WHIM-3D-QSPR APPROACH FOR PREDICTING AQUEOUS SOLUBILITY OF CHLORINATED HYDROCARBONS

Oman Zuas

Analytical Chemistry and Standard Division, Research Centre for Chemistry, Indonesian Institute of Sciences, Kawasan PUSPIPTEK, Serpong, 15314, Tangerang, Banten, Indonesia.

Received 24 October 2007; Accepted 15 December 2007

ABSTRACT

The weighted holistic invariant molecular-three dimensional-quantitative structure property relationship (WHIM-3D-QSPR) approach has been applied to the study of the aqueous solubility (– log Sw) of chlorinated hydrocarbon compounds (CHC's). The obtained QSPR model is predictive and only requires four WHIM-3D descriptors in the calculation. The correlation equation of the model that is based on a training set of 50 CHC's compound has statistical parameters: standard coefficient correlation (R^2) = 0.948; cross-validated correlation coefficients (Q^2) = 0.935; Standard Error of Validation (SEV) = 0.35; and average absolute error (AAE) = 0.31. The application of the best model to a testing set of 50 CHC's demonstrates a reliable result with good predictability. Besides, it was possible to construct new model by applying WHIM-3D-QSPR approach without require any experimental physicochemical properties in the calculation of aqueous solubility.

Keywords: *WHIM-3D; QSPR; aqueous solubility; – Log Sw, chlorinated hydrocarbons, CHC's.*

INTRODUCTION

The aqueous solubility of organic compounds is an essential molecular property that plays a large role in the behaviour of compounds in many areas of concern. The aqueous solubility is also probably one of the most fundamental and deserves attention in the early phases of drug discovery and development. With regard to the importance of solubility, the prediction based solely on molecular structure should prove a useful tool, because the solubility of many existing compounds is not available. Additionally, the solubility of chemicals and drugs in the water phase has also an essential influence on the extent of their absorption and transport in a body. For that reason, the aqueous solubility is considered to be a very important parameter in current Absorption, Distribution, Metabolism, Excretion, and Toxicity (ADMET) research [1-9].

The solubility data of compounds has widespread relevance to many branches and disciplines of science such as medicine, technology, and engineering. This fact has led to the development of several models to predict aqueous solubility of compound interest by using only theoretically derived descriptors without any experimental physicochemical properties. Therefore, reliable computational methods to predict aqueous solubility are more popular in today's research in comparison with time-consuming experimental procedures to determine aqueous solubility [1,7-14].

Predictive models for aqueous solubility are generally based on a diverse set of descriptors such as experimentally based descriptors, molecular properties, and collection of relevant structural features, which are correlated to activity by means of various statistical techniques including multiple linear regression (MLR) and neural networks (NN) [1,7,15]. In a quantitative structure property relationship (QSPR) study is that there is some natures of relationship between the physical property of interest such as aqueous solubility 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. Although the descriptors used to construct a QSPR model can be empirical, it is generally more useful to use descriptors derived mathematically from the three dimensional (3D) molecular structure, because of this allow any relationship so derived to be extended to the prediction of the property for unavailable compounds [9].

Not surprisingly, many of computational methods for the estimation of aqueous solubility has been extensively studied and reported [1-8,15,16]. However, to the best of our knowledge there is no published literature that reported the study on prediction of aqueous solubility of chlorinated hydrocarbon compounds (CHC's) based on the weighted holistic invariant molecular-three dimensional (WHIM-3D) QSPR approach. The aim of this study, therefore, was to investigate the molecular descriptors important based on WHIM-3D-QSPR approach in determining aqueous solubility of a heterogeneous group of 50 CHC's (training set) as listed in Table 1. The stepwise MLR was used to select the most informative descriptors from the calculated descriptors. Leave-one-out (LOO) cross validation method was used to assess the robustness of the model. The best QSPR model obtained was then used to predict the aqueous solubility for testing set of 50 CHC's as also listed in Table 1.

