QSPR Prediction of Aqueous Solubility of Drug-Like Organic Compounds

Jahanbakhsh GHASEMI* and Saadi SAAIDPOUR

Chemistry Department, Faculty of Sciences, Razi University; Kermanshah, 67149, Iran.
Received November 22, 2006; accepted January 10, 2007

A quantitative structure property relationship (QSPR) study was performed to develop a model that relates the structures of 150 drug organic compounds to their aqueous solubility ($\log S_w$). Molecular descriptors derived solely from 3D structure were used to represent molecular structures. A subset of the calculated descriptors selected using stepwise regression that used in the QSPR model development. Multiple linear regression (MLR) is utilized to construct the linear QSPR model. The applied multiple linear regression is based on a variety of theoretical molecular descriptors selected by the stepwise variable subset selection procedure. Stepwise regression was employed to develop a regression equation based on 110 training compounds, and predictive ability was tested on 40 compounds reserved for that purpose. The final regression equation included three parameters that consisted of octanol/water partition coefficient ($\log P$), molecular volume (MV) and hydrogen bond forming ability (HB), of the drug molecules, all of which could be related to solubility property. Application of the developed model to a testing set of 40 drug organic compounds demonstrates that the new model is reliable with good predictive accuracy and simple formulation. The use of descriptors calculated only from molecular structure eliminates the need for experimental determination of properties for use in the correlation and allows for the estimation of aqueous solubility for molecules not yet synthesized. The prediction results are in good agreement with the experimental values. The root mean square error of prediction (RMSEP) and square correlation coefficient (R^2) of prediction of $\log S_w$ were 0.0959 and 0.9954, respectively.

Key words aqueous solubility; quantitative structure property relationship (QSPR); descriptor; multiple linear regression (MLR); prediction

The aqueous solubility of organic compounds is an important molecular property, playing a large role in the behavior of compounds in many areas of interest. Given the importance of solubility, a means of prediction based solely on molecular structure should prove a useful tool, as many compounds exist for which the solubility simply is not available. The solubility of chemicals and drugs in the water phase has an essential influence on the extent of their absorption and transport in a body. That is why solubility is considered to be a very important parameter in current ADMET (Absorption, Distribution, Metabolism, Excretion, and Toxicity) research.[1—5)]

Water solubility plays a key role in areas such as drug dosage, anesthesiology, corrosion of metals, transport fate of pollutants in terrestrial, aquatic and atmospheric ecosystems, deposition of minerals and composition of ground waters, and availability of oxygen and other gases in life support systems. The widespread relevance of water solubility data to many branches and disciplines of science, medicine, technology, and engineering has led to the development of several models to predict water solubility. Hence, it was deemed advantageous to develop a model to predict water solubility using only theoretically derived descriptors.[6—9)] Comparing with the time-consuming experimental procedures to determine aqueous solubility directly, reliable computational methods to predict aqueous solubility are more popular in today's research.[10—12)] There are some reports about the applications of QSPR approaches to predict the aqueous solubility of organic compounds.[13—19)] In our previous papers, we reported on the application of QSPR techniques in the development of a new, simplified approach to prediction of compounds properties.[20—22)] Several articles have published with MLR models for the prediction of aqueous solubility.[23—26)]

In a QSPR study, a mathematical model is developed which relates the structure of a set of compounds to a physical property such as aqueous solubility. In a QSPR study is that there is some sort of relationship between the physical property of interest and structural descriptors. These descriptors are numerical representations of structural features of molecules that attempt to encode important information that causes structurally different compounds to have different physical property values. Even though the descriptors used to build a QSPR model can be empirical, it is generally more useful to use descriptors derived mathematically from the 3D molecular structure, since this allow any relationship so derived to be extended to the prediction of the property for unavailable compounds. In this work a QSPR study is performed, to develop model that relate the structures of a heterogeneous group of 150 drug-like compounds to their aqueous solubility. The stepwise MLR was used to select the most informative descriptors from the calculated descriptors by Molecular Modeling Pro Plus software. The selected descriptors were used to develop a MLR model for predicting the solubility for 40 drug compounds in water at 25 °C. The aim of this work was to investigate molecular descriptors important in determining aqueous solubility.

Data and Methods

The QSPR model for the estimation of the $\log S_w$'s of various drug organic compounds is established in the following six steps: the molecular structure input and generation of the files containing the chemical structures is stored in a computer–readable format; quantum mechanics geometry is optimized with a semi-empirical (AM1) method; structural descriptors are computed; structural descriptors are selected; and the structure-$\log S_w$ model is generated by the multiple linear regression and statistical analysis.

Data Set All solubility data for all 150 compounds was taken from the literature.[27)] These values were converted from mole/liter to logarithm of drug solubility ($\log S_w$). These were measured at 25 °C in aqueous solution. The data set was split into a training set (110 compounds) and a prediction set (40 compounds). The solubility's of these compounds are deposited in Journal log as supporting material (see Table 1).

