

Prediction of Aqueous Solubility of Drug-Like Compounds by Using an Artificial Neural Network and Least-Squares Support Vector Machine

Mohammad Hossein Fatemi,* Afsane Heidari, and Mehdi Ghorbanzade

Chemometrics Laboratory, Faculty of Chemistry, University of Mazandaran, Babolsar 47416-95447, Iran

Received March 15, 2010; E-mail: mhfatemi@umz.ac.ir

In this work the aqueous solubilities of 145 drug-like compounds were predicted from their theoretical derived molecular descriptors. Descriptors which were selected by stepwise multiple subset selection methods are; 1st-order solvation connectivity index, average span R, overall hydrogen bond basicity, and percent of hydrophilic surface area. These descriptors can encode features of molecules which are effected on dispersion, hydrophobic and steric interactions between solute and solvent molecules. To develop quantitative structure–activity relationship (QSAR) models, the methods of multiple linear regressions, least-squares support vector machine, and artificial neural network (ANN) were used by applying the selected descriptors as their inputs. The obtained statistical parameters of these models revealed that ANN model was superior to other methods. The standard error (SE), average error (AE), and average absolute error (AAE) for ANN model are: $SE = 0.714$, $AE = -0.178$, and $AAE = 0.546$, while these values for internal test set are: $SE = 0.830$, $AE = -0.056$, and $AAE = 0.630$ and for external test set are: $SE = 0.762$, $AE = -0.431$, and $AAE = 0.626$, respectively. Moreover the leave-many-out cross validation test was used to further investigate the prediction power and robustness of model, which lead to $R_{L10O}{}^2 = 0.816$ and $SPRESS = 0.32$ for ANN model, which revealed the reliability of this model.

Aqueous solubility is an important property of drugs which will be administrated orally or by injection,[1] due to this fact that oral drugs must be absorbed through the gastrointestinal tract and remain in solution to reach the intended therapeutic target and also, injection drugs must be sufficiently water soluble to transport by blood and lymph.[2,3] Aqueous solubility influences in terms of a drug candidate's absorption, distribution, metabolism, and excretion (ADME). Modeling of ADME properties of drugs is a promising method for recognition of poor drug candidates at the earliest stages of drug design.[4] Aqueous solubility relates to A and D terms in ADME process. The rate of passive drug transport across a biological membrane (the main pathway for drug absorption and distribution) depends on the membrane permeability and concentration gradient, which these values were affected by drug solubility.[5] Moreover aqueous solubility influence on M and E terms, due to this fact that compounds with higher solubilities are more easily metabolized and eliminated from the organism, thus leading to lower probability of adverse effect and bioaccumulation.[6] The aqueous solubility of a solute is affected by many factors that include; the size and shape of the molecule, the polarity and hydrophobicity of the molecule, and the ability of its functional groups to participate in intra- and intermolecular hydrogen bonding.[7] Poor aqueous solubility is caused by two main factors: 1) high lipophilicity and 2) strong intermolecular interactions in a solid structure. Generally, a drug with poor aqueous solubility and membrane permeability is a problematic candidate and needs careful formulation work. Experimentally obtaining high quality solubility data (even though higher throughput assays are used) remains a relatively expensive and time-consuming process. Therefore the development of theoretical models to predict the aqueous solubility of drug candidates from their chemical structures has attracted considerable attention.[5] Numerous different theoretical methods for the prediction of aqueous solubility have been developed. In the following, we will introduce some of these methods. 1) Models based on experimental measurements of some physicochemical parameters such as partition coefficient, melting point, boiling point, etc.,[8–11] 2) models based on amended linear free energy relationship (LFER) approaches,[12] 3) group contribution method (GCM),[13–15] and 4) models based on quantitative structure–property relationship (QSPR) approaches. In QSPR study, a mathematical model is developed which relates the molecular structural descriptors of a set of compounds to their physical properties such as aqueous solubility. These descriptors encode important information about structural features of interested compounds numerically. There are some reports about QSPR modeling of aqueous solubility of drugs. In a pioneering work, Huuskonen et al. established a multiple linear regression (MLR) model to predict the aqueous solubility of 191 drug-like compounds. Their 5-parameters model has the statistics of square correlation coefficient of $R^2 = 0.87$ and standard error of $SE = 0.51$. They examined the predictability of this model to a set of drug-like compounds and a set of agrochemicals which yields the statistics of $R^2 = 0.80$ and $R^2 = 0.88$ for these two independent test sets, respectively.[16] In a recent work, Cheng et al. reported a genetic algorithm–multiple linear regression model for the prediction of aqueous solubility of some organic compounds by using topological descriptors, that resulted statistics were; $R^2 = 0.84$ and root-mean-square error (RMSE) of 0.87.[17] Tetko et al. developed an artificial neural network (ANN) model for estimation of aqueous solubility of organic compounds by using 33 E-state

