Korean J. Chem. Eng., **25**(4), 865-873 (2008)

Aqueous solubility of poorly water-soluble drugs: Prediction using similarity and quantitative structure-property relationship models

Junhyoung Kim*, Dong Hyun Jung*, Hokyoung Rhee*, Seung-Hoon Choi*, Min Jae Sung, and Woo Sik Choi***,***†**

*Insilicotech Co. Ltd., A-1101 Kolontripolis, 210, Geumgok-dong, Bundang-gu, Seongnam-shi 463-943, Korea
**Interdisciplinary Program in Powder Technology, Graduate School,
Pusan National University, 30, Jangjeon-dong, Geumjeong-gu, Busan 609-735, Korea
***Department of Pharmaceutical Manufacturing, Pusan National University,
30, Jangjeon-dong, Geumjeong-gu, Busan 609-735, Korea
(*Received 13 June 2007 • accepted 14 January 2008*)

Abstract−The aqueous solubility of poorly water-soluble drugs is an important property of many factors affecting their bioavailability such as the solubility and rate of dissolution in water. The quantitative structure-property relationship approach using genetic algorithm was applied to make models for predicting some poorly water-soluble drugs such as ursodeoxycholic acid, diphenyl hydrantoin and biphenyl dimethyl dicarboxylate. The experimental solubility data of 3518 chemical structures were collected from the web and used to build a model. Three data sets of 50 compounds were extracted according to their structural similarity with each drug. A fast and predictive similarity based approach was developed and validated with conventional method. This can be used to predict the aqueous solubility for drugs by using a small set of compounds, especially for poorly water-soluble compounds. Moreover, the estimation values of various sets were further compared with fine grinding experiment data.

Key words: Insoluble Drug, Fine Grinding, Solubility, QSPR, UDCA, Phenytoin, DDB

INTRODUCTION

The solubility of a compound is defined as the amount of solute dissolved in a saturated solution under equilibrium conditions. Aqueous solubility is a particularly important property of compounds, especially in the field of pharmaceutical chemistry, biological chemistry, materials, and environmental science. For example, if a drug is to be orally administered, it should have an adequate aqueous solubility. Moreover, the solubility is often determined at all stages of drug discovery. So, a series of attempts to enhance the bioavailability of poorly water-soluble drugs have been made by the fine grinding technique using a planetary ball mill [1,2], and in addition a reliable and cheap prediction method is inevitable for development of new drugs being able to be orally administered.

Appropriate physicochemical properties, e.g., logP and solubility together with pharmacokinetic properties and toxicities, are the major determinants for progressing from a good lead to a good drug. However, experimental assays and animal or clinical tests are expensive and not practical to apply to the large collection of compounds in the early stage. It is also difficult and time consuming to measure the properties accurately, especially for poorly water-soluble compounds. There exists a strong need for fast, reliable, and generally applicable methods for the prediction of aqueous solubility [3]. The quantitative structure-property relationships, QSPR, is a useful tool for prediction of endpoints of interest on compounds that have not been experimentally investigated. In this work, for the collection of aqueous solubility data reported in literature, we applied a typical and similarity based QSPR method for generation of predictive models of aqueous solubility and they are used for predicting the solubility of some insoluble drugs.

†To whom correspondence should be addressed.
E-mail: wschoi@pusan.ac.kr

GENERAL PROCEDURES

The classical QSPR modeling was applied for the prediction of intrinsic aqueous solubility. As described in many studies [4-7], the general procedures for QSPR modeling were involved as follows: the data collection of aqueous solubility and structure information, the classification into training and test set, the building of 3D structure, the calculation of molecular descriptors, the generation and validation of prediction models.

1. Data Collection

The experimental solubility values of 4991 compounds were obtained from ChemIDplus [8] and then identified by CAS registry number. However, the solubility data without the molecular structure information was ruled out from the database. If a compound had duplicates, we took the average value as a representative. These data were also filtered with following conditions and removed from the list: (1) the compounds including the non-organic elements such as Pb, Sn, Hg, and so on.; (2) the solubility of the compounds not having been measured in the appropriate range of the temperature ($15\,^{\circ}C<T<35\,^{\circ}C$); (3) all salt structures; and (4) the compounds which cannot calculate the molecular descriptors due to undefined atom types or molecular sizes.

Finally, a database was prepared with total 3518 compounds and 5 data sets including 1 set of total compounds and 4 drug-like sets were constructed. In this study, the solubility is expressed as logSw, where Sw is the solubility in mg/L.

2. Data Sets for QSPR Models

The ursodeoxycholic acid (UDCA), diphenyl hydrantoin (Pheny-

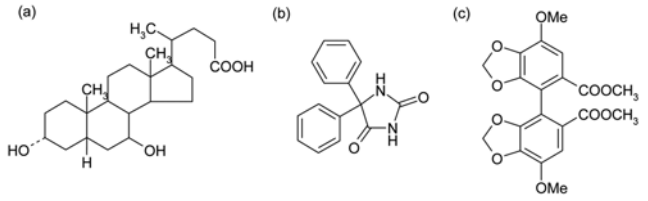

Fig. 1. The chemical structure of (a) UDCA, (b) phenytoin, and (c) DDB.

toin) and biphenyl dimethyl dicarboxylate (DDB) were selected from insoluble drugs and they show low molecular similarity with each other; these compounds have a molecular similarity less than 0.2 for each other. Fig. 1 shows the chemical structure of insoluble drugs, UDCA, phenytoin, and DDB compound.

