 0.86$ and $s = 0.53$ (log units) were achieved.

INTRODUCTION

Aqueous solubility is an important determinant of the usefulness of a drug candidate that may have a marked impact on the whole process of drug discovery and development. A poor aqueous solubility is likely to hamper the bioavailability and cause other problems as well. The need for methods to predict the aqueous solubility on the basis of the chemical structure in the early stages of the discovery process has been accentuated by the advent of combinatorial chemistry. It has been noticed that the synthesis of combinatorial libraries of drug-like organic molecules tends to result in compounds with higher average lipophilicity, and presumably lower water solubility, than conventional synthetic strategies.[1] It has been suggested that computational screens could be utilized to select sublibraries for synthesis to ensure that the distribution of properties relevant to bioavailability corresponds to the range of known values of orally acting drugs.[2]

Theoretical prediction of the aqueous solubility of organic molecules is not feasible at present. Various empirical modeling approaches have been described using nonexperimental structural parameters.[3−8] These methods employ multiple regression or neural network modeling and varying ways of structural parametrization. In their recent review, Lipinski et al.[1] expressed the opinion that none of the published methods can be exploited for a relatively accurate prediction of the solubility of complex pharmaceutical drug candidates. One cause of this failure may be the nature of the training sets used. The training sets have been composed largely of noncomplex organic compounds devoid of, for instance, heterocyclic structures and multiple functional groups.

We have been working on the neural network modeling of aqueous solubility using training sets composed of pharmaceutical compounds. Recently, we concluded that reasonable predictions were attainable within a set of structural analogues using the topological indices of Kier and Hall as structural parameters.[9] The parameters used included connectivity indices, shape indices, and electrotopological state (E-state) indices. However, distinct models were required for each of the three structural classes studied.

Recently, Hall and Kier described an extension of the E-state index, called the atom-type E-state index.[10,11] Using only these indices and artificial neural network (ANN) modeling, Hall and Story[11] were able to predict the boiling points and critical temperatures for two sets heterogeneous, although structurally not very complex, organic compounds.

In an attempt to improve our ANN model we devised a structurally diverse training set of pharmaceuticals and related compounds and included the atom-type E-state indices in the parameter pool. The results of the ANN training and testing of the predictive ability of the model are described in this paper.

METHODS

The aqueous solubilities of 211 drugs and related compounds were taken from the literature.[12−25] The solubility values were expressed as log units of molar solubility (mol/L), and varied from 0.55 (ethambutol) to −5.60 (thioridazine). The data set was divided into a training set of 160 compounds and a randomly chosen test set of 51 compounds. The compounds in the training set and in the test set with the experimental water solubilities are presented in Tables 1 and 2, respectively.

Predictive ability outside the model was evaluated using two external test sets. In addition to the test set just described, the test set designed by Yalkowsky[26] and also used by Klopman et al.[4] was included in the present study to compare our ANN model with the currently available models.

Structural parameters were calculated by Molconn-Z software (Hall Associated Consulting, Quincy, MA). A total of 101 connectivity, shape, and atom-type E-state indices were calculated. The program calculates the number of

* Corresponding author. Telephone: +358 9 70859191. FAX: +358 9 70859556. E-mail: jyrki.taskinen[1]@helsinki.fi.