Published on the web October 21, 2010; doi:10.1246/bcsj.20100074

indices as ANN's input that resulted the statistics were; $R^2 = 0.91$ and $RMSE = 0.62$.[15] Moreover Bergström et al. developed a QSPR model for prediction of aqueous solubility of 85 drug-like compounds from their 2D and 3D descriptors. Their QSPR model yield the statistics of $R^2 = 0.80$ and $RMSE = 0.90$ for training set and $R^2 = 0.89$ and $RMSE = 0.83$ for validation set.[18]

Also a QSPR model for prediction of aqueous solubility of 145 drug-like compounds was developed by Duchowicz et al.[6] They obtained this model by using three parameter multiple linear regression which are; the 1st-order solvation connectivity index, radial distribution function-6.0/unweighted, and Moriguchi octanol–water partition coefficient. The statistics of their model is $R^2 = 0.759$ and $SE = 0.903$ for training set and $R^2 = 0.719$ and $SE = 0.899$ for validation set. In the present work we try to improve this model by using nonlinear feature mapping techniques and investigate some molecular descriptors which are important in determining aqueous solubility of drug-like compounds.

Experimental

Data Set. The experimental aqueous solubility of 145 structurally diverse drug-like organic compounds measured at 298 K and expressed as $(sol) = \text{solubility}/\text{mg mL}^{-1}$ in logarithm units are taken from Ref. 6 and are shown in Table 1. These values ranged from −6 to 3.352 for etofenprox and acetamide, respectively. Figure 1 represents the distribution of the aqueous solubilities of data set. It is obvious that the experimental aqueous solubilities are normally distributed over more than four logarithmic units. These compounds were considered as drug-like compounds according to Lipinski-rule criteria,[19] Veber et al. rule,[20] and rules for evaluating drug-likeness which were extracted from several recent publications.[6] Compounds in the data set were sorted according to their solubility values and internal and external test sets were selected from this set with desirable distance from one another (*y*-ranking method). The training set consist of 87 molecules and was used for model generation, and the internal test set has 29 compounds and used for preventing over training and the external test set has 29 members, which was used to evaluate predictability of the model. The diversity of data set splitting was examined by diversity analysis. In developing the ANN and SVM models, usually a training set was used in optimization of model parameters and training, while during this process the internal test set was used to monitor the extent of model development and prevention of over training. After the training the prediction power of model was evaluated on independent data that was not used during the training step, this data set was named external test set. In the case of MLR model internal and external test sets were considered as test set.

Descriptors Generation and Selection. The chemical structures of molecules were drawn by using Hyperchem package (Version 7)[21] and optimized by the AM1 semiempirical method. Dragon (V.3),[22] Codessa,[23] MOPAC (V.6),[24] and Pro Plus software[25] were used to calculate molecular descriptors by using the Hyperchem output files. Moreover LFER descriptors were obtained by ADME Boxes software.[26] This software calculates these parameters from fragment contributions. A preliminary approach to variable reduction consisting

Table 1. Data Set and Corresponding Observed MLR, LS-SVM, and ANN Predicted Values of log(*sol*)