The 5 sets for QSPR model were prepared by the following approaches: (1) for all sets, the data were separated into 60% the training set and 40% test set randomly. The total compounds data set was classified into 2105 training and 1413 test sets; (2) a UDCA set was created by selection for top 49 most similar molecules in the total set. In order to select the similar molecules, the molecular similarity was calculated with the Tanimoto similarity index [5] by the functional class extended-connectivity fingerprint. Then, the selected molecules were divided into 30 compounds for training and 19 compounds for test. The UDCA compound was added to the test set for model validation; (3) both phenytoin and DDB sets were also created by the same manner as the previous generation of the UDCA set. The maximum, minimum, and average similarity values for UDCA, Phenytoin and DDB sets are listed up in Table 1; (4) finally the UDCA, phenytoin, and DDB sets are merged into a UDCA-Phenytoin-DDB set. The maximum, minimum, and average values for intrinsic aqueous solubility in relation with these 5 data sets are listed up in Table 2.

3. Preparation of Molecular Structure and Calculation of Descriptors

The chemical structures for the most compounds were retrieved from the PubChem [9] with CAS registry number and then exported to Cerius2, software for molecular modeling [10]. All computations using the Cerius2 were performed on a Silicon Graphics Octane IRIX workstation. The energy minimized conformations of 3D structures for all structures were generated by using SMART minimization tool, and the charge of molecular atoms was assigned with the charge-equilibration method, then molecular descriptors were calculated for each compound.

All the descriptors available within Cerius2 were computed as follows: descriptors describing the E-state keys [11], electronic descriptors, topological descriptors, molecular connectivity indices, molecular valence connectivity indices, subgroup count indices, shape indices, information content descriptor, spatial descriptor, shadow indices, structural descriptor, and thermodynamic descriptor.

4. QSPR Models

The Cerius2, version 4.10 package was used to build the QSPR models for the prediction of aqueous solubility. For model construction, a small subset of descriptors that accurately represent the relationship between chemical structures and aqueous solubility is required to be identified. In this study, the descriptor selection was performed by the genetic algorithm routine provided by genetic function approximation of the Cerius2. The genetic algorithm is an iterative improvement optimization technique, in which an analogy with evolution in solving difficult combinatorial problems is used. The genetic algorithm (GA) using R^2 optimization was qualified by the reciprocal of the Friedman's lack-of-fitness function [12]. The 200 population size, 30000 generation, and the multiple linear equations were used for every set.

Regressions were carried out by using an increasing number of descriptors, and both R^2 (coefficient of determination) and Q^2 (leave-one-out cross-validated R^2) were monitored. The leave-one-out validation was performed by holding one compound out and by developing a model based on the rest of the training set. This process was carried out for all the data sets.

RESULTS AND DISCUSSION

1. Model for Total Set

The training model for total set has a coefficient of determination, R^2=0.743, Q^2=0.740, F-test=1009.928, and PRESS=2650.764, where PRESS stands for the squared sum of predicted residuals. The test set of total set has R^2=0.735. The Fig. 2 shows the plot of predicted vs. experimental solubility for the training and test set compounds. The results of all QSPR models are shown in Table 3 with various statistical parameters

Table 1. The spectrums of molecular similarity for each UDCA, Phenytoin, and DDB sets

Class	Model	Similarity max.	Similarity min.	Similarity avg.
I	UDCA set	0.926	0.357	0.449
II	Phenytoin set	0.609	0.346	0.388
III	DDB set	0.500	0.278	0.343

Table 2. The spectra of intrinsic aqueous solubility for each 5 data sets

Class	Model	n[a]	Sw[b] max.	Sw[b] min.	Sw[b] avg.
I	UDCA Set	50	1.300E+06	2.760E-02	1.115E+05
II	Phenytoin Set	50	6.660E+06	3.000E-01	1.546E+05
III	DDB Set	50	4.760E+04	8.200E-02	3.156E+03
IV	UDCA, Phenytoin, DDB Set	150	6.660E+06	2.700E-02	8.975E+04
V	Total Set	3518	6.660E+06	7.430E-06	8.500E+04

[a]The number of compounds.
[b]Aqueous solubility (μg/ml).

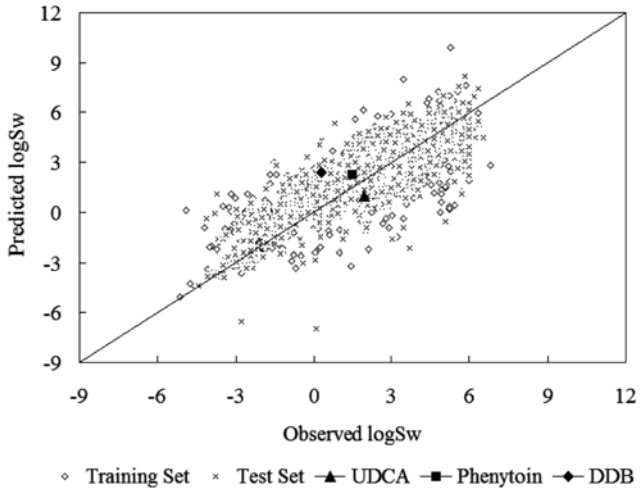

Fig. 2. The predicted vs. experimental solubility for the training set of 2105 compounds and test set of 1413 compounds.