* Corresponding author. Tel: +62-21-7560929, Fax: +62-21-7560549; Email address : oman.zuas@yahoo.co.id

Table 1. Experimental values of the molar aqueous solubility (-log Sw) of the 100 CHC's

Cpd. No.	CAS No.	Name	Experimental -log Sw
Training set			
1	75-09-2	Dichloromethane	0.74
2	67-66-3	Trichloromethane	1.19
3	56-23-5	Tetrachloromethane	2.26
4	79-34-5	1,1,2,2-Tetrachloroethane	1.76
5	540-59-0	1,2-Dichloroethene	1.07
6	79-01-6	Trichloroethene	2.04
7	127-18-4	Tetrachloroethene	2.57
8	108-90-7	Monochloro benzene	2.42
9	541-73-1	1,3-Dichloro benzene	3.04
10	95-50-1	1,2-Dichloro benzene	3.02
11	106-46-7	1,4-Dichloro benzene	3.31
12	120-82-1	1,2,4-Trichloro benzene	3.64
13	87-61-6	1,2,3-Trichloro benzene	4.08
14	108-70-3	1,3,5-Trichloro benzene	4.55
15	634-66-2	1,2,3,4-Tetrachloro benzene	4.38
16	95-94-3	1,2,4,5-Tetrachloro benzene	5.19
17	634-90-2	1,2,3,5-Tetrachloro benzene	4.73
18	608-93-5	Pentachloro benzene	5.37
19	39227-53-7	1-Chloro dibenzo-p-dioxin	5.72
20	39227-54-8	2-Chloro dibenzo-p-dioxin	5.86
21	29446-15-9	2,3-Dichloro dibenzo-p-dioxin	7.23
22	33857-26-0	2,7-Dichloro dibenzo-p-dioxin	7.83
23	39227-58-2	1,2,4-Trichloro dibenzo-p-dioxin	7.53
24	30746-58-8	1,2,3,4-Tetrachloro dibenzo-p-dioxin	8.77
25	2051-60-7	2-Chloro biphenyl	4.63
26	2051-61-8	3-Chloro biphenyl	4.88
27	2051-62-9	4-Chloro biphenyl	5.25
28	2050-68-2	4,4'-Dichloro biphenyl	6.63
29	34883-39-1	2,5-Dichloro biphenyl	5.27
30	33284-50-3	2,4-Dichloro biphenyl	5.29
31	33146-45-1	2,6-Dichloro biphenyl	5.07
32	2050-68-2	2,4'-Dichloro biphenyl	5.60
33	13029-08-8	2,2'-Dichloro biphenyl	5.36
34	37680-65-2	2,2',5-Trichloro biphenyl	5.65
35	35693-92-6	2,4,6-Trichloro biphenyl	6.07
36	15862-07-4	2,4,5-Trichloro biphenyl	6.27
37	32598-13-3	3,3',4,4'-Tetrachloro biphenyl	8.68
38	35693-99-3	2,2',5,5'-Tetrachloro biphenyl	6.44
39	33284-53-6	2,3,4,5-Tetrachloro biphenyl	7.26
40	18259-05-7	2,3,4,5,6-Pentachloro biphenyl	7.78
41	37680-73-2	2,2',4,5,5'-Pentachloro biphenyl	7.44
42	55312-69-1	2,2',3,4,5-Pentachloro biphenyl	7.10
43	74472-44-9	2,3,3',4',5,6-Hexachloro biphenyl	7.83
44	55215-18-4	2,2',3,3',4,5-Hexachloro biphenyl	8.04
45	33979-03-2	2,2',4,4',6,6'-Hexachloro biphenyl	8.48
46	35065-27-1	2,2',4,4',5,5'-Hexachloro biphenyl	8.57
47	38411-22-2	2,2',3,3',6,6'-Hexachloro biphenyl	7.86
48	38380-07-3	2,2',3,3',4,4'-Hexachloro biphenyl	9.00
49	2136-99-4	2,2',3,3',5,5',6,6'-Octachloro biphenyl	9.30
50	40186-72-9	2,2',3,3',4,4',5,5',6-Nonachloro biphenyl	9.93
Testing set			
51	16606-02-3	2,4',5-trichlorobiphenyl	6.25
52	31508-00-6	2,3',4,4',5-Pentachlorobiphenyl	7.39
53	32598-11-1	2,3',4',5-Tetrachlorobiphenyl	7.25
54	35065-28-2	2,2',3,4,4',5'-Hexachlorobiphenyl	8.32
55	35694-08-7	2,2',3,3',4,4',5,5'-octachlorobiphenyl	9.16
56	38380-02-8	2,2',3,4,5'-Pentachlorodiphenyl	7.91
57	38380-08-4	2,3,3',4,4',5-Hexachlorobiphenyl	7.82
58	38444-85-8	2,3,4'-Trichlorobiphenyl	6.26
59	41464-39-5	2,2',3,5'-Tetrachlorobiphenyl	6.47
60	52663-63-5	2,2',3,5,5',6-Hexachlorobiphenyl	7.42
61	52663-69-1	2,2',3,4,4',5',6-Heptachlorobiphenyl	7.92
62	52712-04-6	2,2',3,4,5,5'-Hexachlorobiphenyl	7.68
63	52712-05-7	2,2',3,4,5,5',6-Heptachlorobiphenyl	8.94
64	55215-17-3	2,2',3,4,6-Pentachlorobiphenyl	7.43