No.	Chemical name	Exptl	Predicted		
		log(*sol*)	MLR	LS-SVM	ANN
Training set					
1	Acetamide	3.352	2.558	2.567	3.352
2	Gluconolactone	2.770	2.378	2.246	2.454
3	Arabinose	2.698	2.780	2.204	3.431
4	Hydroxyproline	2.557	2.078	1.854	2.543
5	Ascorbic acid	2.522	1.784	1.542	1.233
6	Glycine	2.396	2.547	2.431	3.150
7	Alanine	2.214	2.627	2.427	3.289
8	Hymexazol	1.929	1.077	1.542	1.893
9	Acrylonitrile	1.872	1.077	1.585	2.188
10	Hydroquinone	1.857	1.092	1.771	1.590
11	Methomyl	1.763	0.848	1.088	1.283
12	Guaifenesin	1.698	0.556	0.677	0.656
13	Phthalazine	1.698	0.569	0.903	0.701
14	Histidine	1.658	1.779	1.662	1.774
15	Ethohexadiol	1.623	1.134	0.843	1.188
16	Aniline	1.556	1.534	1.370	1.777
17	Hexazinone	1.519	−0.238	0.939	1.348
18	Adipic acid	1.414	0.819	1.121	1.414
19	Hydroxyphenamate	1.397	0.621	0.497	0.706
20	Azidamfenicol	1.301	−0.155	0.098	0.328
21	Isoflurophate	1.187	2.033	1.471	2.343
22	Picric acid	1.103	−0.176	−0.654	0.885
23	Imazapyr	1.053	0.016	0.737	0.410
24	Caproic acid	1.012	0.938	1.004	1.010
25	Glutamic acid	0.933	1.766	1.366	1.780
26	Barbital	0.873	1.303	1.130	1.647
27	3,4-Dinitrobenzoic acid	0.826	−0.650	−0.906	−0.383
28	Amicarbalide	0.700	−0.678	0.302	−0.240
29	Azintamide	0.699	−1.298	−0.322	0.843
30	*p*-Hydroxybenzoic acid	0.699	0.784	1.074	1.340
31	Ganciclovir	0.633	1.455	0.204	1.024
32	Acetamiprid	0.623	−0.001	−0.132	0.085
33	Carmustine	0.602	−0.014	−0.078	0.158
34	4-Amino-2-sulfobenzoic acid	0.477	0.856	1.154	0.808
35	Pirimicarb	0.431	0.021	0.539	0.687
36	Ethinamate	0.398	0.546	0.473	0.588
37	Cyanuric acid	0.301	1.848	0.554	1.251
38	Iridomyrmecin	0.301	0.802	0.301	0.365
39	Allobarbital	0.258	0.772	0.974	0.928
40	Imazethapyr	0.146	−0.542	0.163	0.030
41	Dimethenamid	0.079	−1.12	−0.583	−0.634
42	Dimethirimol	0.079	−0.257	−0.211	−0.028
43	*p*-Fluorobenzoic acid	0.079	0.529	0.808	0.892
44	Khellin	0.017	−0.532	−0.996	−0.796
45	1,6-Cleve's acid	0.000	−0.045	0.155	0.093
46	Cyclizine	0.000	−0.813	−0.645	−0.499
47	Acetazolamide	−0.009	0.332	0.207	0.025
48	2-Ethyl-1-hexanol	−0.056	0.322	0.745	−0.048
49	Dehydroacetic acid	−0.161	1.228	1.339	1.563
50	Ancymidol	−0.187	−0.780	−0.339	0.002
51	ANTU	−0.222	0.016	−0.431	−0.054
52	Picloram	−0.367	−0.345	−0.453	−0.128
53	EPTC	−0.426	0.202	0.263	0.376

Continued on next page.

Continued.