The regression equation is as follows:

$$\begin{aligned}\text{logSw}=&4.84243+2.84427*[\text{Jurs-RNCG}]-0.673154*[\text{AlogP98}]\\&-0.267192*[\text{S\_aaaC}]-0.003611*[\text{Jurs-TPSA}]\\&+0.171372*[\text{Rotlbonds}]-0.119782*[\text{Shadow-Xlength}]\end{aligned}\quad(1)$$

where [x] represents the name of descriptor. There are six significant descriptors and the coefficient of descriptor illustrates the relative importance of each descriptors.

The Jurs-RNCG descriptor [13,14] expressing the charge of most negative atom divided by the total negative charge is a relative negative charge. The Jurs-TPSA is a total hydrophobic surface area which is the sum of solvent-accessible surface areas of atoms with absolute value of partial charges less than 0.2. The AlogP98 descriptor is the significant contributor to the solubility prediction. It is not surprising that this term carries a negative sign since solubility and water-octanol partition coefficient represent two opposite properties of a compound: hydrophilicity and hydrophobicity. The higher value of the AlogP98 means that the compound is more hydrophobic and has lower solubility. E-state key, S_aaaC, is related to the aromatic nature of a compound. The number of rotatable bonds, Rotlbonds, is a parameter closely related to the size of a compound. The Shadow-Xlength descriptor [15] expresses the length of molecule in the x dimension and is also related to the size of compound.

2. Model for UDCA Set

For the development of *in silico* predictive models for UDCA compounds set, 50 compounds including UDCA itself were selected among the total data of 3518 compounds by Tanimoto similarity score, and they were sub-divided randomly into two classes, training 30 compounds and test 20 compounds. A test set of 20 compounds was provided for a validation test of the UDCA-like intrinsic aqueous solubility QSPR model.

The training model for the UDCA set has a coefficient of determination, R^2=0.969, Q^2=0.950, F-test=195.035, PRESS=4.823, and the remaining test set has R^2=0.927. Fig. 3(a) shows the predicted versus experimental solubility for training and test compounds of UDCA set.

The regression equation for UDCA-like model is as follows:

$$\begin{aligned}\text{LogSw}=&7.58122-0.277775*[\text{IAC-Total}]+0.038091*[\text{MW}]\\&-1.09645*[\text{AlogP98}]+0.151451*[\text{Jurs-WPSA-3}]\end{aligned}\quad(2)$$

There are four significant descriptors. The IAC-Total descriptor is the information of atomic composition index. The atoms in the molecule are partitioned into equivalence classes corresponding to their atomic numbers. The partition then yields the descriptor IAC-mean as the mean quantity of information. The descriptor IAC-Total is defined as N x IAC-mean, where N is the number of atoms in the molecule. The descriptor MW is the molecular weight and the descriptor Jurs-WPSA-3 is the surface-weighted positive charged partial surface areas. The observed and calculated aqueous solubility and Tanimoto similarity score for compounds in the UDCA set are presented in Table 4.

The molecular structures of UDCA set consist of hydrophobic and hydrophilic moiety which are composed of linear chain, cyclic ring chain, carboxylic acid, and hydroxyl group. The above descriptors are chosen based on the molecular size and the lipophilicity property. It appears that the solubility was determined by contribution of these properties.

3. Model for Phenytoin Set

The training model for the phenytoin set has a coefficient of determination, R^2=0.944, Q^2=0.920, F-test=105.726, PRESS=5.047, and the remaining test set has R^2=0.865. Fig. 3(b) shows the comparison between the predicted and the experimental solubility for Phenytoin set.

The regression equation for Phenytoin-like model is as follows:

$$\begin{aligned}\text{logSw}=&14.2052+0.293378*[\text{Dipole-mag}]-0.089897*[\text{MolRef}]\\&-8.31318*[\text{Density}]-0.008465*[\text{Jurs-WNSA-2}]\end{aligned}\quad(3)$$

The Dipole-mag descriptor is the magnitude of dipole moment and a 3D electronic descriptor that indicates the strength and orien-

Table 3. Statistical parameters for the all models

Class	Model	Training Set					Test Set	
		n^a	R^2	F-test	PRESS	Q^2	n^a	R^2
I	Total set	2105	0.743	1009.928	2650.764	0.740	1413	0.735
II	UDCA set	30	0.969	195.035	4.823	0.950	20	0.927
III	Phenytoin set	30	0.944	105.726	5.047	0.920	20	0.865
IV	DDB set	30	0.925	107.538	5.004	0.902	20	0.903
V	UDCA, phenytoin, DDB set	90	0.775	47.596	57.726	0.732	60	0.646

[a]The number of compounds.

Table 4. Observed and calculated aqueous solubility for compounds in the UDCA set. All solubility data are logSw, where Sw is aqueous solubility in the unit of μg/ml