Table 1. (cont.)

Cpd. No.	CAS No.	Name	Experimental -log Sw
65	55702-45-9	2,3,6-Trichlorobiphenyl	6.29
66	56558-16-8	2,2',4,6,6'-Petachlorobiphenyl	7.32
67	74472-42-7	2,3,3',4,4',6-Hexachlorobiphenyl	7.66
68	75-09-2	dichloromethane	0.63
69	67-66-3	trichloromethane	1.17
70	56-23-5	tetrachloromethane	2.31
71	75-34-3	1,1-dichloroethane	1.29
72	107-06-2	1,2-dichloroethane	1.06
73	109-69-3	1-chlorobutane	2.03
74	78-86-4	2-chlorobutane	1.96
75	513-36-0	1-chloro-2-methylpropane	2.00
76	541-33-3	1,1-dichlorobutane	2.40
77	7581-97-7	2,3-dichlombulane	2.70
78	543-59-9	1-chloropentane	2.73
79	625-29-6	2-chloropenlane	2.63
80	616-20-6	3-chloropentane	2.63
81	594-36-5	2-chloro-2-methylbutane	2.51
82	544-10-5	1-chlorohexane	3.12
83	319-86-8	α -hexachlorocyclohexane	4.51
84	75-35-4	1,1-dichlorcethylene	1.64
85	156-59-2	1,2-dichlorcethylene	1.30
86	107-05-1	3-chloropropylene	1.36
87	87-68-3	hexachloro-1,3-butadiene	4.92
88	77-47-4	hexachlorocyclo-pentadiene	5.18
89	95-49-8	2-chlorotoluene	3.52
90	100-44-7	α-chlorotoluene	2.39
91	106-43-4	p-chlorotoluene	3.08
92	38444-93-8	2,2',3,3'-tetrachlorobiphenyl	7.28
93	32598-10-0	2,3',4,4'-tetrachlorobiphenyl	7.80
94	41464-40-8	2,2',4,5'-tetrachlorobiphenyl	6.57
95	15968-05-5	2,2',6,6'-tetrachlorobiphenyl	8.03
96	52704-70-8	2,2',3,3',5,6-pentachlorobiphenyl	8.60
97	50-29-3	DDT	7.15
98	72-55-9	DDE	6.90
99	91-58-7	2-chloronaphthalene	4.14
100	90-13-1	1-chloronaphthalene	3.93

EXPERIMENTAL SECTION

Data set

In this work, a set of 100 CHC's were studied. Their chemical names are listed in Table 1. All CHC's together with their experimental – log Sw values were taken from the work of Wang *et al*. [5]. The CHC's were divided into training set (50 compounds) and a testing set (50 compounds). Both the training and the testing set contain saturated, unsaturated, aliphatic and aromatic compounds, dioxins and polychlorinated biphenyls.

Descriptors calculations

Firstly, the chemical structures of all molecules were built using HyperChem Release 7.0 for Windows [17] and were implemented in the DRAGON version 5.4 software [18], for the WHIM-3D descriptors calculation. The WHIM-3D descriptors used in this work are listed in Table 2.