Computer Hardware and Software All calculations were run on a Pentium IV personal computer with windows XP operating system. The

* To whom correspondence should be addressed. e-mail: Jahan.ghasemi@gmail.com

Table 1. Experimental Values of log S_w for Drug-Like Organic Compounds in Water at 25 °C for Training (a) and Pediction (b) Sets

No.	Name	log S_w	No.	Name	log S_w
1	Naphthacene[a]	−8.69	76	5-Methyl-5-(3-methylbut-2-enyl)barbiturate[a]	−2.602
2	Chrysene[a]	−8.06	77	5-i-Propyl-5-(3-methylbut-2-enyl)barbiturate[a]	−2.593
3	3,4-Benzopyrene[a]	−7.82	78	Propylparaben[b]	−2.557
4	Hexachlorobenzene[a]	−7.76	79	Lomefloxacin[a]	−2.533
5	Benzanthracene[a]	−7.21	80	5,5-Dipropylbarbiturate[b]	−2.527
6	Etomidate[a]	−6.735	81	Carbofuran[a]	−2.5
7	Anthracene[a]	−6.38	82	Acetazolamide[a]	−2.489
8	Fenbuconazole[b]	−6.226	83	Amobarbital[a]	−2.47
9	Pyrene[a]	−6.18	84	Isocarboxazid[b]	−2.461
10	Fenbufen[a]	−5.301	85	Heptobarbital[a]	−2.38
11	Fludioxonil[a]	−5.21	86	Phenacetin[b]	−2.371
12	Phenanthrene[a]	−5.15	87	Phenobarbital[b]	−2.366
13	Diclofenac[b]	−5.097	88	Pteridine-4-methyl-thiol[b]	−2.365
14	Fenpiclonil[a]	−5.074	89	Cyclobutane-spirobarbiturate[a]	−2.349
15	Fluorene[a]	−4.92	90	Ethylparaben[a]	−2.346
16	Indoprofen[a]	−4.824	91	5-Allyl-5-phenylbarbiturate[a]	−2.346
17	Fenoxycarb[b]	−4.719	92	5-Ethyl-5-pentylbarbiturate[a]	−2.34
18	Flufenamic acid[a]	−4.623	93	Glutethimide[a]	−2.337
19	G-BHC (Lindane)[a]	−4.6	94	Secbutabarbital[b]	−2.333
20	Acenaphthene[b]	−4.59	95	Cyclobarbital[b]	−2.273
21	Iopanoic acid[a]	−4.58	96	Sulfamethazine[a]	−2.268
22	Diflunisal[a]	−4.479	97	5-Ethyl-5-(3-methylbut-2-enyl)barbiturate[a]	−2.253
23	5-Ethyl-5-nonylbarbiturate[a]	−4.462	98	Propylthiouracil[a]	−2.185
24	Amitriptyline[a]	−4.456	99	Idobutal[a]	−2.172
25	1,3,5-Trichlorobenzene[a]	−4.44	100	2-Naphthol[a]	−2.159
26	Haloperidol[a]	−4.429	101	Probarbital[b]	−2.153
27	Diphenyl[b]	−4.34	102	Atropine[a]	−2.124
28	Phenytoin[a]	−4.226	103	Butalbital[a]	−2.119
29	5,5-Diphenylbarbiturate[a]	−4.196	104	Camphor[a]	−2.086
30	Naproxen[a]	−4.155	105	Minoxidil[a]	−1.978
31	1,4-Dibromobenzene[b]	−4.07	106	Salicylamide[a]	−1.836
32	Oxazepam[b]	−3.952	107	7-Butyltheophylline[a]	−1.805
33	5-Ethyl-5-octylbarbiturate[a]	−3.943	108	Salicylic acid[a]	−1.804
34	Fenchlorphos[a]	−3.905	109	Allobarbital[b]	−1.796
35	Fenclofenac[a]	−3.854	110	Pteridine-2-methyl-thiol[a]	−1.754
36	Methyclothiazide[a]	−3.778	111	7-Butyl-8-methyltheophylline[a]	−1.745
37	Mefenamic acid[a]	−3.77	112	Aspirin[b]	−1.733
38	1,2,3-Trichlorobenzene[a]	−3.76	113	Saccharin[a]	−1.725
39	Diuron[a]	−3.76	114	Aprobarbital[a]	−1.71
40	Flurbiprofen[a]	−3.74	115	Methyl-*p*-hydroxybenzoate[b]	−1.705
41	Naphthalene[b]	−3.61	116	Baclofen[a]	−1.696
42	Lorazepam[a]	−3.604	117	Butobarbitone (Butethal)[a]	−1.686
43	Bumetanide[b]	−3.562	118	1-Butyltheobromine[b]	−1.625
44	5-*t*-Butyl-5-(3-methylbut-2-enyl)barbiturate[b]	−3.551	119	5-Ethyl-5-allylbarbiturate[b]	−1.614
45	Linuron[a]	−3.521	120	Cimetidine[b]	−1.613
46	Atrazine[a]	−3.489	121	7-Isobutyl-8-methyltheophylline[a]	−1.599
47	Melphalan[b]	−3.485	122	Benzoic acid[a]	−1.555
48	Isoproturon[a]	−3.469	123	Pteridine-7-methyl-thiol[a]	−1.551
49	Fluometuron[b]	−3.463	124	5-Ethyl-5-propylbarbiturate[b]	−1.491
50	Ibuprofen[a]	−3.42	125	5,5-Diethylbarbiturate[b]	−1.41
51	Nalidixic acid[a]	−3.366	126	Acetanilide[a]	−1.398
52	Carbamazepine[a]	−3.294	127	Salbutamol[a]	−1.224
53	5-Ethyl-5-heptylbarbiturate[a]	−3.218	128	Sulfamerazine[a]	−1.218
54	Cyclohexane-spirobarbiturate[a]	−3.168	129	1-Propyltheobromine[b]	−1.207
55	Ketoprofen[a]	−3.155	130	5-Methyl-5-ethylbarbiturate[a]	−1.162
56	Butamben[a]	−3.131	131	5-Methyl-5-allylbarbiturate[b]	−1.16
57	Alclofenac[a]	−3.125	132	6-Chlorpteridine[b]	−1.124
58	Butylparaben[b]	−3.101	133	2-Methoxypteridine[a]	−1.112
59	Hexethal[a]	−3.049	134	Acetaminophen[a]	−1.074
60	Hydroflumethiazide[a]	−3.043	135	4-Dimethylaminopteridine[b]	−1.021
61	Heptabarbital[b]	−3.000	136	Methocarbamol[b]	−0.985
62	Cycloheptane-spirobarbiturate[a]	−2.982	137	Benzamide[b]	−0.953
63	Methaqualone[a]	−2.921	138	7-Isobutyltheophylline[a]	−0.942
64	Praziquantel[b]	−2.893	139	Didanosine[a]	−0.937
65	Chlorzoxazone[a]	−2.831	140	7-Chlorpteridine[b]	−0.876
66	Dichlorprop[a]	−2.827	141	7-Methylpteridine[a]	−0.854
67	Sulfathiazole[a]	−2.805	142	Nicotinic acid[a]	−0.85
68	Diatrizoic acid[a]	−2.788	143	Propranolol[a]	−0.714
69	5,5-Di-i-propylbarbiturate[a]	−2.766	144	2-Chlorpteridine[a]	−0.699
70	Pteridine-7-thiol[a]	−2.706	145	Aminopyrine[a]	−0.619
71	Sulfamethoxazole[a]	−2.705	146	Guaifenesin[a]	−0.598
72	Pteridine-4-thiol[a]	−2.646	147	Ethambutol[b]	−0.565
73	Phenylbutazone[a]	−2.644	148	Methyprylon[a]	−0.382
74	Primidone[a]	−2.64	149	2-Methylpteridine[a]	−0.094
75	Ethyl-4-aminobenzoate (Benzocaine)[a]	−2.616	150	7-Dimethylaminopteridine[a]	−0.021