No.	IUPAC NAME	Observed	Predicted	Residuals	Similarities
Train 1	Cholic acid	2.24	1.41	0.83	0.93
Train 2	Lithocholic acid	−0.42	−0.61	0.19	0.88
Train 3	Glycoursodeoxycholic acid	0.13	0.56	−0.43	0.73
Train 4	Glycochenodeoxycholic acid	0.50	0.67	−0.17	0.73
Train 5	Epiandrosterone	1.21	1.17	0.04	0.60
Train 6	Leucine	4.22	4.36	−0.14	0.46
Train 7	3-Methyladipic acid	5.39	5.38	0.01	0.44
Train 8	Cholesterol	−1.02	−1.48	0.45	0.43
Train 9	D-Isoleucine	4.51	4.26	0.25	0.41
Train 10	Allo-DL-isoleucine	4.73	4.25	0.48	0.41
Train 11	Norleucine	4.08	4.27	−0.19	0.41
Train 12	Prasterone	1.80	1.79	0.01	0.41
Train 13	d-Camphoric acid	3.74	3.78	−0.04	0.40
Train 14	Gamma-aminobutyric acid	6.11	6.27	−0.15	0.39
Train 15	Cyclobutaneacetic acid, 3-acetyl-2,2-dimethyl-	4.69	4.57	0.12	0.39
Train 16	4-Methylcyclohexanol	4.18	4.17	0.00	0.38
Train 17	Aminocaproic acid	5.70	4.92	0.79	0.38
Train 18	Testosterone	1.37	1.74	−0.37	0.38
Train 19	Acetylleucine	3.91	4.16	−0.25	0.38
Train 20	Lauric acid	0.68	1.12	−0.44	0.37
Train 21	Caproic acid	4.01	4.16	−0.15	0.37
Train 22	2-Methylheptan-4-ol	3.21	3.09	0.12	0.37
Train 23	Octanoic acid	2.90	3.02	−0.13	0.37
Train 24	6-Methylheptan-3-ol	3.19	3.11	0.08	0.37
Train 25	Undecanoic acid	1.72	1.59	0.13	0.37
Train 26	Decanoic acid	1.79	2.06	−0.27	0.37
Train 27	Heptanoic acid	3.45	3.58	−0.13	0.37
Train 28	Pelargonic acid	2.45	2.54	−0.09	0.37
Train 29	Corticosterone	2.30	2.82	−0.52	0.37
Train 30	Menthol	2.66	2.70	−0.04	0.36
Test 1	UDCA	2.20	0.39	1.81	1.00
Test 2	Hyocholic acid	1.26	1.17	0.09	0.85
Test 3	Glycocholic acid	0.52	1.78	−1.26	0.69
Test 4	Glycodeoxycholic acid	0.43	0.51	−0.08	0.61
Test 5	Cyclopentanepropanoic acid	3.38	3.73	−0.35	0.46
Test 6	3-Aminobutyric acid	6.00	6.13	−0.13	0.44
Test 7	Bicyclo(2.2.1)heptan-2-ol, 1,7,7-trimethyl-, (1R-endo)- (9CI)	2.87	3.68	−0.81	0.43
Test 8	Cyclohexaneacetic acid	3.46	3.62	−0.16	0.41
Test 9	DL-Valine	4.82	5.04	−0.22	0.39
Test 10	Butyric acid	4.78	5.48	−0.71	0.38
Test 11	11-Hydroxyundecanoic acid	2.60	2.84	−0.24	0.38
Test 12	11-Aminoundecanoic acid	3.00	2.24	0.76	0.38
Test 13	Acetoacetic acid	6.00	7.20	−1.20	0.37
Test 14	Margaric acid	−1.57	−0.47	−1.10	0.37
Test 15	Isovaleric acid	4.61	4.94	−0.33	0.37
Test 16	Myristic acid	0.03	0.36	−0.33	0.37
Test 17	Pentadecanoic acid	−0.49	0.04	−0.53	0.37
Test 18	Palmitic acid	−1.40	−0.22	−1.18	0.37
Test 19	n-Pentanoic acid	4.38	4.79	−0.41	0.37
Test 20	Propionic acid	6.00	6.24	−0.24	0.36

Table 5. Observed and calculated aqueous solubility for compounds in the phenytoin set. All solubility data are logSw, where Sw is aqueous solubility in the unit of μg/ml