Modeling and prediction

Milano Chemometric and QSAR Research Group of Professor Roberto Todeschini firstly developed the WHIM descriptor [19]. The WHIM descriptors are three-dimensional descriptor based on the calculation of principal component axes calculated from a weighted covariance matrix obtained by the molecule geometric coordinated. Six different weighting schemes are used for weighted covariance matrix i.e., *u* (unweighted); *m* (atomic mass), *p* (atomic polarizability), *v* (van der Waals volume), *e* (atomic electronegativity), and *s* (atomic electronegativity state). The WHIM descriptors are consisted of 99 descriptors and contain chemical information concerning size, symmetry, shape and distribution of the molecule atoms, more complete definition of the WHIM-3D descriptors can be found in the literatures [20-24]. After the calculation of eight WHIM-3D descriptors (Table 2), the stepwise MLR was used build the QSPR model by means of the SPSS Release 12.0 for Windows [25]. The classical QSPR regression equation can be obtained by the use of the scaled regression coefficients, mean and standard deviation of each original descriptor. The statistical parameters used to assess the quality of the models are the Standard Error of Validation (SEV) and the standard correlation coefficients (R^2), and cross-validated correlation coefficients (Q^2) are given by Eq. 1, 2 and 3, respectively [5, 26-29]. The best model

Table 2. The notation of the WHIM-3D descriptors involved in the QSPR model.

Notation	Descriptors
L1m	1st component size directional WHIM index / weighted by atomic masses
L2m	2nd component size directional WHIM index / weighted by atomic masses
L3m	3rd component size directional WHIM index / weighted by atomic masses
P1m	1st component shape directional WHIM index / weighted by atomic masses
P2m	2nd component shape directional WHIM index / weighted by atomic masses
G1m	1st component symmetry directional WHIM index / weighted by atomic masses
G2m	2st component symmetry directional WHIM index / weighted by atomic masses
G3m	3st component symmetry directional WHIM index / weighted by atomic masses
E1m	1st component accessibility directional WHIM index / weighted by atomic masses
E2m	2nd component accessibility directional WHIM index / weighted by atomic masses
E3m	3rd component accessibility directional WHIM index / weighted by atomic masses

derived from the MLR analysis was used to predict the –log Sw of the testing set compounds (Table 1) which were not included in the training set.

$$SEV = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{n}} \qquad (1)$$

$$R^2 = 1 - \left(\frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2_{experimental}}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \right) \qquad (2)$$

$$Q^2 = 1 - \left(\frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n}\left((y_i - \bar{y})^2\right)} \right) \qquad (3)$$

In these equations, n is the number of compounds used for cross-validation, y^j is the experimental value of the physicochemical property for the i^{th} sample. $\hat{y}_i$ is the value predicted by the model built without sample i. y^j is the mean value of experimental physicochemical property. The average absolute error (AAE) (Eq. 4) was calculated as the following equation.

$$AAE = \frac{\sum\left|\left(-\log Sw_{predicted}\right) - \left(-\log Sw_{experimental}\right)\right|}{n} \qquad (4)$$

Where –log $Sw_{predicted}$ are predicted values of the aqueous solubility, –log $Sw_{experimental}$ are the experimental values of the aqueous solubility, and n is number of compounds

RESULT AND DISCUSSION

The experimental data of 50 CHC's listed as training set in Table 1 were used to construct the regression models and set as dependent variable. Eleven WHIM-3D descriptors (Table 2) are set as independent variables. The values of all descriptors are listed in Table 3 except for L3m, G3m and E3m since these three descriptors have constant values and will have no effect on the statistical calculation. That why these three descriptors were removed from the table. To obtain the best QSPR model all possible combinations of the WHIM-3D descriptors were investigated. The stepwise MLR was used to select each independent variable for deriving a QSPR model by considering the correlation between each variable with the dependent variable. Of eight descriptors (Table 3), four descriptors have been automatically selected to model – log Sw *i.e.* L1m, L2m, G1m, and G2m. It should be mentioned that more models were obtained from the MLR analysis, but they were ruled out by the stepwise MLR procedure. The general purpose of MLR is to quantify the relationship between several independent variables (WHIM-3D descriptors) and a dependent variable (-log Sw) of CHC's. A set of coefficients defines the single linear combination of independent variables that best describes aqueous solubility of CHC's. The MLR equation used for the QSPR model developed is as follows:

