Kragujevac J. Sci. 30 (2008) 65-72. UDC 541.123

QSPR MODELLING OF WATER SOLUBILITY OF MINERALS BY OPTIMAL DESCRIPTORS CALCULATED WITH SMILES

Alla P. Toropova[a], Andrey A. Toropov[a,*] and Ivan Gutman[b]

[a]*Uzbekistan Academy of Sciences, Institute of Geology and Geophysics, 700041, Khodzhibaev Street 49, Tashkent, Uzbekistan*
e-mail: aatoropov@yahoo.com

[b]*Faculty of Science, P.O. Box 60, 34000 Kragujevac, Republic of Serbia*
e-mail: gutman@kg.ac.yu

(Received February 11, 2008)

ABSTRACT. Quantitative structure–property relationships (QSPR) for water solubility at 293 K (g per 100 g of water) of forty one minerals are obtained. Fluorides, chlorides, nitrates, sulfates, carbonates, and perchlorates have been examined. Optimal descriptors calculated with the Simplified molecular input line entry system (SMILES) have been utilized in one-varaible correlations with water solubility. Statistical characteristics of the best model are the following: n=31, r^2=0.9678, s=7.72 g, F=870 (training set) and n=10, r^2=0.9324, Q^2=0.8713, s=13.6 g, F=110.

INTRODUCTION

Correlation between molecular structure and water solubility is not simple. Multi-molecular complexes with water and different kinds of intermolecular interactions considerably influence the value of this parameter [1, 5]. However, even preliminary information on the physicochemical parameters is important for chemistry, geochemistry, and ecology [8].

Quantitative structure–property/activity relationships (QSPR/QSAR) can be utilized as a tool for predicting water solubility [2-4]. Recently, an approach based on the simplified molecular input line entry system (SMILES) in QSPR/QSAR analyses has been put forwards [6,7].

The present study is aimed to estimate the ability of optimal descriptors in QSPR modeling of water solubility of fluorides, chlorides, sulfides, nitrides, and carbonates. Numerical data have been taken from web site of http://wulfenite.fandm.edu/wired.html.

METHOD

The optimal descriptors are defined as

$$DCW = \prod_{k=1}^{N} CW(SF_k) \quad (1)$$

where SF_k is a fragment of the SMILES string: there is a hierarchy of four-, three-, two- and one-characters SMILES fragments. $CW(SF_k)$ is the correlation weight of SF_k; N is the number of different SF_k's, occurring in the given SMILES. Numbers of the brackets iof the form "(xxx", in particular, chlorine atoms "Clxx", fluorine "Fxxx", nitrogen "Nxxx", sulfur "Sxxx", and oxygen "Oxxx" also have been utilized as SF_k's. For instance, "Cl02" is an indicator of the presence of two chlorine atoms, "N003" is an indicator of the presence of three nitrogen atoms, etc. By Monte Carlo optimization one can calculate values of the $CW(SF_k)$ providing maximal values of the correlation coefficient between water solubility and DCW. Having numerical data on the $CW(SF_k)$ one can calculate the model of water solubility as follows:

$$S = C_0 + C_1 \cdot DCW \quad (2)$$

The predictive potential of the model can be validated with an external test set.

The SMILES strings used in the present study were obtained wby means of the ACD/ChemSketch Software (http://www.acdlabs.com/).

RESULTS

Statistical characteristics of the model for water solubility in three probes of the Monte Carlo method optimization are given in Table 1.

Table 1. Statistical quality of the mineral water solubility models

	Training set, n=31			**Test set, n=10**		
Probe	**R^2**	**S**	**F**	**R^2**	**S**	**F**
1	0.9678	7.72	870	0.9324	13.6	110
2	0.9689	7.58	905	0.8660	15.4	52
3	0.9620	8.39	734	0.9054	17.0	77

From the Table 1 one can see that their statistical quality is quite good. The correlation weights of the SF_k's are listed in Table 2.