No.	Chemical name	Exptl	Predicted		
		log(*sol*)	MLR	LS-SVM	ANN
54	Carbofuran	−0.495	−0.392	−0.783	−0.380
55	Carisoprodol	−0.523	1.231	0.948	0.758
56	Heptabarbital	−0.602	0.031	0.067	0.742
57	Alochlor	−0.620	−0.812	−0.476	−0.599
58	Furametpyr	−0.648	−1.546	−1.453	−1.349
59	Cyanazine	−0.767	0.075	−0.413	−0.052
60	Cumic acid	−0.821	−0.319	0.240	−0.599
61	Cyproconazole	−0.854	−1.082	−1.171	−1.055
62	Imazaquin	−1.045	−1.072	−0.475	−0.492
63	Linuron	−1.124	−0.448	−0.583	−0.266
64	2,6-Dibromoquinone-4-chlorimide	−1.230	−0.922	−0.818	−0.823
65	Ethofumesate	−1.301	−0.932	−0.663	−0.291
66	2,4-DB	−1.337	−1.535	−1.315	−1.500
67	Furazolidone	−1.397	−0.389	−0.767	0.025
68	Carfentrazone-ethyl	−1.657	−2.02	−2.072	−1.602
69	Dichlobenil	−1.673	−0.871	−1.068	−1.016
70	Dimethomorph	−1.728	−2.593	−2.464	−2.613
71	Haloperidol	−1.853	−2.202	−2.007	−2.177
72	Barban	−1.958	−1.438	−1.611	−0.911
73	Flufenamic acid	−2.041	−0.614	−1.194	−1.234
74	Acibenzolar-*S*-methyl	−2.113	−0.462	−0.700	−0.331
75	Lenacil	−2.221	−0.485	−0.938	−0.407
76	Diniconazole	−2.397	−1.722	−2.755	−1.820
77	Flumioxazin	−2.747	−1.369	−1.438	−1.232
78	Folic acid	−2.795	−2.120	−2.468	−2.596
79	Equilin	−2.850	−2.051	−1.729	−2.709
80	Amitraz	−3.000	−2.589	−2.576	−3.215
81	Diclofop-methyl	−3.096	−3.718	−3.583	−3.264
82	Carbosulfan	−3.522	−2.160	−2.624	−3.030
83	Dichlofenthion	−3.610	−2.350	−2.797	−3.343
84	Fenbuconazole	−3.699	−2.662	−3.910	−3.395
85	Acequinocyl	−4.173	−4.012	−4.081	−3.954
86	Etofenprox	−6.000	−5.063	−4.778	−6.074
87	Fluspirilene	−2.000	−3.682	−2.149	−2.327
	Internal test set				
88	Carnosine	1.914	1.247	0.968	0.786
89	Cycloleucine	1.698	1.921	1.444	2.226
90	Isoleucine	1.536	−0.274	0.550	1.088
91	Glyphosate	1.079	2.028	0.881	0.844
92	Isophorone	1.079	1.814	1.457	1.652
93	Cystine	0.951	1.067	0.401	0.350
94	Aspartic acid	0.912	2.207	1.868	2.550
95	Asulam	0.699	0.107	0.266	0.214
96	Idoxuridine	0.301	0.528	0.547	0.510
97	2,4,5-Trichlorophenol	0.079	−2.052	−1.205	0.236
98	Ethoprop	−0.125	−0.885	−0.694	−0.086
99	Badische acid	−0.225	0.122	0.192	0.182
100	Phthalimide	−0.444	1.228	1.339	1.550
101	Benzidine	−0.495	0.101	−0.123	0.168
102	Acetochlor	−0.652	−1.077	−1.140	−1.240
103	Fludrocortisone	−0.854	−1.147	−2.497	−0.715
104	Dexamethasone	−1.051	−1.291	−2.577	−0.746
105	Capric acid	−1.209	−0.698	0.130	−1.576
106	2-Cyclohexyl-4,6-dinitrophenol	−1.823	−1.387	−1.543	−2.295
107	Difenoconazole	−1.823	−3.312	−3.628	−3.751
108	Diallate	−1.853	−1.537	−1.123	−1.671
109	Cyprodinil	−1.886	−0.948	−1.587	−1.340
110	Ketanserin	−2.000	−1.996	−2.416	−2.474
111	Fenbufen	−2.656	−1.593	−2.171	−1.366
112	Bifenox	−3.397	−2.981	−3.611	−3.607
113	Bifenthrin	−4.000	−4.197	−3.546	−3.788
114	Acrylamide	2.806	1.799	2.342	2.622
115	Hydrocortisone	−0.495	−2.456	−1.532	0.578
116	Dimorpholamine	2.698	1.217	1.800	1.502
	External test set				
117	α-Acetylbutyrolactone	2.301	1.545	1.642	2.052
118	Acetylacetone	2.221	1.147	2.121	2.173
119	Crotonic acid	1.934	1.389	1.633	2.352
120	Allidochlor	1.294	0.343	0.740	0.581
121	Phthalic acid	0.846	0.6	0.727	0.960
122	Fumaric acid	0.845	1.228	1.276	1.646
123	Acetanilide	0.806	0.893	1.177	1.468
124	Aldicarb	0.780	0.561	0.507	0.803
125	PABA	0.769	1.129	1.143	2.044
126	Cyclobarbital	0.204	0.138	0.435	0.886
127	Thionazin	0.057	−0.179	0.110	0.022
128	Adenine	0.013	1.584	0.832	1.109
129	Cymoxanil	−0.051	0.734	0.740	0.443
130	Dicamba	−0.080	−0.261	−0.026	−0.066
131	Amobarbital	−0.220	0.407	0.634	1.173
132	Digallic acid	−0.301	−0.485	−0.497	0.189
133	Hydroflumethiazide	−0.523	0.354	0.633	0.613
134	Bendiocarb	−0.585	0.352	−0.179	0.097
135	Biotin	−0.658	−0.111	−0.154	−0.037
136	Ethirimol	−0.699	−0.205	−0.429	−0.089
137	Carboxin	−0.701	−0.121	−0.635	−0.333
138	Ethoprop	−0.125	−1.412	−1.341	−0.717
139	Flufenacet	−1.252	−2.927	−2.218	−2.951
140	Azoxystrobin	−2.000	−0.846	−1.191	−1.312
141	Fenoxaprop-ethyl	−3.046	−3.133	−2.366	−2.643
142	Fluthiacet-methyl	−3.070	−2.126	−1.666	−1.632
143	Aminopromazine	−3.239	−1.54	−1.781	−2.174
144	Aconitic acid	2.698	1.396	1.263	2.019
145	Isoniazid	2.146	1.037	1.255	1.995