$$Y = a_1b_1 + a_2b_2 + a_3b_3 .....a_nb_n + c \qquad (5)$$

Where Y is dependent variable. $a_1, a_2, a_3 .....a_n$ is the regression coefficients of independent variables. $b_1, b_2, b_3 .....b_n$ are independent variables. c is the regression constant obtained from the model fit. To avoid self correlation between the variables used for the derivation of the QSPR model, the correlation matrix of eight selected descriptor was calculated and the result shown in Table 4. The best QSPR model obtained from the MLR analysis is shown in Equation 6. While the statistical parameters values of SEV and AAE of prediction set for the MLR model were 0.35 and zero, respectively.

$$\begin{aligned} -\log Sw = {} & 1.581\,(\pm 0.0831)\,L1m + \\ & 4.374\,(\pm 0.294)\,L2m + 0.911\,(\pm 0.225)\,G1m + \\ & 0.664\,(\pm 0.227)\,G2m - 1.952\,(\pm 0.337) \end{aligned} \qquad (6)$$

(n=50, R^2=0.948, Q^2 = 0.935, F = 204.968, s= 0.564)

Based on this relationship the –log Sw of 50 CHC's were predicted and the accuracy of the predictions was then assessed by the residuals between the experimental and predicted values. The plot of experimental and predicted –log Sw of compounds in the training set based on the QSAR equation above is given in Fig 1. The results indicate a

Table 3. The calculated WHIM-3D descriptors values of CHC's as training compounds used in QSPR models.