S0095-2338(97)00100-5 CCC: $15.00
Published on Web 02/24/1998

Table 1. Observed and Calculated Aqueous Solubilities for the Training Set

ID	name	log S_{obs}	log S_{calc}	resid	ref
1	ethambutol	0.545	0.616	−0.071	20
2	antipyrine	0.530	0.086	0.444	20
3	2-dimethylaminopteridine	0.360	−0.524	0.884	19
4	pteridine	0.021	0.157	−0.136	19
5	7-dimethylaminopteridine	−0.021	−0.624	0.603	19
6	2-methylpteridine	−0.094	−0.547	0.453	19
7	cefamandole	−0.143	−0.452	0.309	20
8	2-chlorpteridine	−0.699	−0.896	0.197	19
9	propranolol	−0.714	−1.073	0.359	20
10	7-ethyltheophylline	−0.757	−1.056	0.299	25
11	7-methylpteridine	−0.854	−0.662	−0.192	19
12	7-chlorpteridine	−0.876	−1.005	0.129	19
13	caffeine	−0.877	−0.931	0.054	20
14	paracetamol	−0.878	−2.167	1.289	24
15	7-methoxypteridine	−0.910	−1.319	0.409	19
16	7-*i*-butyltheophylline	−0.942	−1.459	0.517	25
17	4-dimethylpteridine	−1.021	−0.508	−0.513	19
18	2-methoxypteridine	−1.112	−1.198	0.086	19
19	4-methoxypteridine	−1.112	−1.121	0.009	19
20	6-chlorpteridine	−1.124	−1.013	−0.111	19
21	cytosine	−1.159	−1.802	0.643	25
22	5-methyl-5-allylbarbiturate	−1.160	−1.709	0.549	22
23	1-propyltheobromine	−1.207	−1.294	0.087	25
24	5-methyl-5-ethylbarbiturate	−1.228	−1.664	0.436	22
25	theophylline	−1.347	−1.513	0.166	25
26	5,5-diethylbarbiturate	−1.396	−1.754	0.358	22
27	5-methylcytosine	−1.444	−2.144	0.700	25
28	4-hydroxypteridine	−1.471	−1.675	0.204	19
29	5-ethyl-5-propylbarbiturate	−1.491	−2.049	0.558	18
30	uracil	−1.493	−0.951	−0.542	25
31	thymine	−1.499	−1.337	−0.162	25
32	pteridine-7-methyl thiol ether	−1.551	−2.183	0.632	19
33	7-*i*-butyl-8-methyltheophylline	−1.599	−1.907	0.308	25
34	5-ethyl-5-allylbarbiturate	−1.614	−1.801	0.187	22
35	1-butyltheobromine	−1.625	−1.581	−0.044	25
36	cyclopropane-1',5-spirobarbiturate	−1.655	−1.880	0.225	22
37	butethal	−1.661	−2.375	0.714	18
38	disopyramide	−1.701	−2.302	0.601	20
39	7-butyl-8-methyltheophylline	−1.745	−2.107	0.362	25
40	pteridine-2-methyl thiol ether	−1.754	−2.090	0.336	19
41	salicylic acid	−1.804	−1.581	−0.223	20
42	cefroxanide	−1.889	−2.079	0.190	20
43	orotic acid	−1.935	−2.313	0.378	25
44	talbutal	−2.016	−2.114	0.098	18
45	butatbital	−2.019	−2.222	0.203	18
46	5,5-diallylbarbiturate	−2.077	−1.866	−0.211	22
47	5-allyl-5-butylbarbiturate	−2.110	−2.458	0.348	18
48	cyclobarbital	−2.116	−2.507	0.391	18
49	7-hydroxypteridine	−2.124	−1.891	−0.233	19
50	butabarbital	−2.132	−2.035	−0.097	18
51	5-ethyl-5-*i*-propylbarbiturate	−2.148	−1.884	−0.264	18
52	5,5-diethyl-2-thiobarbiturate	−2.167	−2.328	0.161	22
53	chlordiazepoxide	−2.176	−2.100	−0.076	20
54	adenine	−2.177	−2.386	0.209	25
55	secobarbital	−2.223	−2.414	0.191	18
56	5-ethyl-5-(3-methylbut-2-enyl)barbiturate	−2.253	−2.085	−0.168	22
57	hypoxanthine	−2.289	−1.836	−0.453	25
58	vinbarbital	−2.296	−1.881	−0.415	18
59	2-aminopteridine	−2.298	−2.653	0.355	19
60	7-aminopteridine	−2.313	−2.706	0.393	19
61	benzocain	−2.320	−2.375	0.055	20
62	5-ethyl-5-pentylbarbiturate	−2.340	−2.744	0.404	18
63	5-ethyl-5-phenylbarbiturate	−2.340	−2.740	0.400	18
64	6-aminopteridine	−2.343	−2.709	0.366	19
65	cyclobutane-1',5-spirobarbiturate	−2.349	−2.180	−0.169	22
66	pteridine-4-methyl thiol ether	−2.365	−2.101	−0.264	19
67	5-allyl-5-phenylbarbiturate	−2.369	−2.846	0.477	18
68	5-methyl-5-phenylbarbiturate	−2.380	−2.703	0.323	22
69	5-ethyl-5-(1-methylbutyl)barbiturate	−2.387	−2.316	−0.071	22
70	5,5-dipropylbarbiturate	−2.410	−2.357	−0.053	18
71	5-methyl-5-(3-methylbut-2-enyl)barbiturate	−2.602	−2.019	−0.583	22
72	RTI12[a]	−2.620	−3.208	0.588	21
73	pteridine-4-thiol	−2.646	−2.282	−0.364	19
74	butallylonal	−2.647	−3.016	0.369	18

Table 1 (Continued)