ChemDraw Ultra version 9.0 (ChemOffice 2005, CambridgeSoft Corporation) software was used for drawing the molecular structures.[28] The optimizations of molecular structures were done by the MOPAC 7.0 (AM1 method) and descriptors were calculated by Molecular Modeling Pro Plus (MMPP) Version 6.0 (ChemSW, Inc.) softwares.[29,30] A stepwise MLR procedure was used for selection of descriptors using the SPSS/PC software package.[31] MLR was performed by using a routine from the Unscrambler version 7.6 package[32] and other calculations were performed in the MATLAB (version 7.0, MathWorks, Inc.) environment.

Molecular Modeling and Theoretical Molecular Descriptors The derivation of theoretical molecular descriptors proceeds from the chemical structure of the compounds. In order to calculate the theoretical descriptors, molecular structures were constructed with the aid of ChemDraw Ultra version 9.0 and molecular structures were optimized using AM1 algorithm.[33,34] The computational chemistry software Chem3D Ultra version 9.0 with MOPAC was used to build the molecules and perform the necessary geometry optimizations. A gradient cutoff of 0.01 was used for all geometry optimizations and the COSMO (COnductor-like Screening MOdel) solvation model was applied for calculations of molecular geometry in water. We have chosen descriptors associated with the neutral molecules of drug in our calculations. As a result, a total of 20 theoretical descriptors were calculated for each compound in the data sets (150 compounds).

The molecular weight, van der Waals volume, surface area, molecular volume,[35] molar volume,[36] density, molecular length, molecular width, molecular depth, octanol–water partition coefficient ($\log P$),[37,38] molar refractivity (MR), $Q \log P$,[39] Hansen's solubility parameters (dispersion, polarity and hydrogen bonding), mean water of hydration,[40] hydrophilic–lipophilic balance (HLB), hydrophilic surface area, % hydrophilic surface area and polar surface area[41] descriptors were calculated by Molecular Modeling Pro Plus (MMPP) Version 6.0 (ChemSW, Inc.) software.

Stepwise Regression for Descriptor Selection The selection of relevant descriptors, which relate the solubility to the molecular structure, is an important step to construct a predictive model. In this work, the stepwise multiple linear regression was used as the feature selection method to select the best calculated descriptors among 20 theoretical descriptors using Molecular Modeling Pro Plus software. All descriptors with zero values or constant and near constant values for all the molecules in the data set were eliminated. The correlation matrix was calculated between the descriptors, one of the two descriptors which has the pair wise correlation coefficient above 0.8 ($r > 0.8$) and it has a large correlation coefficient with the other descriptors was eliminated.