Cpd. No	-log Sw	L1m	L2m	P1m	P2m	G1m	G2m	E1m	E2m
1	0.74	0.426	0.202	0.657	0.312	1.000	0.301	1.742	0.747
2	1.19	0.357	0.303	0.541	0.459	0.301	1.000	0.617	0.882
3	2.26	0.314	0.314	0.500	0.500	1.000	1.000	0.531	0.531
4	1.76	0.599	0.443	0.575	0.425	1.000	1.000	1.504	0.522
5	1.07	0.559	0.471	0.543	0.457	1.000	0.279	0.712	0.710
6	2.04	0.679	0.355	0.656	0.344	0.679	0.279	1.024	0.204
7	2.57	0.607	0.437	0.582	0.418	1.000	1.000	0.116	1.097
8	2.42	1.162	0.316	0.786	0.214	0.218	1.000	0.655	0.039
9	3.04	1.189	0.574	0.674	0.326	1.000	0.218	0.701	0.132
10	3.02	1.239	0.557	0.690	0.310	0.218	1.000	0.774	0.124
11	3.31	1.505	0.242	0.862	0.138	1.000	1.000	1.418	0.023
12	3.64	1.413	0.527	0.728	0.272	0.218	0.218	1.248	0.114
13	4.08	1.018	0.963	0.514	0.486	0.218	1.000	0.523	0.460
14	4.55	0.963	0.963	0.500	0.500	0.436	1.000	0.460	0.460
15	4.38	1.240	0.844	0.595	0.405	1.000	0.218	1.000	0.359
16	5.19	1.455	0.595	0.710	0.290	1.000	1.000	1.433	0.145
17	4.73	1.251	0.810	0.607	0.393	0.218	1.000	1.020	0.325
18	5.37	1.255	0.894	0.584	0.416	1.000	0.218	1.066	0.411
19	5.72	2.206	0.737	0.750	0.250	0.781	0.183	0.243	0.151
20	5.86	3.126	0.462	0.871	0.129	0.183	0.781	0.488	0.072
21	7.23	3.784	0.536	0.876	0.124	0.183	1.000	0.731	0.108
22	7.83	3.875	0.414	0.903	0.097	1.000	1.000	0.818	0.054
23	7.53	2.877	0.994	0.743	0.257	0.183	0.531	0.416	0.394
24	8.77	3.424	0.998	0.774	0.226	0.183	1.000	0.598	0.469
25	4.63	1.683	0.739	0.695	0.305	0.781	0.183	0.178	0.196
26	4.88	2.578	0.591	0.813	0.187	0.183	0.183	0.463	0.096
27	5.25	3.166	0.377	0.894	0.106	0.183	1.000	0.675	0.051
28	6.63	3.947	0.318	0.925	0.075	1.000	1.000	1.147	0.036
29	5.27	2.173	0.837	0.722	0.278	0.702	0.635	0.328	0.200
30	5.29	2.704	0.630	0.811	0.189	0.183	0.198	0.493	0.149
31	5.07	1.437	0.943	0.604	0.396	0.183	1.000	0.128	0.356
32	5.60	2.711	0.632	0.811	0.189	0.183	0.198	0.496	0.150
33	5.36	1.500	0.883	0.630	0.370	1.000	1.000	0.154	0.214
34	5.65	1.888	1.021	0.649	0.351	0.198	0.183	0.249	0.305
35	6.07	2.357	0.817	0.743	0.257	0.531	1.000	0.374	0.267
36	6.27	2.962	0.771	0.793	0.207	0.183	0.531	0.618	0.212
37	8.68	4.155	0.581	0.877	0.123	1.000	1.000	1.400	0.108
38	6.44	2.361	0.937	0.716	0.284	1.000	1.000	0.465	0.189
39	7.26	3.092	0.959	0.763	0.237	0.183	0.198	0.666	0.390
40	7.78	2.773	1.071	0.721	0.279	0.183	1.000	0.534	0.521
41	7.44	2.894	0.938	0.755	0.245	0.183	0.781	0.646	0.273
42	7.10	2.797	1.062	0.725	0.275	0.183	0.531	0.547	0.501
43	7.83	2.940	1.120	0.724	0.276	0.183	0.183	0.633	0.528
44	8.04	2.993	1.083	0.734	0.660	0.183	0.183	0.671	0.456
45	8.48	2.506	0.969	0.721	0.279	1.000	1.000	0.462	0.427
46	8.57	3.365	0.886	0.792	0.208	1.000	1.000	0.912	0.283
47	7.86	2.133	0.954	0.691	0.309	1.000	1.000	0.481	0.139
48	9.00	2.133	0.954	0.691	0.309	1.000	1.000	0.481	0.139
49	9.30	2.134	1.462	0.593	0.407	1.000	1.000	0.299	1.326
50	9.93	3.194	1.204	0.726	0.274	0.183	0.253	0.804	0.824

good linear regression of correlation between experimental and predicted –log Sw.

In previous study, Wang *et al*. [5] used molecular connectivity indices to develop a QSPR model for CHC's. They achieved in correlating the three connectivity indices descriptors that reflect the contribution of clusters in a molecule to aqueous solubility that are important in describing the aqueous solubility of CHC's. Another study reported by Delgado [16] demonstrated that CODESSA has been successfully in applying to develop QSPR model and carried out a correlation analysis to find the best QSPR model using a heuristic method. Delgado attained in obtaining the two descriptors that have definite physical meaning corresponding to different intermolecular interactions.

Table 4. The correlation matrix of all WHIM-3D descriptors used in QSPR studies

	-log Sw	L1m	L2m	P1m	P2m	G1m	G2m	E1m	E2m
-log Sw	1	0.825	0.705	0.449	-0.325	-0.135	0.097	-0.238	-0.049
L1m		1	0.299	0.814	-0.671	-0.269	0.023	-0.083	-0.325
L2m			1	-0.194	0.265	-0.214	-0.096	-0.454	0.314
P1m				1	-0.887	-0.195	-0.001	0.151	-0.604
P2m					1	0.103	-0.086	-0.154	0.567
G1m						1	0.205	0.263	0.077
G2m							1	-0.004	0.003
E1m								1	-0.131
E2m									1