ID	name	log S_{obs}	log S_{calc}	resid	ref
75	5-ethyl-5-(3-methylbutyl)barbiturate	−2.658	−2.346	−0.312	22
76	pteridine-7-thiol	−2.706	−2.441	−0.265	19
77	6-hydroxupteridine	−2.714	−1.894	−0.820	19
78	RTI10	−2.877	−3.816	0.939	21
79	phenolphthalein	−2.900	−3.149	0.249	12
80	heptabarbital	−2.906	−2.773	−0.133	18
81	hydrocortisone	−2.970	−3.657	0.687	12
82	cycloheptane-1',5-spirobarbiturate	−2.982	−3.424	0.442	22
83	hexethal	−3.049	−3.160	0.111	18
84	alclofenac	−3.125	−3.605	0.480	23
85	ketoprofen	−3.155	−3.650	0.495	23
86	cyclohexane-1',5-spirobarbiturate	−3.168	−2.961	−0.207	22
87	RTI14	−3.193	−4.632	1.439	21
88	5-ethyl-5-heptylbarbiturate	−3.218	−3.614	0.396	18
89	RTI8	−3.324	−3.649	0.325	21
90	RTI16	−3.360	−3.832	0.472	21
91	nitrofurantoin	−3.380	−2.668	−0.712	20
92	isoguanine	−3.401	−3.010	−0.391	25
93	ibuprofen	−3.420	−3.242	−0.178	23
94	deoxycorticosterone	−3.450	−4.144	0.694	12
95	cyclopentane-1',5-spirobarbiturate	−3.462	−3.347	−0.115	22
96	5-ethyl-5-(cyclohexylidene-2-ethyl)barbiturate	−3.529	−3.111	−0.418	22
97	RTI24	−3.535	−4.462	0.927	21
98	RTI18	−3.536	−4.129	0.593	21
99	bumetadine	−3.562	−3.397	−0.165	20
100	guanine	−3.577	−3.228	−0.349	25
101	trimazosin	−3.638	−3.583	−0.055	20
102	desipramine	−3.658	−4.505	0.847	13
103	5-ethyl-5-(1-methylbutyl)barbiturate	−3.679	−2.694	−0.985	22
104	triamcinolone	−3.680	−3.425	−0.255	12
105	RTI7	−3.680	−3.879	0.199	21
106	flurbiprofen	−3.740	−3.697	−0.043	23
107	RTI6	−3.762	−4.468	0.706	21
108	betamethasone	−3.770	−3.907	0.137	12
109	pentazocin	−3.803	−3.932	0.129	24
110	11α-hydroxyprogesterone	−3.820	−4.050	0.230	12
111	uric acid	−3.925	−3.542	−0.383	25
112	RTI22	−3.928	−4.380	0.452	21
113	oxazepam	−3.952	−3.763	−0.189	24
114	methyltestosterone	−3.990	−4.279	0.289	16
115	griseofulvin	−4.072	−4.159	0.087	24
116	testosterone	−4.080	−4.084	0.004	12
117	triazolam	−4.090	−4.685	0.595	20
118	RTI20	−4.114	−4.102	−0.012	21
119	triamcinolone diacetate	−4.130	−5.021	0.891	12
120	naproxen	−4.155	−4.215	0.060	23
121	perphenazine	−4.155	−4.564	0.409	13
122	5,5-diphenylbarbiturate	−4.196	−3.858	−0.338	22
123	RTI23	−4.207	−4.476	0.269	21
124	cortisone acetate	−4.210	−4.614	0.404	12
125	promazine	−4.301	−4.547	0.246	13
126	prednisolone acetate	−4.370	−3.951	−0.419	12
127	prochlorperazine	−4.398	−4.971	0.573	13
128	dihydroequilin	−4.402	−4.525	0.123	15
129	progesterone	−4.420	−4.214	−0.206	12
130	amitriptylin	−4.456	−4.074	−0.382	13
131	prasterone acetate	−4.458	−4.308	−0.150	14
132	5-ethyl-5-nonylbarbiturate	−4.462	−4.481	0.019	18
133	ethinylestradiol	−4.484	−5.035	0.551	15
134	RTI19	−4.554	−4.485	−0.069	21
135	cyclodecane-1',5-spirobarbiturate	−4.585	−4.704	0.119	22
136	deoxycorticosterone acetate	−4.630	−4.766	0.136	12
137	RTI13	−4.634	−3.714	−0.920	21
138	dihydroequilenin	−4.642	−5.100	0.458	15
139	thiopropazate	−4.699	−4.906	0.207	13
140	betamethasone-17-valerate	−4.710	−4.601	−0.109	12
141	RTI15	−4.741	−4.807	0.066	21
142	stanolone	−4.743	−4.483	−0.260	14
143	pecazine	−4.745	−4.826	0.081	13
144	RTI17	−4.749	−3.903	−0.846	21
145	RTI25	−4.799	−4.880	0.081	21
146	noretindrone acetate	−4.800	−4.838	0.038	14
147	indoprofen	−4.824	−4.752	−0.072	23
148	RTI2	−4.849	−4.145	−0.704	21