In order to select the subset of descriptors that best explain drug solubility, we have used stepwise regression.[42—44] This method combines the forward and backward procedures. Stepwise model-building techniques for regression designs with a single dependent variable involve identifying an initial model, repeatedly altering the model from the previous step by adding (forward stepwise) or removing (back stepwise) a predictor variable and terminating the search when stepping does not further improve the model. The forward stepwise method employs a combination of the forward entry of independent variables and backward removal of insignificant variables. The best single predictor, which is the most significant variable, was used for the initial linear regression step. Next, descriptors were added one at a time, always adding the one that most improved the fit, until the fit was not significantly improved. Once all the significant variables were determined, the regression equation was constructed. The number of variables retained in the model is based on the levels of significance assumed for inclusion and exclusion of variables from the model.

By using these criteria, 17 out of 20 original descriptors were eliminated and the remaining descriptors were used to generate the models using the SPSS/PC software package. The result shows that three calculated descriptors are the most feasible ones. The selected descriptors are octanol–water partition coefficient ($\log P$), molecular volume (MV) and Hansen's hydrogen bond forming ability (HB).

Multiple Linear Regression Modeling The general purpose of multiple regressions is to quantify the relationship between several independent or predictor variables and a dependent variable. A set of coefficients defines the single linear combination of independent variables (molecular descriptors) that best describes drug solubility. The solubility value for each drug would then be calculated as a composite of each molecular descriptor weighted by the respective coefficients. A multilinear model can be represented as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k + \varepsilon \quad (1)$$

where k is the number of independent variables, $\beta_1, \ldots, \beta_k$ are the regression coefficients and y is the dependent variable. Regression coefficients represent the independent contributions of each calculated molecular descriptor. The algebraic MLR model is defined in Eq. 1 and in matrix notation:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{e} \quad (2)$$

When $\mathbf{X}$ is of full rank the least squares solution is: $\hat{\mathbf{b}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ where $\hat{\mathbf{b}}$ is the estimator for the regression coefficients in $\hat{\mathbf{b}}$.

A single MLR model was developed for drug organic compounds using the Unscrambler version 7.6 software. MLR model was constructed with remaining descriptors based on stepwise feature selection. The MLR model was built using a training set and validation using an external prediction set. Multiple linear regression (MLR) techniques based on least-squares procedures are very often used for estimating the coefficients involved in the model equation.[45,46]

Results and Discussion

All descriptors were calculated for the neutral species. The $\log S_w$ is assumed to be highly dependent upon the octanol–water partition coefficient ($\log P$), molecular volume (MV) and hydrogen bond forming ability (HB). The correlation coefficients between experimental $\log S_w$ and the $\log P$, MV and HB are -0.9229, -0.6215 and 0.5501, respectively. Figure 1 shows the excellent correlation between the experimental $\log S_w$ of the all drug compounds with the $\log P$.

In the present study, the QSPR model was generated using a training set of 110 molecules. The test set of 40 molecules (Table 2) with regularly distributed $\log S_w$ values was used to assess the predictive ability of the QSPR model produced in the regression.

MLR Analysis The software package used for conducting MLR analysis was Unscrambler 7.6. Multiple linear regression (MLR) analysis has been carried out to derive the best QSPR model. The MLR technique was performed on the molecules of the training set shown in Table 1. After regression analysis, a few suitable models were obtained among which the best model was selected and presented in Eq. 3. A small number of molecular descriptors ($\log P$, MV and HB) proposed were used to establish a QSPR model. Additional validation was performed on an external data set consisting of 40 organic compounds. Multiple linear regression analysis provided a useful equation that can be used to predict the $\log S_w$ of drug based upon these parameters. The best equation obtained for the solubility of the drug compounds is:

$$\log S_w = -0.3359 - 1.0563 \log P - 0.0062\mathrm{MV} + 0.0378\mathrm{HB} \quad (3)$$

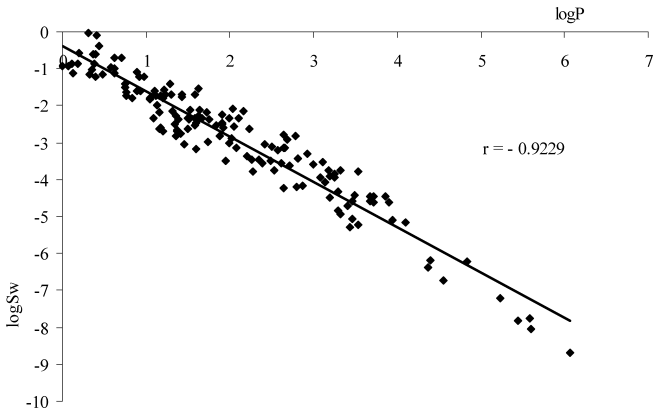

Fig. 1. The Experimental $\log S_w$ Values Drug-Like Compounds Correlate Well with the Octanol–Water Partition Coefficients ($\log P$)