Table 2. Correlation weights for calculating DCW with Eq. (1)
over three probes of the Monte Carlo optimization

SF_k	**$CW(SF_k)$ in Probe 1**	**$CW(SF_k)$ in Probe 2**	**$CW(SF_k)$ in Probe 3**
(	1.0179239	0.9291620	1.0026448
+	1.1736004	1.2377331	1.0050353
-	0.9865009	1.2111346	0.9890795

1	0.9789420	0.9682479	1.0376053
2	1.1950718	1.2537364	1.0109897
C	0.7415111	0.8902343	0.7653571
Cl	0.9084931	0.9027433	1.0302563
Co	1.0000000	1.0000000	1.0000000
Cu	1.1354664	1.0083906	1.0999443
F	1.1540900	1.1727901	1.1743495
Hg	1.3131877	1.1725905	1.2600849
K	0.8228276	0.9518912	0.9165604
O=	0.9880474	1.0188094	0.9716955
Cd	1.3430724	1.3250789	1.5299832
Mg	0.9319417	0.9253417	1.1410990
O	0.9030923	0.9738433	0.9781454
S	0.9832715	1.0153507	1.0361186
[	1.0458236	0.9081927	0.9792182
]	0.9519969	0.9754658	0.9960142
(000	1.0130855	1.1848373	0.9246453
(001	0.9729371	1.2476762	0.9318532
(002	0.8482164	0.9465379	0.7796312
(004	1.0371161	0.9879883	0.8596136
(006	1.1311182	1.0124663	0.9394737
F000	0.8960846	1.0285802	0.9359520
F001	1.0203790	1.1657322	1.1664141
Cl00	0.9625744	0.8867248	0.9824899
Cl01	0.9067350	0.9919463	1.0362622
Cl02	0.9787824	0.9773312	1.0199183
Cl03	1.0343771	0.9155597	0.9013054
N000	0.8223189	0.9378077	0.8944542
N001	0.7027616	0.7274991	0.7821577

N002	1.3756126	1.4078366	1.4224405
N003	0.8903920	0.9045474	1.0175186
O000	0.9161081	0.9913185	0.9354264
O002	0.6497540	0.6567631	0.8085218
O003	1.0356683	0.9819452	1.1216401
O004	1.1154570	1.3977252	1.1747286
O006	0.8869454	0.8888375	1.0491176
O009	1.0769127	1.0447932	1.2230952
O010	1.0000000	1.0000000	1.0000000
O011	1.0000000	1.0000000	1.0000000
O012	0.8916166	1.2072244	1.0473189
O015	1.0000000	1.0000000	1.0000000
S000	1.1194504	1.1238423	1.0997881
S001	0.7779598	0.9056306	0.8459749
S003	1.0121472	0.9964419	1.0526072
[Ag]	1.2149942	1.2280848	1.2879668
[Al]	0.9073082	1.1279163	0.9317839
[Co]	1.0000000	1.0000000	1.0000000
[Cs]	0.6659884	0.7192348	0.7522824
[Fe]	0.8700280	0.8695492	0.7843779
[N+]	0.9402310	1.0607259	1.0201286
[O-]	1.0483734	1.1600206	0.9948525
[Li]	1.0102469	1.0004765	1.0338469
[Mg]	0.9280325	0.9323149	0.9208065
[Na]	0.6971120	0.7250161	0.7536031
[Mn]	1.0000000	1.0000000	1.0000000
[Ni]	1.1477707	1.2689967	1.0973409
[Pb]	0.6982905	0.6985430	0.7174975

Experimental water solubilities as well as those calculated by means of the correlation equation:

$$S = -(56.5 \pm 1.2) + (192.3 \pm 1.9)\, DCW \qquad (3)$$

are listed in Table 3. Calculation of the DCW with $CW(SF_k)$ of the first probe is shown in Table 4. The quality of the model is illustrated graphically for the training set and for the test in Figs. 1 and 2, respectively.

Our final conclusion is that water solubility of minerals can be reliably calculated by means QSPR models based on SMILES.

References:

[1] K. V. BALAKIN, N. P. SAVCHUK, I. V. TETKO, In silico approaches to prediction of aqueous and DMSO solubility of drug-like compounds: Trends, problems and solutions, *Curr. Med. Chem.*, **13** (2006) 223-241.