Table 1 (Continued)

ID	name	log S_{obs}	log S_{calc}	resid	ref
149	RTI3	−4.871	−4.383	−0.488	21
150	dexamethasone acetate	−4.900	−4.669	−0.231	12
151	estriol	−4.955	−4.639	−0.316	15
152	diclofenac	−5.097	−4.406	−0.691	23
153	RTI1	−5.153	−4.381	−0.772	21
154	equilenin	−5.249	−5.118	−0.131	15
155	fenbufen	−5.301	−4.307	−0.994	23
156	fluopromazine	−5.301	−4.957	−0.344	13
157	stanolone acetate	−5.348	−4.874	−0.474	14
158	RTI21	−5.360	−4.079	−1.281	21
159	cyclododecane-1',5-spirobarbiturate	−5.796	−4.989	−0.807	22
160	thioridazine	−5.824	−4.782	−1.042	13

[a] RTIs are reverse transcriptase inhibitors according to Morelock et al.,[21] and the compound numbering is same than in the original paper.

Table 2. Observed and Predicted Aqueous Solubilities for the Test Set 1.

ID	name	log S_{obs}	log S_{pred}	resid	ref
1	7-propyltheophylline	0.017	−1.079	1.096	25
2	aminopyrine	−0.360	−0.251	−0.109	20
3	4-methylpteridine	−0.466	−0.721	0.255	19
4	1-ethyltheobromine	−0.719	−0.865	0.146	25
5	6-methoxypteridine	−1.139	−1.124	−0.015	19
6	acetylsalicylic acid	−1.600	−2.099	0.499	20
7	5-*i*-propyl-5-allylbarbiturate	−1.708	−1.887	0.179	18
8	5,5-dimethylbarbiturate	−1.742	−1.919	0.177	22
9	7-butyltheophylline	−1.805	−1.327	−0.478	25
10	cycloethane-1',5-spirobarbiturate	−1.886	−2.144	0.258	22
11	2-hydroxypteridine	−1.947	−1.707	−0.240	19
12	4-aminopteridine	−2.313	−2.640	0.327	19
13	phenobarbital	−2.322	−2.965	0.643	18
14	prostaglandin E2	−2.470	−2.061	−0.409	12
15	xanthine	−2.483	−2.491	0.008	25
16	theobromine	−2.523	−1.664	−0.859	25
17	5-*i*-propyl-5-(3-methylbut-2-enyl)barbiturate	−2.593	−2.294	−0.299	22
18	pteridine-2-thiol	−2.629	−1.964	−0.665	19
19	reposal	−2.696	−2.697	0.001	18
20	5,5-di-*i*-propylbarbiturate	−2.766	−2.210	−0.556	18
21	quinidine	−2.812	−4.102	1.290	20
22	RTI11	−2.860	−3.094	0.234	21
23	RTI9	−3.043	−4.056	1.013	21
24	cyclopentane-1',5-spirobarbiturate	−3.060	−2.762	−0.298	22
25	prednisolone	−3.180	−3.175	−0.005	12
26	corticosterone	−3.240	−3.665	0.425	12
27	cortisone	−3.270	−3.795	0.525	12
28	5-*t*-butyl−5-(3-methylbut−2-enyl)barbiturate	−3.551	−3.039	−0.512	22
29	dexamethasone	−3.590	−3.788	0.198	12
30	lorazepam	−3.604	−4.274	0.670	20
31	diazepam	−3.754	−4.280	0.526	20
32	fenclofenac	−3.854	−3.990	0.136	23
33	5-ethyl-5-octylbarbiturate	−3.943	−3.926	−0.017	18
34	phenytoin	−3.990	−3.340	−0.650	20
35	prasterone	−4.064	−3.374	−0.690	14
36	imipramine	−4.187	−4.560	0.373	13
37	promethazine	−4.260	−3.613	−0.647	13
38	RTI5	−4.272	−4.835	0.563	21
39	triamcinolone acetonide	−4.310	−4.359	0.049	12
40	hydrocortisone acetate	−4.340	−4.172	−0.168	12
41	trifluoperazine	−4.523	−4.954	0.431	13
42	noretindrone	−4.630	−4.295	−0.335	16
43	RTI4	−4.706	−3.453	−1.253	21
44	estradiol	−4.845	−4.734	−0.111	15
45	sulindac	−5.000	−4.461	−0.539	23
46	chlorpromazine	−5.097	−4.856	−0.241	13
47	testosterone acetate	−5.184	−4.674	−0.510	14
48	equilin	−5.282	−4.681	−0.601	15
49	methyltestosterone acetate	−5.284	−4.837	−0.447	14
50	danazol	−5.507	−4.699	−0.808	24
51	estrone	−5.530	−5.009	−0.521	15