Table 2. Experimental log S_w, Molecular Descriptors, Predicted log S_w, Residuals and Percent Relative Errors Values for External Prediction Set

No.	Name	Exp. (log S_w)	Log P	MV	HB	Pred. (log S_w)	Residuals	%RE
1	Fenbuconazole	−6.226	4.8400	180.7364	7.6632	−6.2717	0.0457	0.7338
2	Diclofenac	−5.097	3.9400	143.4893	8.9003	−5.0448	−0.0522	−1.0248
3	Fenoxycarb	−4.719	3.4028	166.1030	7.7500	−4.6602	−0.0588	−1.2460
4	Acenaphthene	−4.590	3.4660	124.0000	2.5700	−4.6636	0.0736	1.6030
5	Diphenyl	−4.340	3.3000	126.0000	3.0700	−4.4816	0.1416	3.2636
6	1,4-Dibromobenzene	−4.070	3.1400	126.0000	7.3100	−4.1522	0.0822	2.0194
7	Oxazepam	−3.952	3.0832	141.6930	12.3718	−3.9973	0.0453	1.1471
8	Naphthalene	−3.610	2.7100	111.0000	5.2600	−3.6832	0.0732	2.0267
9	Bumetanide	−3.562	2.6112	164.4000	12.0828	−3.6496	0.0876	2.4598
10	5-*t*-Butyl-5-(3-methylbut-2-enyl)barbiturate	−3.551	2.3944	145.4307	7.5771	−3.4743	−0.0767	−2.1611
11	Melphalan	−3.485	2.4829	156.9261	10.6779	−3.5212	0.0362	1.0393
12	Fluometuron	−3.463	2.3500	114.3074	7.6518	−3.2328	−0.2302	−6.6481
13	Butylparaben	−3.101	2.5020	110.8953	12.8699	−3.1748	0.0738	2.3813
14	Heptabarbital	−3.000	2.0000	138.6097	7.9317	−3.0022	0.0022	0.0742
15	Praziquantel	−2.893	2.0224	142.0000	7.4934	−3.0634	0.1704	5.8888
16	Propylparaben	−2.557	2.0460	100.9936	12.6500	−2.6405	0.0835	3.2659
17	5,5-Dipropylbarbiturate	−2.527	1.5834	118.5753	8.4545	−2.4190	−0.1080	−4.2749
18	Isocarboxazid	−2.461	1.5946	123.4382	9.9000	−2.4061	−0.0549	−2.2322
19	Phenacetin	−2.371	1.7570	103.2280	7.2320	−2.5540	0.1830	7.7202
20	Phenobarbital	−2.366	1.5196	121.1266	8.7466	−2.3563	−0.0097	−0.4121
21	Pteridine-4-methyl-thiol	−2.365	1.6325	88.3642	10.7589	−2.1975	−0.1675	−7.0824
22	Secbutabarbital	−2.333	1.4961	118.6515	8.4605	−2.3270	−0.0060	−0.2568
23	Cyclobarbital	−2.273	1.3589	128.6793	8.2893	−2.2503	−0.0227	−0.9966
24	Probarbital	−2.153	1.3100	108.6790	9.4100	−2.0331	−0.1199	−5.5706
25	Allobarbital	−1.796	1.0444	112.3838	10.8000	−1.7227	−0.0733	−4.0785
26	Aspirin	−1.733	1.1769	92.6674	11.6759	−1.7081	−0.0249	−1.4379
27	Methyl-*p*-hydroxybenzoate	−1.705	1.3040	81.0940	14.0800	−1.6801	−0.0249	−1.4633
28	1-Butyltheobromine	−1.625	0.7700	129.8055	11.1612	−1.5266	−0.0984	−6.0565
29	5-Ethyl-5-allylbarbiturate	−1.614	0.8966	105.5769	9.1275	−1.5880	−0.0260	−1.6115
30	Cimetidine	−1.613	0.9319	124.0000	9.2716	−1.7333	0.1203	7.4598
31	5-Ethyl-5-propylbarbiturate	−1.491	0.7488	98.5588	9.4045	−1.3782	−0.1128	−7.5685
32	5,5-Diethylbarbiturate	−1.410	0.7488	98.6599	9.4045	−1.3788	−0.0312	−2.2144
33	1-Propyltheobromine	−1.207	0.3900	119.8542	11.7121	−1.0430	−0.1640	−13.5833
34	5-Methyl-5-allylbarbiturate	−1.160	0.4793	95.4815	10.6000	−1.0293	−0.1307	−11.2678
35	6-Chlorpteridine	−1.124	0.6200	84.0000	11.8881	−1.0584	−0.0656	−5.8341
36	4-Dimethylaminopteridine	−1.021	0.3530	94.8208	11.8202	−0.94564	−0.0754	−7.3807
37	Methocarbamol	−0.985	0.6288	115.9000	15.3616	−1.0328	0.0478	4.8543
38	Benzamide	−0.953	0.5810	67.3657	11.2000	−0.9408	−0.0122	−1.2817
39	7-Chlorpteridine	−0.876	0.3830	76.0465	11.8881	−0.7591	−0.1169	−13.3458
40	Ethambutol	−0.565	0.2000	119.3000	14.6500	−0.6278	0.0628	11.1097

Positive values in the regression coefficients indicate that the indicated descriptor contributes positively to the value of log S_w, whereas negative values indicate that the greater the value of the descriptor the lower the value of log S_w. In other words, increasing the log P and MV will decrease log S_w and increasing the HB increases extent of log S_w of the drug organic compounds.