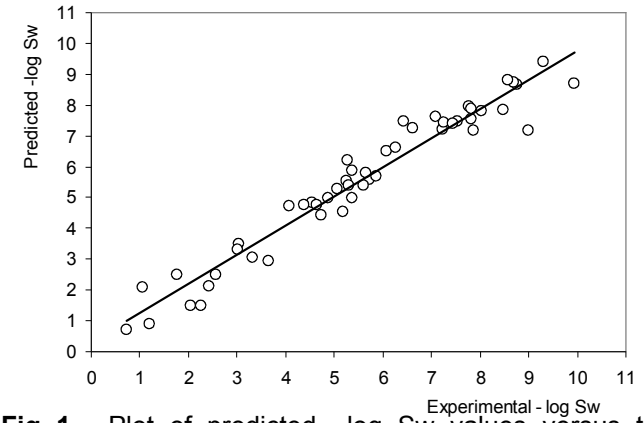

Fig 1. Plot of predicted –log Sw values versus the experimental –log Sw values of CHC's in the training set.

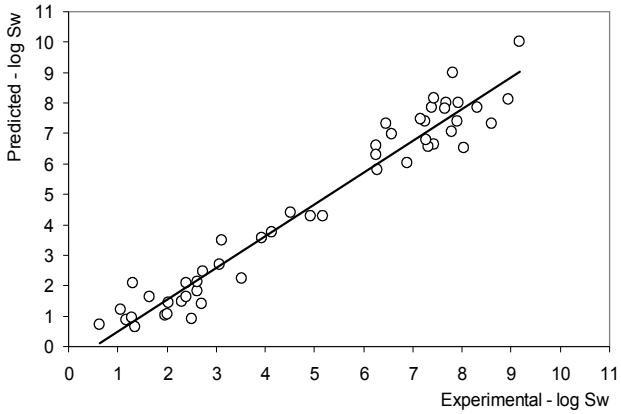

Fig 2. Plot of predicted –log Sw values versus the experimental –log Sw values of CHC's in the testing set.

In this study, the WHIM-3D descriptors were used to predict the aqueous solubility of CHC's. WHIM descriptors are the molecular descriptors based on statistical indices calculated on the projections of the atoms along principal axes. They are built in such a way as to capture relevant molecular 3-dimensional information regarding molecular size, shape, symmetry, and atom distribution with respect to invariant reference frames [21, 22, 24]. The fact that the WHIM descriptor is derived from three-dimensional representing of a molecule, seems to indicate a connection between the molecular structures of the physicochemical properties of compounds. The result of this study demonstrated that the L1m and L2m that reflect the size of molecule is the most significant descriptor, as can be seen by its highest correlation values with –log Sw of CHC's (Table 4). The other descriptors G1m and G2m that reflect the contribution of symmetry of the molecule are also important in describing the aqueous solubility of CHC's compound. The positive value of the coefficient for all descriptors implies that a high size and symmetry of the CHC's molecule correlates with decreased the solubility of the CHC's. To test the predictive ability of the model obtained in this study (Eq. 6), the aqueous solubility data for 50 CHC's taken from the Wang *et al*. [5] were predicted. The scatter plot of obtained predictive results together with their experimental values is given in Fig 2.

The prediction results are in good agreement with the experimental values. The high cross-validated coefficient correlation (Q^2= 0,935) and low average absolute error (AAE = 0.31) observed indicated that the developed QSPR model is reliable and has good predictability.

CONCLUSION

QSPR studies are an important tool for research and knowledge of chemical compounds and it has been frequently used in medicinal chemistry and molecular design to investigate new drugs. Predictive QSPR model that is based on WHIM-3D is suggested in this study to correlate the aqueous solubility of 50 CHC's. The application of the best model obtained to a testing set of 50 CHC's demonstrates that the new model is reliable with good predictability. Besides, it was possible to construct new model by applying WHIM-3D approach without require any experimental physicochemical properties in the calculation of aqueous solubility.

ACKNOWLEDGEMENT

The authors thank Umi Karomah Yaumidin, S.E. M.Econ.St. Research Centre for Economic-LIPI for

assistance with the statistical calculations. The author's thank also go to Nandang, B.Sc. for assistance in gaining access to the Computational Centre at the Research Centre for Chemistry–LIPI, and to Hery Kresnadi, M.Eng., for assistance in providing the valuable computational software.