hydrogen bonding acceptors and donors, and indicates whether the compound is aromatic or not (aromaticity indicator). These values were also included in our parameter pool. The number of parameters was first reduced by excluding all the parameters that had nonzero values for a few compounds only ($n < 3$). Several combinations of ~30

nonintercorrelated ($r^2 < 0.5$) indices were then tested as inputs in ANNs, and final determination of the most useful parameters was performed by the neural network itself.

The neural network simulations were carried out using NeuDesk (v2.20, Neural Computional Sciences, UK). A three-layered, fully connected neural network was trained by the standard back-propagation learning algorithm. Before the training was started, the input and output values were scaled between 0.1 and 0.9, and the adjustable weights between neurons were given random values of between −0.5 and 0.5. The learning rate and momentum parameters were set at 0.1 and 0.9, respectively. The training endpoint was determined on the basis of the average training error (E), which is the mean square error between the target and actual outputs. The optimal training endpoint was searched for by overtraining the network. Networks with 2−6 neurons in the hidden layer were studied. The network architecture and the training endpoint giving the highest coefficient of determination, r^2_{pred} and lowest standard error s for the predictions of the test set were then used in the predictions. The predicted aqueous solubilities are the averaged log S values from 10 independent ANN runs.

The most significant parameters (inputs) in ANN were identified by the sensitivity method.[27] The sensitivity of each input was calculated as the sum of the absolute magnitude of the weights connected to it. The least sensitive, and therefore the least significant, inputs were excluded. The sensitivity method was tested in two ways, with and without a noise.[28] One uniformly distributed random data point was included as input in ANN and was generated by the MS Excel 5.0 random number generator tool. The purpose of the random data point was to give the network a choice between real data and random data.

RESULTS AND DISCUSSION

Data on the aqueous solubility of 211 pharmaceutical compounds or compounds with related complex structures were collected from published papers (Tables 1 and 2). The data set was divided into a training set of 160 compounds for developing the ANN models, and a randomly chosen test set of 51 compounds (Test set 1) for evaluating the predictive ability of the models. Another test set of 21 compounds (Test set 2), used earlier by other authors,[4,26] was also used to allow comparison of the predictions with earlier results.

The Molconn-Z program calculated 101 topological indices for the compound set studied, including 75 connectivity indices and 26 atom-type E-state indices. A subset of 31 parameters was selected as inputs for training the ANN model. All the 21 atom-type E-state indices that were represented by three compounds at least were included. Following a number of trials, 10 additional parameters were chosen to describe properties such as aromaticity, hydrogen bonding, size, branching, etc. Using these inputs, the best performance of the network was achieved with five neurons in the hidden layer. Networks with fewer hidden neurons were unable to reach the desired training level, and if the number of neurons was higher, the networks gave poorer predictions. The optimal training endpoint, 0.07, required 200−300 training epochs when an ANN architecture of 31−5−1 was used.

The number of inputs could be decreased without a loss of predictive ability of ANN by using the sensitivity method.

Table 3. Structural Parameters in the 31-5-1 ANN Model

(A) topological indices

symbol	explanation
Ar	aromaticity indicator[a]
HBD	number of hydrogen bond donors
HBA	number of hydrogen bond acceptors
$^2\chi$	path 2 simple connectivity index
$^3\chi_c$	cluster 3 simple connectivity index
$^6\chi^v_{ch}$	chain 6 valence connectivity index
$^7\chi^v_{ch}$	chain 7 valence connectivity index[b]
$^9\chi^v_{ch}$	chain 9 valence connectivity index
$^{10}\chi^v_{ch}$	chain 10 valence connectivity index
$^3\kappa_\alpha$	valence kappa 3 index