For evaluation of the predictive power of the generated MLR, the optimized model was applied for prediction of log S_w values of 40 compounds in the prediction set which were not used in the optimization procedure. For the constructed model, the predictive ability of the MLR model was evaluated by calculation of statistical parameters. The predicted values of log S_w, residuals and the percent relative errors (%RE) of prediction obtained by the MLR method are presented in Table 2. The plots of predicted log S_w *versus* experimental log S_w and the residuals (experimental log S_w − predicted log S_w) *versus* experimental log S_w value, obtained by the MLR modeling, and the random distribution of residuals about zero mean are shown in Fig. 2. The stability and validity of model was tested by prediction of the response values for the prediction set. This model is applicable for prediction of log S_w from −8.690 to −0.021. The average relative errors (RE%) of prediction and squares of correlation coefficients (R^2) are −1.2996% and 0.9954 for MLR model, respectively.

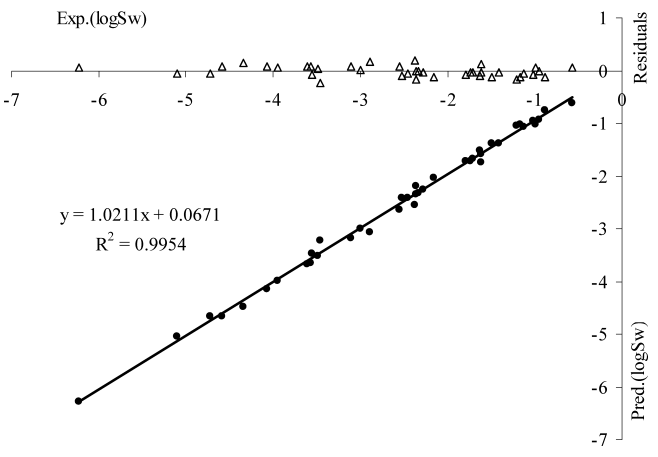

Fig. 2. Plots of Predicted log S_w and Residuals log S_w Estimated by MLR Modeling *versus* Experimental log S_w for Test Molecules in Prediction Set

Interpretation of Descriptors The QSPR developed indicated that octanol–water partition constant ($\log P$), molecular volume and hydrogen bond forming ability significantly influence drug aqueous solubility.

The *n*-octanol–water partition coefficient, respectively its logarithmic value is called $\log P$. The $\log P$ frequently used to estimate the membrane permeability and the bioavailability of compounds, since an orally administered drug must be enough lipophilic to cross the lipid bilayer of the membranes, and on the other hand, must be sufficiently water soluble to be transported in the blood and the lymph. The $\log P$ is frequently used in quantitative structure–property relationships as a measure of the lipophilic character of the molecules. Octanol–water partition coefficient ($\log P$) is used in QSPR studies and rational drug design as a measure of molecular hydrophobicity. Hydrophobicity affects drug absorption, bioavailability, hydrophobic drug–receptor interactions, metabolism of molecules, as well as their toxicity. Lipophilicity is approximately correlated to passive transport across cell membranes and the ability of a compound to partition through a membrane since membranes are composed largely of lipids. Log *P* is well established as a key parameter to describe lipophilicity, uptake and distribution in biological systems. With increases octanol/water partition coefficients, water solubility decreases.

Molecular volume determines transport characteristics of molecules, such as intestinal absorption or blood-brain barrier penetration. Volume is therefore often used in QSPR studies to model molecular properties and biological activity. The steric effects characterise bulk properties of a molecule and can be described with molecular volume. The molecular volume is clearly the most important descriptor for aqueous solubility. In order for a solute to enter into aqueous solution, a cavity must be formed in the solvent for the solute molecule to occupy. Water as a solvent would much prefer to interact with itself or other hydrogen bonding or ionic species than with a nonpolar solute, so there is an increasing penalty (and thus lower solubility) for larger solutes. By increasing molecular volume leads to increasing cavity formation energy in water, the larger the solute, the greater the energy demand to make cavity and the lower the solubility.