No.	IUPAC NAME	Observed	Predicted	Residuals	Similarities
Train 1	Benzilic acid	3.15	2.94	0.21	0.46
Train 2	Phenylmethylbarbituric acid	2.88	2.85	0.03	0.61
Train 3	Ethylphenylhydantoin	3.14	2.97	0.17	0.54
Train 4	Phthalimide	2.56	2.83	−0.27	0.45
Train 5	2-Bromo-3,3-dimethyl-N-N-(alpha-alpha-dimethylbenzyl) butyramide	0.55	0.25	0.30	0.41
Train 6	Benzamide	4.13	3.92	0.21	0.41
Train 7	Saccharin	3.60	3.49	0.11	0.40
Train 8	5-Isopropylbarbituric acid	3.78	3.67	0.10	0.39
Train 9	Norbormide	1.78	1.90	−0.12	0.38
Train 10	Anthraquinone	0.13	0.17	−0.04	0.38
Train 11	Benzaldehyde	3.84	4.18	−0.34	0.38
Train 12	Glutethimide	3.00	2.95	0.05	0.38
Train 13	Uracil	3.56	3.03	0.53	0.38
Train 14	Barbital	3.87	3.38	0.50	0.38
Train 15	Hippuric acid	3.57	3.69	−0.12	0.37
Train 16	Isoprocarb	2.60	2.23	0.38	0.37
Train 17	2-Naphthalenesulfonic acid	4.78	4.28	0.50	0.36
Train 18	Thymine	3.58	3.52	0.06	0.36
Train 19	1,8-Anthracenedisulfonic acid, 9,10-dihydro-9,10-dioxo-	6.82	6.78	0.04	0.36
Train 20	6-Methyluracil	3.85	3.48	0.36	0.36
Train 21	5-Aminouracil	2.70	2.95	−0.25	0.36
Train 22	m-Acetotoluide	3.08	3.30	−0.22	0.36
Train 23	Acetanilide, 2'-ethyl-	3.62	3.16	0.46	0.36
Train 24	Benzoic acid	3.53	4.37	−0.84	0.35
Train 25	Probarbital	3.08	3.33	−0.24	0.35
Train 26	Butethal	3.69	3.80	−0.11	0.35
Train 27	Barbituric acid, 5-ethyl-5-pentyl-	3.18	3.99	−0.81	0.35
Train 28	1,4-Diaminoanthraquinone	−0.48	−0.27	−0.21	0.35
Train 29	Allobarbital	3.26	3.24	0.01	0.35
Train 30	1,5-Dihydropyrimido(5,4-d)pyrimidine-2,4,6,8(3H,7H)tetrone	−0.16	0.29	−0.45	0.41
Test 1	Phenytoin	1.76	2.52	−0.76	1.00
Test 2	Phenobarbital	3.05	2.74	0.30	0.56
Test 3	N,N'-Carbonylbis(acetamide)	4.85	4.65	0.20	0.45
Test 4	Benzenesulfonamide	3.63	3.99	−0.36	0.43
Test 5	Phthalamide	2.78	3.44	−0.66	0.41
Test 6	1,2-Benzoisothiazoline 1,1-dioxide	3.71	3.86	−0.15	0.40
Test 7	2-Toluenesulfonamide	3.21	4.08	−0.87	0.39
Test 8	Benzophenone	2.14	2.33	−0.19	0.38
Test 9	m-Toluenesulphonamide	3.89	4.23	−0.34	0.38
Test 10	Benzamide, o-methoxy-	3.40	3.50	−0.10	0.37
Test 11	Acetophenone	3.79	3.99	−0.20	0.36
Test 12	Uric acid	1.78	0.95	0.83	0.36
Test 13	1,5-Anthracenedisulfonic acid, 9,10-dihydro-9,10-dioxo-	5.82	5.01	0.81	0.36
Test 14	Barbituric acid, 5,5-dipropyl-	2.78	3.65	−0.87	0.36
Test 15	Thalidomide	2.74	2.60	0.14	0.35
Test 16	Urea, N-(4-methylphenyl)-N'-(1-methyl-1-phenylethyl)-	0.08	0.83	−0.75	0.35
Test 17	4-Toluenesulfonamide	3.50	4.36	−0.86	0.35
Test 18	Benzoyl peroxide	0.96	0.94	0.02	0.35
Test 19	1-Aminoanthraquinone	−0.52	0.15	−0.68	0.35
Test 20	Phthalic acid	3.85	3.31	0.53	0.35