[2] E. A. CASTRO, A. A.TOROPOV, A. I. NESTEROVA, O. M. NABIEV, O.M., QSPR modeling aqueous solubility of polychlorinated biphenyls by optimization of correlation weights of local and global graph invariants, *Centr. Eur. J. Chem.* **2** (2004) 500-523.

[3] P. R. DUCHOWICZ, E. A. CASTRO, A. A. TOROPOV, I. V. NESTEROV, O. M. NABIEV, O.M., QSPR modeling the aqueous solubility of alcohols by optimization of correlation weights of local graph invariants, *Mol. Div.* **8** (2004) 325-330.

[4] K. ROY, A. A. TOROPOV, QSPR modeling of the water solubility of diver functional aliphatic compounds by optimization of correlation weights of local graph invariants, *J. Mol. Model.* **11** (2005) 89-96.

[5] I. V. TETKO, V. Y. TANCHUK, T. N. KASHEVA, A. E. P. VILLA, Internet software for the calculation of the lipophilicity and aqueous solubility of chemical compounds, *J. Chem. Inf. Comput. Sci*. **41** (2001) 246-252.

[6] A. TOROPOV, K. NESMERAK, I. RASKA, K. WAISSER, K. PALAT, QSPR modeling of the half-wave potentials of benzoxazines by optimal descriptors calculated with the SMILES, *Comput. Biol. Chem*. **30** (2006) 434–437.

[7] A. A. TOROPOV, A. P. TOROPOVA, D. V. MUKHAMEDZHANOVA, I. GUTMAN, I., Simplified molecular input line entry system (SMILES) as an alternative for constructing quantitative structure-property relationships (QSPR), *Indian. J. Chem.* **44A** (2005) 1545-1552.

[8] A. P. TOROPOVA, A. A. TOROPOV, S. K. MAKSUDOV, QSPR modeling mineral crystal lattice energy by optimal descriptors of the graph of atomic orbitals, *Chem. Phys. Lett*. **428** (2006) 183–186.

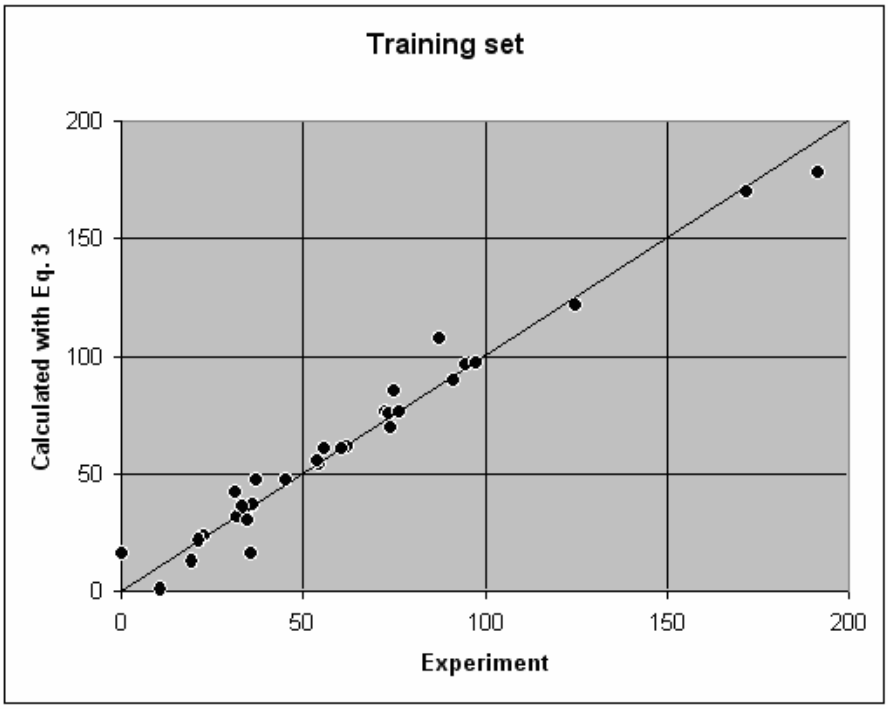
Table 3. Experimental values and those calculated by Eq. (3) of water solubility of minerals (g in 100 g of water). By an asterisk (*) are indicated the minerals that contain $SF^{\#}_k$ absent in the training set. Their $CW(SF^{\#}_k)$-values have been set equal to 1.