A particularly strong type of polar interaction occurs in molecules where a hydrogen atom is attached to an extremely electron-hungry atom such as oxygen, nitrogen, or fluorine. In such cases, the hydrogen's sole electron is drawn toward the electronegative atom, leaving the strongly charged hydrogen nucleus exposed. In this state the exposed positive nucleus can exert a considerable attraction on electrons in other molecules, forming a protonic bridge that is substantially stronger than most other types of dipole interactions. This type of polarity is so strong compared to other van der Waals interactions, that it is given its own name: hydrogen bonding. Understandably, hydrogen bonding plays a significant role in solubility behavior. Hydrogen bonding not a true bond, but a very strong form of dipole–dipole attraction. The O–H and N–H bonds in molecular structures are strongly polarized and the positive charge is located on $H^{\delta+}$. In this study we have a dipolar protic solvent (water) containing hydrogen bond donor (O–H bonds) and hydrogen bond acceptor (lone pairs of oxygen atom). Hydrogen bond donor solutes are simply those containing a hydrogen atom bound to an electronegative atom. Hydrogen bond acceptors solutes are that have a lone pair available for donation, and include N and O atoms in their structures. The hydrogen bonding (Hansen) a measure of the tendency of a molecule to form hydrogen bonds. As the hydrogen bond formation increases, water solubility increases, this is agreed to the fact that water has large dipolarity/polarizability. As polarity increases, water solubility increases. The intermolecular hydrogen bonding can dramatically influence solubility properties.

Statistical Parameters For evaluation of the predictive power of the generated MLR, the optimized models was applied for prediction of $\log S_w$ values of test compounds in the prediction set, which were not used in the optimization procedure. For the constructed models, five general statistical parameters were selected to evaluate the prediction ability of the model for $\log S_w$. For this case, the predicted $\log S_w$'s of each sample in prediction step were compared with the experimental $\log S_w$. PRESS (predicted residual sum of squares) appears to be the most important parameter accounting for a good estimate of the real predictive error of the models. Its small value indicates that the model predicts better than chance and can be considered statistically significant.

$$\mathrm{PRESS} = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 \tag{4}$$

Root mean square error of prediction (RMSEP) is a measurement of the average difference between predicted and experimental values, at the prediction stage. RMSEP can be interpreted as the average prediction error, expressed in the same units as the original response values. The RMSEP was obtained by the following formula:

$$\mathrm{RMSEP} = \left[\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \right]^{0.5} \tag{5}$$

The third statistical parameter was relative error of prediction (REP) that shows the predictive ability of each component, and is calculated as:

$$\mathrm{REP}\ (\%) = \frac{100}{\bar{y}} \left[\frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 \right]^{0.5} \tag{6}$$

The predictive applicability of a regression model is described in various ways. The most general expression is the standard error of prediction (SEP) which is given in the following formula:

$$\mathrm{SEP} = \left[\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n-1} \right]^{0.5} \tag{7}$$

The square of the correlation coefficient (R^2), which is, indicated the quality of fit of all the data to a straight line is calculated for the checking of test set, and is calculated as:

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{8}$$

where y_i is the experimental $\log S_w$ of the drug in the sample i, $\hat{y}_i$ represented the predicted $\log S_w$ of the drug in the sample i, $\bar{y}$, is the mean of experimental $\log S_w$ in the prediction set and n is the total number of samples used in the prediction set.

The statistical parameters values of PRESS, RMSEP, REP (%), SEP and R^2 of prediction set for the MLR model were equal to 0.3677, 0.0959, −3.7619, 0.0971 and 0.9954 respectively.

Conclusions

Predictive QSPR model which is based on molecular descriptors is proposed in this study to correlate the aqueous solubility of drug compounds. Application of the developed model to a testing set of 40 compounds demonstrates that the new model is reliable with good predictive accuracy and simple formulation. Since the QSPR was developed on the basis of theoretical molecular descriptors calculated exclusively from molecular structure, the proposed model could potentially provide useful information about the solubility of drug compounds.

We have developed here a useful QSPR equation derived from theoretical descriptors associated with solubility property. A MLR is successfully presented for prediction aqueous solubility property ($\log S_w$) of various drug compounds with diverse chemical structures using a linear quantitative structure–property relationship. A model with high statistical quality and low prediction errors was obtained. The model could predict the solubility property of the drug compounds accurately. The macroscopic (bulk) activities/properties of chemical compounds clearly depend on their microscopic (structural) characteristics. Development of quantitative structure property/activity relationships (QSPR/QSAR) on theoretical descriptors is a powerful tool not only for prediction of the chemical, physical and biological properties/activities of compounds, but also for deeper understanding of the detailed mechanisms of interactions in complex systems that predetermine these properties/activities. MLR analysis provided useful equation that can be used to predict the $\log S_w$ of chemicals based upon $\log P$, MV and HB parameters. The results indicate that a strong correlation exists between the $\log S_w$ and $\log P$ for drug compounds. This procedure allowed us to achieve a precise and relatively fast method for determination of $\log S_w$ of different series of drug compounds and to predict with sufficient accuracy the $\log S_w$ of new drug derivatives.