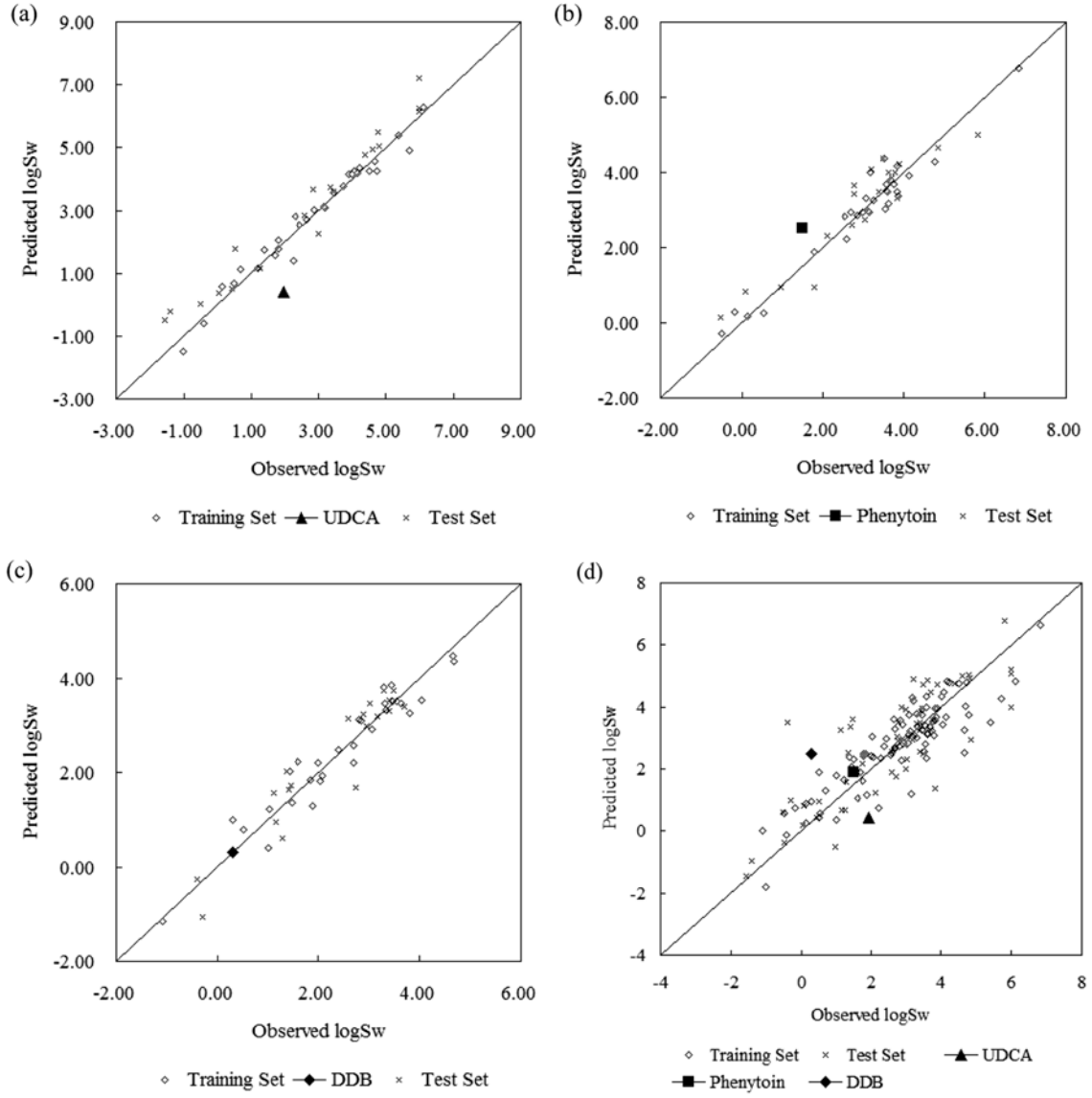

Fig. 3. The observed vs. predicted solubility for the (a) UDCA set, (b) Phenytoin set, (c) DDB set, (d) UDCA-Phenytoin-DDB set.

tation behavior of a molecule in an electrostatic field. The MolRef is the molar refractivity molecular descriptors that can be used to relate chemical structure to observed chemical behavior. The molecular refractivity index of a substituent is a combined measure of its size and polarizability. The density descriptor reflects the types of atoms and how tightly they are packed in a molecule and can be related to transport and melt behavior. The Jurs-WNSA-2 descriptor is the surface-weighted charged negative partial surface areas. The observed and calculated aqueous solubility for compounds in the phenytoin set is presented in Table 5.

The phenytoin set is the combinations of amide group, carboxylic acid, sulfonic, sulfonamide, hydratoin and uracil derivatives. That means this set has largely carbonyl and sulfonyl moiety in common. So, the electronic and polarizability property may affect the solubility of these compounds and is reflected in the Dipole-mag and MolRef descriptors.

4. Model for DDB Set

The training model for the DDB set has a coefficient of determination, R^2=0.925, Q^2=0.902, F-test=107.538, PRESS=5.004, and the remaining test set has R^2=0.903. Fig. 3(c) shows the predicted and experimental solubility for DDB set.

The regression equation for the DDB-like model is as follows:

$$\log Sw = 4.1411 - 0.876718*[AlogP98] + 1.42576*[S\_dssC] + 3.57233*[Jurs\text{-}RPCG] \quad (4)$$

E-state key, descriptor S_dssC, is related to the nature of sp^2 carbon atom. The descriptor Jurs-RPCG is the relative positive charge of the most positive atom divided by the total positive charge. The observed and calculated aqueous solubility for compounds in the DDB set are presented in Table 6.

This set has mainly aromatic rings and conjugated systems. Thus, the lipophilicity and the nature of sp^2 carbon atoms are important factors in these compounds. This nature can be quantitatively determined by S_dssC and AlogP98 descriptors. And, the UDCA, Phenytoin, and DDB sets are commonly influenced by the hydrophilic descriptor such as Jurs-WPSA-3, Jurs-WNSA-2, and Jurs-RPCG

Table 6. Observed and calculated aqueous solubility for compounds in the DDB set. All solubility data are logSw, where Sw is aqueous solubility in the unit of μg/ml