Mineral	**SMILES**	**DCW**	**Expr**	**Calc**	**Expr-Calc**
Training set					
KF	F[K]	0.7933709	94.9	96.05	-1.15
AgF	[Ag]F	1.1766510	172.0	169.73	2.27
$AlCl_3$	Cl[Al](Cl)Cl	0.5361112	45.8	46.59	-0.79
NH_4Cl	Cl.[NH4+]	0.5389602	37.2	47.13	-9.93
$BaCl_2*2H_2O$	O.O.Cl[Ba]Cl	0.4799627	35.8	35.79	0.01
$CaCl_2*6H_2O$	O.O.O.O.O.O.Cl[Ca]Cl	0.6551721	74.5	69.48	5.02
$CuCl_2$	Cl[Cu-]Cl	0.6897348	73.0	76.12	-3.12
$FeCl_2$	Cl[Fe]Cl	0.6127201	62.5	61.32	1.18
$FeCl_3*6H_2O$	O.O.O.O.O.O.Cl[Fe-](Cl)Cl	0.7571803	91.8	89.09	2.71
$MgCl_2$	Cl[Mg]Cl	0.5739573	54.6	53.86	0.74
$HgCl_2$	Cl[Hg-]Cl	0.7976909	98.0	96.88	1.12
$NiCl_2$	Cl[Ni]Cl	0.6066515	60.8	60.15	0.65
NaCl	Cl[Na]	0.3757157	35.9	15.75	20.15
$Al(NO_3)_3$	[O-][N+](=O)O[Al](O[N+]([O-])=O)O[N+](=O)[O-]	0.6833388	73.9	74.89	-0.99
NH_4NO_3	[O-][N+](O)=O.[NH4+]	1.2197782	192.0	178.02	13.98
$CsNO_3$	[O-][N+](=O)O[Cs]	0.4150108	23.0	23.31	-0.31
$Cu(NO_3)_2$	[O-][N+](=O)O[Cu-]O[N+]([O-])=O	0.9256266	125.0	121.47	3.53
$Pb(NO_3)_2$	[O-][N+](=O)O[Pb]O[N+]([O-])=O	0.5795704	54.3	54.94	-0.64
KNO_3	[O-][N+](=O)O[K]	0.5104998	31.6	41.66	-10.06
$NaNO_3$	[O-][N+](=O)O[Na]	0.8503220	87.6	106.99	-19.39
$Al_2(SO_4)_3$	O=S(=O)(O[Al]1OS(=O)(=O)O1)O[Al]2OS(=O)(=O)O2	0.4844273	36.4	36.65	-0.25
$(NH_4)SO_4$	OS(O)(=O)=O.[NH4+].[NH4+]	0.7340923	75.4	84.65	-9.25
$CdSO_4$	[Cd+]OS(O)(=O)=O	0.6886657	76.6	75.92	0.68
$CaSO_4*2H_2O$	O.O.[Ca+]OS(O)(=O)=O	0.3749009	0.26	15.59	-15.33
$CuSO_4*5H_2O$	O.O.O.O.O.[Cu]OS(O)(=O)=O	0.4551977	32.0	31.03	0.97
Li_2SO_4	[Li]OS(=O)(=O)O[Li]	0.4478678	34.8	29.62	5.18
$MgSO_4$	[Mg+]OS(O)(=O)=O	0.4778568	33.7	35.39	-1.69
K_2SO_4	[K]OS(=O)(=O)O[K]	0.2945103	11.1	0.14	10.96