(B) atom type E-state indices[c]

symbol[d]	atom type[e]
$SsCH_3$	- CH_3
$SdCH_2$[b]	= CH_2
$SssCH_2$	- CH_2 -
SdsCH	= CH -
SaasCH	.. CH ..
SsssCH	> CH -
SdssC	= C <
SaasC	.. C ..
SssssC	> C <
$SsNH_2$[b]	- NH_2
SssNH[b]	- NH -
SdsN[b]	= N -
SaaN	.. N ..
SsssN	> N -
SsOH	- OH
SdO	= O
SssO	- O -
SsF[b]	- F
SsSH[b]	- SH
SdS[b]	= S
SssS	- S -
SsCl[b]	- Cl

[a] For aromatic compounds, Ar = 1 and for nonaromatic Ar = 0. [b] These parameters were eliminated by the sensitivity method in ANN modeling. [c] According to Kier and Hall.[10] [d] S states for the sum of the E-state values for a certain atom type or group; the hydroxyl group is SsOH, the ether or ester oxygen is SssO, and the keto oxygen is SdO. [e] The formula of the atom type or group; the bond types between the heavy atoms are s = single (−), d = double (=), and a = aromatic (..).

When a noise was added the differences between important and noimportant inputs became clearer and the choice of most significant parameters could be made. The final ANN model contained 23 structural parameters (Table 3) as inputs and five neurons in the hidden layer. Elimination of the inputs not contributing significantly to the prediction did not improve the generalization ability of the network we used as it has done in some studies.[28]

The estimated aqueous solubilities of the compounds in the training set are presented in Table 1. The predicted aqueous solubilities for test set 1 are presented in Table 2 and those for test set 2 in Table 4. The calculated and experimental solubilities of the training and test sets are plotted in Figures 1−3.

The present model was able to estimate, with a reasonable degree of accuracy, most of the aqueous solubilities of the training set ($r^2 = 0.90$ and $s = 0.46$, $n = 160$) and the test set 1 ($r^2_{pred} = 0.86$, $s = 0.53$, $n = 51$). There were only four compounds in the training set (paracetamol, RT inhibitors 14 and 21, and thioridazine) and four compounds in the

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

ID	name	log S_{obs}	log S_{pred}	resid
1	antipyrine	0.390	−0.342	0.732
2	theophylline	−1.370	−1.757	0.387
3	aspirin	−1.610	−1.735	0.125
4	benzocain	−2.320	−2.567	0.247
5	phenobarbital	−2.340	−2.714	0.374
6	prostaglandin E2	−2.470	−1.590	−0.880
7	phenolphthalein	−2.900	−3.245	0.345
8	malathion	−3.360	−2.054	−1.306
9	nitrofurantoin	−3.380	−2.488	−0.892
10	atrazine	−3.550	−2.446	−1.084
11	diuron	−3.760	−3.728	−0.032
12	diazepam	−3.760	−4.193	0.433
13	diazinon	−3.760	−2.808	−0.952
14	phenytoin	−3.990	−3.843	−0.147
15	testosterone	−4.070	−4.050	−0.020
16	parathion	−4.290	−2.651	−1.639
17	lindane	−4.600	−3.403	−1.197
18	chlordane	−5.350	−4.965	−0.385
19	chlorpyrifos	−5.670	−3.007	−2.663
20	2,2',4,5,5'-PCB	−6.770	−4.836	−1.934
21	DDT	−8.080	−4.756	−3.324

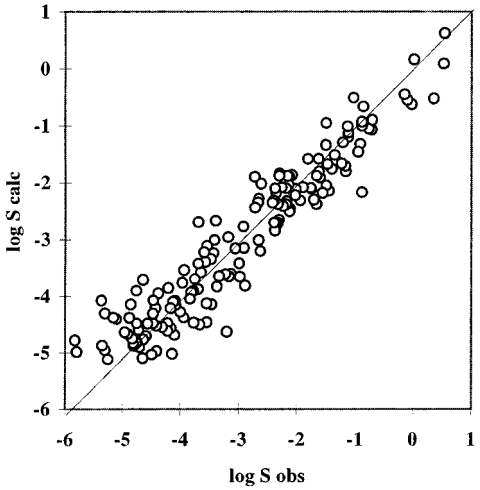

Figure 1. The calculated versus observed aqueous solubilities for the training set ($n = 160$).