No.	IUPAC NAME	Observed	Predicted	Residuals	Similarities
Train 1	Benzoic acid, 4-methoxy-, methyl ester	2.81	3.12	−0.31	0.50
Train 2	Veratraldehyde	3.80	3.26	0.54	0.41
Train 3	Picropodophyllin	2.00	2.20	−0.20	0.46
Train 4	Piperonal	3.54	3.52	0.02	0.45
Train 5	Hydrastine	1.48	1.36	0.12	0.41
Train 6	Methyl benzoate	3.32	3.47	−0.15	0.41
Train 7	trans-3,4-(Methylenedioxy)cinnamic acid	1.43	2.01	−0.58	0.37
Train 8	3,3-Dimethylphthalide	3.34	3.33	0.02	0.37
Train 9	Benzaldehyde, 4-ethoxy-3-methoxy-	3.06	2.91	0.15	0.36
Train 10	Benzoic acid, 2-benzoyl-, methyl ester	1.90	1.29	0.61	0.35
Train 11	Methyl anthranilate	3.45	3.85	−0.39	0.35
Train 12	Methyleugenol	2.70	2.20	0.50	0.34
Train 13	Methyl salicylate	2.85	3.10	−0.25	0.34
Train 14	4-Anisaldehyde	3.63	3.47	0.16	0.33
Train 15	Vanillin	4.04	3.54	0.50	0.32
Train 16	Benzoic anhydride	1.00	0.40	0.60	0.32
Train 17	Benzene, 1,2-dichloro-4,5-dimethoxy-	1.86	1.84	0.01	0.31
Train 18	Oxolinic acid	0.51	0.79	−0.29	0.31
Train 19	Piperine	1.60	2.22	−0.62	0.31
Train 20	Methyl 3-chloro-4-hydroxybenzoate	2.40	2.48	−0.08	0.31
Train 21	Benzoic acid, 3-amino-2,5-dichloro-, methyl ester	2.08	1.93	0.15	0.31
Train 22	Phenyl phthalate	−1.09	−1.15	0.07	0.29
Train 23	Methanone, (4-methoxy-3-methylphenyl)(3-methylphenyl)-	0.30	1.00	−0.70	0.29
Train 24	Benzene, 1,2,3-trichloro-4,5-dimethoxy-	1.01	1.22	−0.21	0.28
Train 25	Benzoic acid, 2-(acetyloxy)-, (methylsulfonyl)methyl ester	2.04	1.82	0.22	0.30
Train 26	Vanillyl alcohol	3.30	3.81	−0.51	0.28
Train 27	Methylnicotinate	4.68	4.36	0.31	0.34
Train 28	Colchicine	4.65	4.46	0.19	0.31
Train 29	Acetophenone, 2,4'-dihydroxy-3'-methoxy-	3.48	3.50	−0.02	0.29
Train 30	Methylphenylsulfide	2.70	2.58	0.12	0.30
Test 1	DDB	0.29	0.32	−0.02	1.00
Test 2	Benzaldehyde, 3,4,5-trimethoxy-	3.17	3.19	−0.01	0.48
Test 3	beta-Peltatin	1.11	1.57	−0.46	0.44
Test 4	Meconin	3.40	3.29	0.10	0.44
Test 5	alpha-Peltatin	1.48	1.72	−0.24	0.40
Test 6	Benzoic acid, 2-(acetyloxy)-, (methylthio)methyl ester	2.74	1.68	1.06	0.38
Test 7	Benzaldehyde, 2,5-dimethoxy-	2.90	3.24	−0.34	0.37
Test 8	Chlorthal-dimethyl	−0.30	−1.06	0.76	0.34
Test 9	Piperonyl butoxide [BAN]	1.16	0.96	0.20	0.33
Test 10	Methoxymethyl salicylate	2.96	2.98	−0.02	0.32
Test 11	Methiocarb	1.43	1.62	−0.19	0.31
Test 12	Acetovanillone	3.70	3.40	0.30	0.31
Test 13	Trimethoprim [USAN:BAN:INN:JAN]	2.60	3.14	−0.54	0.31
Test 14	Carbamic acid, methyl-, 4-methylthio-m-tolyl ester	1.38	2.03	−0.65	0.31
Test 15	Phenyl acetylsalicylate	1.30	0.60	0.70	0.29
Test 16	Bifenox	−0.40	−0.27	−0.13	0.28
Test 17	Benzamide, o-methoxy-	3.40	3.53	−0.13	0.30
Test 18	Benzenesulfonic acid, methyl ester	3.49	3.75	−0.25	0.31
Test 19	Anise alcohol	3.30	3.75	−0.44	0.28
Test 20	Anisole	3.02	3.47	−0.45	0.30

which represent the hydrophilic groups positioned on the surface of molecule.

5. Model for UDCA-Phenytoin-DDB Set

The training model for the DDB set has a coefficient of determination, R^2=0.775, Q^2=0.732, F-test=47.596, PRESS=57.726, and the remaining test set has R^2=0.646. Fig. 3(d) shows the predicted and experimental solubility for the merged UDCA-Phenytoin-DDB set.

The regression equation for DDB-like model is as follows:

$$\begin{aligned}\log Sw = {} & 16.0671 - 0.130984*[MolRef] + 0.116783*[IAC\text{-}Total] \\ & + 0.709195*[S\_dssC] + - 0.005361*[Jurs\text{-}PNSA\text{-}2] \\ & - 1.18848*[RadOfGyration] - 7.99104*[Density] \end{aligned} \quad (5)$$

The radius of gyration is calculated by using the following equation:

$$\text{Radius of gyration} = \sqrt{\left(\sum \frac{(x_i^2 + y_i^2 + z_i^2)}{N}\right)} \quad (6)$$

where N is the number of atoms and x, y, and z are the atomic coordinates relative to the center of mass, respectively.

6. Limitations of this Approach

In Fig. 4, the histogram for log scale of aqueous solubility is listed with observed and each predicted value of various models. In the case of DDB molecule, predicted solubility by similarity models showed good agreement with observed logSw, but UDCA molecule was underestimated by all prediction models. One major weakness of this model is that it is not considered the ionizable state of compounds. In the deprotonated state the carboxylic acid group of UDCA compound will affect the solubility appreciably. This suggests that the solubility of weak acid or base compounds in water must require further consideration of the ionized species because the solute will determine the pH of the solution. Secondly, these are modeled in a temperature range between 15 °C and 35 °C. Such an extension of the temperature, for sufficient amount of experimental data, also may affect the accuracy of solubility prediction.

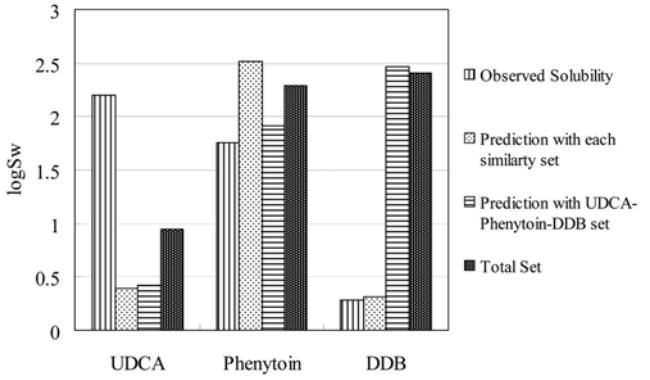

Fig. 4. Histogram of observed and predicted aqueous solubilities for the UDCA, Phenytoin and DDB molecules, respectively.