Na_2SO_4	[Na]OS(=O)(=O)O[Na]	0.3567433	19.5	12.10	7.40
Na_2CO_3	[Na]OC(=O)O[Na]	0.4027033	21.5	20.94	0.56
$LiClO_4$	[Li]OCl(=O)(=O)=O	0.6074491	56.1	60.30	-4.20
Test set					
$CoCl_2$	Cl[Co]Cl *	0.6184668	52.9	62.42	-9.52
LiCl	Cl[Li]	0.6371131	83.5	66.00	17.50
$MnCl_2$	Cl[Mn]Cl *	0.6184668	73.9	62.42	11.48
$Ca(NO_3)_2*4H_2O$	O.O.O.O.[O-][N+](=O)O[Ca]O[N+]([O-])=O *	1.0174735	129.0	139.13	-10.13
$Co(NO_3)_2$	[O-][N+](=O)O[Co]O[N+]([O-])=O *	0.8299847	97.4	103.09	-5.69
$Fe(NO_3)_2*6H_2O$	O.O.O.O.O.O.[O-][N+](=O)O[Fe]O[N+]([O-])=O	1.0330320	140.0	142.12	-2.12
$Fe(NO_3)_2*9H2O$	O.O.O.O.O.O.O.O.O.[O-][N+](=O)O[Fe]O[N+]([O-])=O *	1.1586056	138.0	166.26	-28.26
$LiNO_3$	[O-][N+](=O)O[Li]	0.6295356	70.1	64.55	5.55
$CoSO_4$	[Co+]OS(O)(=O)=O *	0.5127539	36.1	42.10	-6.00
$FeSO_4*7H_20$	O.O.O.O.O.O.O.[Fe+]OS(O)(=O)=O *	0.4813179	48.0	36.05	11.95

Table 4. Calculation of the DCW for KF (SMILES=F[K]; DCW=0.7933709)

SF_k	**$CW(SF_k)$ of first probe**	**SF_k**	**$CW(SF_k)$ of first probe**
F	1.1540900	F001	1.0203790
[	1.0458236	Cl00	0.9625744
K	0.8228276	S000	1.1194504
]	0.9519969	N000	0.8223189
(000	1.0130855	O000	0.9161081



[9]

Figure 1. Experimental versus calculated (by Eq. (3)) values of water solubility of minerals for the training set.

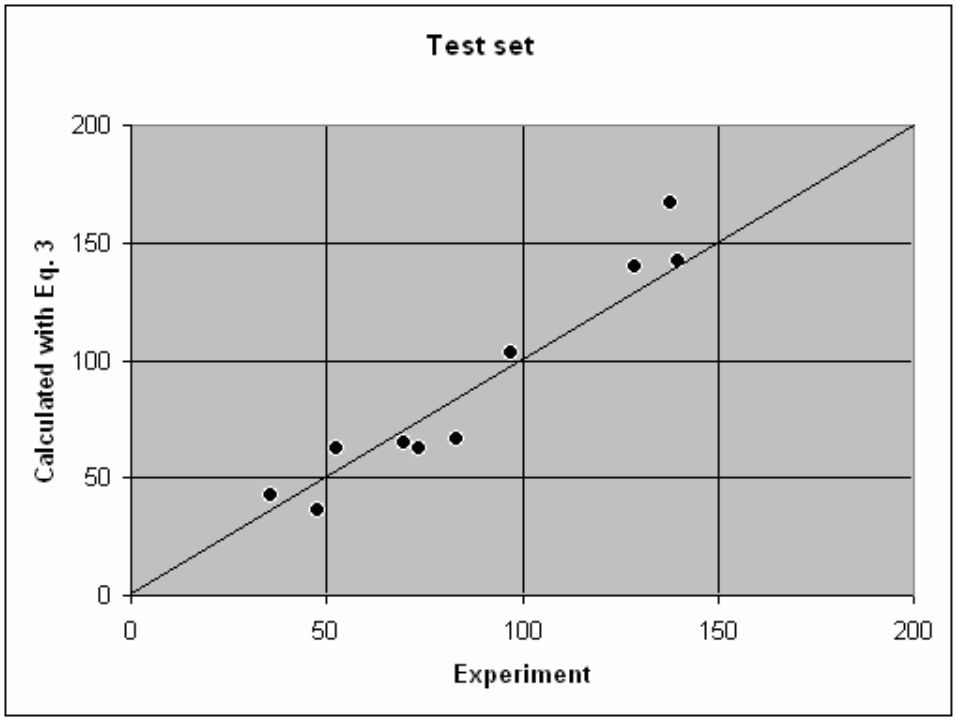


Figure 2. Experimental versus calculated (by Eq. (3)) values of water solubility of minerals for the test set.