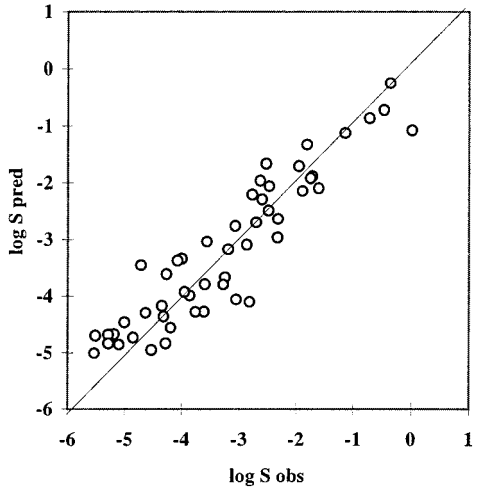

Figure 2. The predicted versus observed aqueous solubilities for test set 1.

test set (7-propyltheophylline, quinidine, and RT inhibitors 4 and 9) that had an estimated error >1.0 log units. The greatest deviations were found in the prediction of quinidine and RT inhibitor 4, with an error of 1.2 log unit; therefore, these compounds could not be regarded as outliers.

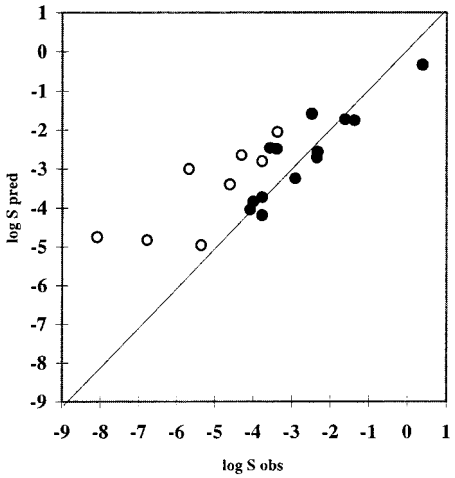

Figure 3. The predicted versus observed aqueous solubilities for test set 2. Key: (○) for the subgroup of eight compounds containing phosphate or thiophosphate group and polychlorinated hydrocarbons; (●) remaining 13 compounds.

Klopman et al.[4] tested the general applicability of their group contribution model for the prediction of aqueous solubility by the test set designed by Yalkowsky.[26] This test set contains 21 commonly used compounds of pharmaceutical and enviromental interest. The present ANN model and Klopman's model gave the same standard deviations, $s = 1.25$, for this test set. However, it is obvious from Figure 3 that our ANN model gave poor predictions for the subgroup of pesticides containing phosphate or thiophosphate group and polychlorinated hydrocarbons ($s = 1.90$, $n = 8$). On the other hand, the ANN model was successful with the other subgroup ($s = 0.55$, $n = 13$), which was composed mainly of pharmaceutical compounds from the chemical classes represented by the training set. The three compounds chlorpyrifos, 2,2',4,5,5'-PCB, and DDT, which were most poorly predicted by the ANN model, have a very poor aqueous solubility (log $S < -5.5$) that is outside the range of our training set.

We are well aware of the shortcomings of the present model. The data set is limited in number, in the range of aqueous solubilities, and in structural variation. Many functional groups common to drug molecules were totally absent or represented only by a single compound. Topological indices cannot account for three-dimensional and conformational effects, which may have a major role, as suggested recently by Palm and co-workers,[29,30] in the case of some solubility-dependent properties. Topological indices, however, are attractive because they can be calculated easily and rapidly. The results of this study show that a practical solubility-predicting model can be constructed for a structurally diverse set of drug compounds with neural network modeling.

ACKNOWLEDGMENT

The authors thank the Technology Development Centre in Finland (TEKES) for financial support.

REFERENCES AND NOTES

(1) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computional Approach to Estimate Solubility and