It may not be easy to make a robust prediction model for aqueous solubility of whole molecules with molecular descriptors set at present. The similarity factor was chosen for good model generation because this will eliminate the structurally common information about the extracted molecule set and also enable building prediction models with a few representations. The derived models with structurally similar molecules were validated with the UDCA, phenytoin, and DDB compounds included in the respective test set. For each model, the QSPR modeling process produced a set of descriptors that are significantly related to each set, in the statistical sense. The charged surface area, the size of molecule, and the lipophilicity descriptors always have been used as essential terms in various solubility prediction approaches.

The similarity model showed good agreement with observed aqueous solubility, except for UDCA molecule. The cholic acid series are almost the same structure except for the number of hydroxyl groups and their positions. The cholic acid and hyocholic acid are known to have different experimental solubility by about one log-unit (2.24, 1.26), but they are only distinguished from the positions of three hydroxyl groups. The solubilities of other cholic acid series are −0.42-0.52 log-unit and less two or three log-units than cholic acid. These molecules have 1-3 hydroxyl groups or additive amide groups. Therefore, the prediction would become more accurate when the descriptors which have the abilities to discriminate the structural isomers by the 3-dimensional distribution of functional groups are available.

Another characteristic of poorly water-soluble drugs is that they can be more dissolved in water by fine grinding. In Table 7, the experimental and calculated aqueous solubility data at 25 °C are listed together with data of particle size distribution. The experimental data

Table 7. Comparison of observed and predicted aqueous solubility at 25 °C

Drug	Median diameter (μm)		Sw[a]				Remark
	Intact	Ground	Observed		Calculated		
			Intact	Ground	Ref. [14]	This work	
UDCA	42.10	0.48[b]	160.00	338.00	2.41	2.29	Underestimated
Phenytoin	58.00	3.30[c]	58.00	106.00	178.60	331.10	Overestimated
DDB	22.70	0.30[d]	1.96	4.20	1.59	2.09	Good
					1.29[e]	-	

Experimental conditions for fine grinding of wet process:

[a]Aqueous solubility (μg/ml).

[b]Grinding time; 16 h with no additive, D_B=ϕ1.0 mm of alumina.

[c]Additive; PVA10; 20 wt%, D_B=ϕ1.0 mm of alumina, Grinding time; 30 min.

[d]Additive; DCNa; 0.5 g, Polysorbate80; 2.75 g, PVP10; 3.0 g, D_B=ϕ1.0 mm, Grinding time; 24 h.

[e]Calculation with COSHO-RS [16].

show that the grinding enhances the solubility to be about two-fold, so the discrepancy between the experimental and calculated values is enlarged. This effect could not be explained in this work either because the grinding effect cannot be considered quantitatively by the currently available descriptors.

CONCLUSION

In conclusion, a QSPR model for fast evaluation of aqueous solubility of organic compounds was developed based on a similarity set of literature data. The quality of the model was limited by the accuracy of the experimental measurements on the training set compounds. The modeling approach described here has produced powerful QSAR models for aqueous solubility and can also be used to predict more accurately the aqueous solubility of unknown compounds with a similar structure.

ACKNOWLEDGMENT

This work was supported by the Korea Science and Engineering Foundation (KOSEF) NRL Program grant funded by the Korea government (MEST) (R0A-2008-000-20024-0) and from an academic research program fund of Pusan National University from 2004 to 2006.

REFERENCES

1. M. J. Sung, B. J. Cha, W. S. Choi, J. H. Kim and S. H. Choi, *The 43rd symposium on powder science and technology*, Busan, Korea (2005).
2. J. H. Kim, D. H. Jung, H. K. Rhee, S. H. Choi, M. J. Sung and W. S. Choi, *Korean J. Chem. Eng.*, **25**, 171 (2008).
3. C. A. Lipinski, F. Lombardo, B. W. Dominy and P. J. Feeney, *Adv. Drug Deliv. Rev.*, **23**, 3 (1997).
4. J. C. Dearden, *Chemom. Intell. Lab. Syst.*, **24**, 77 (1994).
5. W. L. Jorgensen and E. M. Duffy, *Adv. Drug Deliv. Rev.*, **54**, 355 (2002).
6. J. Taskinen and J. Yliruusi, *Adv. Drug Deliv. Rev.*, **55**, 1163 (2003).
7. R. D. Cramer, *J. Med. Chem.*, **46**, 374 (2003).
8. ChemIDplus, http://chem.sis.nlm.nih.gov/chemidplus/.
9. PubChem, http://pubchem.ncbi.nlm.nih.gov/.
10. Cerius2., Accelrys, San Diego, CA.
11. L. H. Hall and L. B. Kier, *J. Chem. Inf. Comput. Sci.*, **35**, 1039 (1995).
12. J. Devillers, *Genetic algorithms in molecular modeling*, Academic Press, San Diego (1996).
13. B. E. Mitchel and P. C. Jurs, *J. Chem. Inf. Comput. Sci.*, **38**, 489 (1998).
14. D. T. Stanton and P. C. Jurs, *Anal. Chem.*, **62**, 2323 (1990).
15. R. H. Rohrbaugh and P. C. Jurs, *Anal. Chim. Acta*, **199**, 99 (1987).
16. W. Arlt, *Private communication on the prediction of thermo-physical data*, University of Erlangen, Germany (2005).