REFFERENCES

1. Butina, D., and Gola, J.M.R., 2003, *J. Chem. Inf. Comput. Sci.,* 43, 837-841
2. Cal, K., 2006, *Yakugaku Zasshi*, 126, 307-309.
3. Loftsson, T., and Hreinsdóttir, D., 2006, *AAPS Pharm.Sci.Tech*. 7, E1-E4.
4. Johnson, S.R., and Zheng, W., 2006, *The AAPS Pharm.Sci.Tech.,* 8, E27-E40.
5. Wang, Y.L., Hu, L.D., and Wu, L.Y., 2006, *Int. J. Mol. Sci.,* 7, 47-58.
6. Cheng, A.J., and Merz, K.M.J., 2003, *J. Med. Chem.,* 46, 3572-3580.
7. Erös, D., Kéria, G., Kövesdi, I., Szántai-Kis, C., Mészáros, G., and Örfi, L., 2004, *Mini-Rev. Med. Chem.,* 4, 167-177.
8. Delaney, J.S., 2005, *Drug Discov. Today*, 10, 289-295.
9. Ghasemi, J., and Saaidpour, S., 2007, *Chem. Pharm. Bull.,* 55, 669-674.
10. Huuskonen, J., Salo, M., and Taskinen, J., 1997, *J. Pharm. Sci.,* 86, 450-454.
11. Huuskonen, J., Salo, M., and Taskinen, J., 1998, *J. Chem. Inf. Comput. Sci.,* 38, 450-456.
12. William, L.J., and Erin, M.D., 2000, *Med. Chem. Lett.*, 10, 1155-1158.
13. William L. J., and Erin, M.D., 2002, *Adv. Drug Del. Rev.,* 54, 355-366.
14. Huuskonen, J., 2000, *J. Chem. Inf. Comput. Sci.,* 40, 773-777.
15. Catana, C., Gao, H., Orrenius, C., and Stouten, P. F. W., 2005, *J. Chem. Inf. Model,* 45, 170-176.
16. Delgado, E.J., 2002, *Fluid. Phase. Equilibr.,* 199, 101-107.
17. Hypercube, I. 2002, *HyperChem Release 7.0 for Windows*, Hypercube. Inc., Gainesville, Florida, USA.
18. Telete, Srl., 2006, *DRAGON for Windows (Software for Molecular Descriptor Calculations) Version 5.4,* Talete, Srl., Milano, Italy, http://www.telete.mi.it.
19. Todeschini, R., 2007, *Milano Chemometric and QSAR Research Group*, University of Milano, Milano Italy, http://michem.disat.unimib.it/chm/index.htm. Accessed on 30 September 2007.
20. Gramatica, P., 2006, *QSAR & Combinatorial Sciences*, 25, 327-332.
21. Todeschini, R., and Gramatica, P., 2006, *Quantitative Structure-Activity Relationships,* 16, 113 – 119.
22. Todeschini, R., Bettiol, C., Giurin, G., Gramatica, P., Miana, P., and Argese, E., 1996, *Chemosphere*, 33, 71-79.
23. Chiorboli, C., Gramatica, P., Piazza, R., Pino, A., and Todeschini, R., 1997, *SAR and QSAR in Environmental Research*, 7, 133-150.
24. Todeschini, R., Vighi, M., Finizio, A., and Gramatica, P, 1997, *SAR and QSAR in Environmental Research*, 7, 173 – 193.
25. Foundation, A.S., 2004, SPSS Release 12.0 for Windows, SPSS, Inc., Chicago, Illinois, USA, http://www.spss.com/.
26. Hawkins, D.M., Basak, S.C., and Mills, D., 2003, *J. Chem. Inf. Comput. Sci*., 43, 579-586.
27. Kohavi, R., 1995., A study of cross-validation and bootstrap for accuracy estimation and model selection. *In Proceedings of the Fourteenth International Conference on Artificial Intelligence (IJCAI), San Mateo*, 1137-1143.
28. Wold, S., 1991, *Quant. Struct.-Act. Relat.*, 10, 191-193.
29. Ribeiro, F.A.D., and Ferreira, M.M.C., 2003, *Journal of Molecular Structure: THEOCHEM.*, 663, 109-126.