Permeability in Drug Discovery and Development Settings. *Adv. Drug. Del. Rev.* **1997**, *23*, 3−25.
(2) Salemme, F. R.; Spurlino, J.; Bone, R. Serendipity Meets Precision: The Integration of Structure-Based Drug Design and Combinatorial Chemistry for Efficient Drug Discovery. *Structure* **1997**, *5*, 319−324.
(3) Meylan, W. M.; Howard, P. H.; Boethling, R. S. Improved Method for Estimating Water Solubility from Octanol/Water Partition Coefficient. *Environ. Toxicol. Chem.* **1996**, *15*, 100−106.
(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. Comp. Sci.* **1992**, *32*, 474−482.
(5) Patil, G. S. Prediction of Aqueous Solubility and Octanol−Water Partition Coefficient for Pesticides Based on Their Molecular Structure. *J. Hazard. Mater.* **1994**, *36*, 35−43.
(6) 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.
(7) Nelson, T. M.; Jurs, P. C. Prediction of Aqueous Solubility of Organic Compounds. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 601−609.
(8) Sutter, J. M.; Jurs, P. C. Prediction of Aqueous Solubility for a Diverse Set of Heteroatom- Containing Organic Compounds Using a Quantitative Structure−Property Relationship. *J. Chem. Inf. Comp. Sci.* **1996**, *36*, 100−107.
(9) 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.
(10) 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. Comp. Sci.* **1995**, *35*, 1039−1045.
(11) 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.
(12) Yalkowsky, S. H.; Valvani, S. C. Solubility and Partitioning I: Solubility of Nonelectrolytes in Water. *J. Pharm. Sci.* **1980**, *69*, 912−922.
(13) Green, A. L. Ionization Constants and Water Solubilities of Some Aminoalkylphenothiazine Tranquillizers and Related Compounds. *J. Pharm. Pharmacol.* **1966**, *19*, 10−16.
(14) Bowen, D. B.; James, K. C.; Roberts, M. J. An Investigation of the Distribution Coefficients of Some Androgen Esters Using Paper Chromatography. *J. Pharm. Pharmacol.* **1970**, *22*, 518−522.
(15) Hurwitz, A. R.; Liu, S. T. Determination of Aqueous Solubility and pK_a Values of Estrogens. *J. Pharm. Sci.* **1977**, *66*, 624−627.
(16) Higuchi, T.; Shih, F.-M.; Kimura, T.; Rytting, J. H. Solubility Determination of Barely Aqueous-Soluble Organic Solids. *J. Pharm. Sci.* **1979**, *68*, 1267−1272.
(17) Yalkowsky, S. H.; Valvani, S. C.; Roseman, T. J. Solubility and Partitioning VI: Octanol Solubility and Octanol−Water Partition Coefficients. *J. Pharm. Sci.* **1983**, *72*, 866−870.
(18) Pinal, R.; Yalkowsky, S. H. Solubility and Partitioning VII: Solubility of Barbiturates in Water. *J. Pharm. Sci.* **1987**, *76*, 75−85.
(19) Grant, D. J. W.; Higuchi, T. in *Solubility Behavior of Organic Compounds*; Saunders, Jr., W. H., Ed.; Wiley: New York, 1990; p 26.
(20) Herman, R. A.; Veng-Pedersen, P. Quantitative Structure-Pharmacokinetic Relationships for Systemic Drug Distribution Kinetics not Confined to a Congeneric Series. *J. Pharm. Sci.* **1994**, *83*, 423−428.
(21) Morelock, M. M.; Choi, L. L.; Bell, G. L.; Wright, J. L. Estimation and Correlation of Drug Water Solubility with Pharmacological Parameters Required for Biological Activity. *J. Pharm. Sci.* **1994**, *83*, 948−952.
(22) Prankerd, R. J.; McKeown, R. H. Physico-Chemical Properties of Barbituric Acid Derivatives: IV. Solubilities of 5,5-disubstituted Barbituric Acids in Water. *Int. J. Pharm.* **1994**, *112*, 1−15.
(23) Fini, A.; Fazio, G.; Feroci, G. Solubility and Solubilization Properties of Non-Steroidal Antiinflammatory Drugs. *Int. J. Pharm.* **1995**, *126*, 95−102.
(24) Mithani, S. D.; Bakatselou, V.; tenHoor, C. N.; Dressman, J. B. Estimation of the Increase in Solubility of Drugs as a Function of Bile Salt Concentration. *Pharm. Res.* **1996**, *13*, 163−167.
(25) Pogliani, L. Modeling Purines and Pyrimidines With the Linear Combination of Connectivity Indices − Molecular Connectivity LCCI-MC Methodol. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1082−1091.
(26) Yalkowsky, S. H.; Banerjee, S. In *Aqueous Solubility, Methods for Estimation for Organic Compounds*; Marcel Dekker: New York, 1992; pp 128−148.
(27) 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.
(28) Maddalena, D. J.; Johnston, G. A. R. Prediction of Receptor Properties and Binding Affinity of Ligands to Benzodiazepine/GABA Receptors Using Artificial Neural Networks. *J. Med. Chem.* **1995**, *38*, 715−724.
(29) Palm, K.; Luthman, K.; Ungell, A. L.; Strandlund, G.; Artursson, P. Correlation of Drug Absorption with Molecular Surface Properties. *J. Pharm. Sci.* **1996**, *85*, 32−39.
(30) Palm, K.; Stenberg, P.; Luthman, K.; Artursson, P. Polar Molecular Surface Properties Predict the Intestinal Absorption of Drugs in Humans. *Pharm. Res.* **1997**, *14*, 568−571.

CI970100X