References and Notes

1) Morelock M. M., Choi L. L., Bell G. L., Wright J., *J. Pharm. Sci.*, **83**, 948—952 (1994).
2) Forbes R. T., York P., Davidson J. R., *Int. J. Pharm.*, **126**, 199—208 (1995).
3) O'Connor K. M., Corrigan O. I., *Int. J. Pharm.*, **226**, 163—179 (2001).
4) Huuskonen J., *Comb. Chem. High Throughput Screen*, **4**, 311—316 (2001).
5) Catana C., *J. Chem. Inf. Model.*, **45**, 170—176 (2005).
6) Huuskonen J., Salo M., Taskinen J., *J. Pharm. Sci.*, **86**, 450—454 (1997).
7) Huuskonen J., Salo M., Taskinen J., *J. Chem. Inf. Comput. Sci.*, **38**, 450—456 (1998).
8) William L. J., Erin M. D., *Bioorg. Med. Chem. Lett.*, **10**, 1155—1158 (2000).
9) William L. J., Erin M. D., *Adv. Drug Del. Rev.*, **54**, 355—366 (2002).
10) Butina D., Gola J. M. R., *J. Chem. Inf. Model.*, **43**, 837—841 (2003).
11) Eros D., Keria G., Kovesdi I., Szantai C. K., Meszaros G., Orfi L., *Mini-Rev. Med. Chem.*, **4**, 167—177 (2004).
12) Delaney J. S., *Drug Discov. Today*, **10**, 289—295 (2005).
13) Huuskonen J., *J. Chem. Inf. Comput. Sci.*, **40**, 773—777 (2000).
14) Huuskonen J., Rantanen J., Livingstone D., *Eur. J. Med. Chem.*, **35**, 1081—1088 (2000).
15) Huuskonen J., *Environ. Tox. Chem.*, **20**, 491—497 (2001).
16) Tetko I. V., Tanchuk V. Y., Kasheva T. N., Villa A. E., *J. Chem. Inf. Comput. Sci.*, **41**, 1488—1493 (2001).
17) Ran Y., Jain N., Yalkowsky S. H., *J. Chem. Inf. Comput. Sci.*, **41**, 1208—1217 (2001).
18) McElroy N. R., Jurs P. C., *J. Chem. Inf. Comput. Sci.*, **41**, 1237—1247 (2001).
19) Yaffe D., Cohen Y., Espinosa G., Arenas A., Giralt F., *J. Chem. Inf. Comput. Sci.*, **41**, 1177—1207 (2001).
20) Ghasemi J., Shahmirani S., Farahani E. V., *Annali di Chimica*, **96**, 327—337 (2007).
21) Ghasemi J., Saaidpour S., Brown S. D., *J. Mol. Struct.* (*Theochem*), **805**, 27—32 (2006).
22) Ghasemi J., Ahmadi Sh., *Annali di Chimica*, **97**, 69—83 (2007).
23) Chunsheng Y., Xinhui L., Weimin G., Teng L., Xiaodong W., Liansheng W., *Water Res.*, **36**, 2975—2982 (2002).
24) Schaper K. J., Kunz B., Raevsky O. A., *QSAR Comb. Sci.*, **22**, 943—958 (2003).
25) Hou T. J., Xia K., Zhang W., Xu X. J., *J. Chem. Inf. Comput. Sci.*, **44**, 266—275 (2004).
26) Raevsky O. A., Raevskaja O. E., Schaper K. J., *QSAR Comb. Sci.*, **23**, 327—343 (2004).
27) Rytting E., Lentz K. A., Chen X. Q., Qian F., Venkatesh S., *AAPS J.*, **7**, 78—105 (2005).
28) ChemOffice 2005, CambridgeSoft Corporation, Web: http://www.cambridgesoft.com.
29) Web: http://www.psu.ru/science/soft/winmopac/
30) Web: http://www.chemsw.com/
31) Web: http://www.spss.com/
32) The Unscrambler version 7.6, 2000, Web: http://www.camo.com.
33) Dewar M. J. S., Zoebisch E. G., Healy E. F., Stewart J. J. P., *J. Am. Chem. Soc.*, **107**, 3902—3909 (1985).
34) Young D. C., "Computational Chemistry," John Wiley & Sons, Inc., New York, 2001.
35) Potts R. O., Guy R. H., *Pharm. Res.*, **10**, 635—637 (1993).
36) Genty M., González G., Clere C., Desangle-Gouty V., Legendre J. Y. D., *Eur. J. Pharm. Sci.*, **12**, 223—229 (2001).
37) Hansch C., Leo A., "Substituent Constants for Correlation Analysis in Chemistry and Biology," Wiley, New York, 1979.
38) Rekker R. F., De Kort H. M., *Eur. J. Med. Chem.*, **6**, 565—616 (1979).
39) Bodor N., Buchwald P., *Curr. Med. Chem.*, **5**, 353—380 (1998).
40) Meng Z., Carper W. R., *J. Mol. Struct.* (*Theochem*), **531**, 89—98 (2000).
41) Ertl P., Rohde B., Selzer P., *J. Med. Chem.*, **43**, 3714—3717 (2000).
42) Darlington R. B., "Regression and Linear Models," McGraw-Hill Higher Edu., New York, 1990.
43) Massart D. L., Vandeginste B. G. M., Buydens L. M. C., De Jong S., Lewi P. J., Smeyers-Verbeke J., "Handbook of Chemometrics and Qualimetrics," Part A, Elseiver, Amsterdam, 1997.
44) Xu L., Zhang W. J., *Anal. Chim. Acta*, **446**, 477—483 (2001).
45) Martens H., Naes T., "Multivariate Calibration," Wiley, Chichester, 1989.
46) Darlington R. B., "Regression and Linear Models," McGraw-Hill, New York, 1990.