Continued on next column.

of the elimination of all the variables that take the same value for all objects in the data set. Also near-constant variables, i.e., variables that assume the same value except in one or very few cases, would be excluded. Pairs of variables with a correlation coefficient greater than 0.90 were classified as intercorrelated, and only one of them was considered in developing the model. Then the method of stepwise multiple linear regression was performed to select the most relevant descriptors. In order to determine the optimum number of descriptors, the correlation coefficient of cross-validation (R_{cv}^2) value for each subset was calculated and were plotted versus the number of descriptors in the models for the 1–13-parameter models (break point

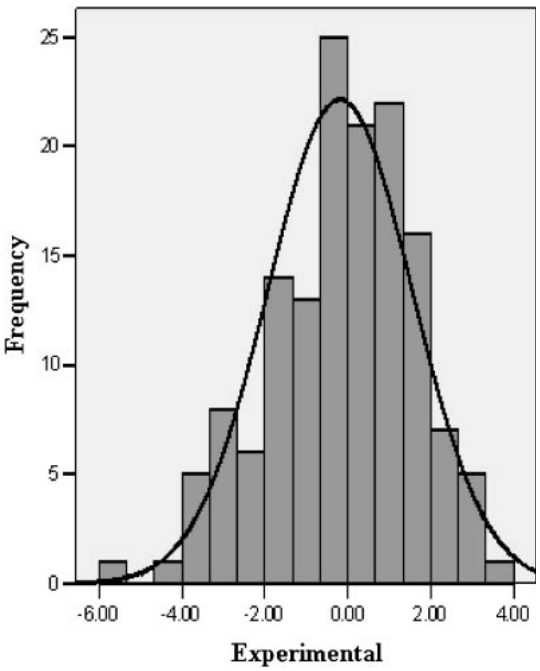


Figure 1. Distribution of the experimental log(*sol*) values of data set.

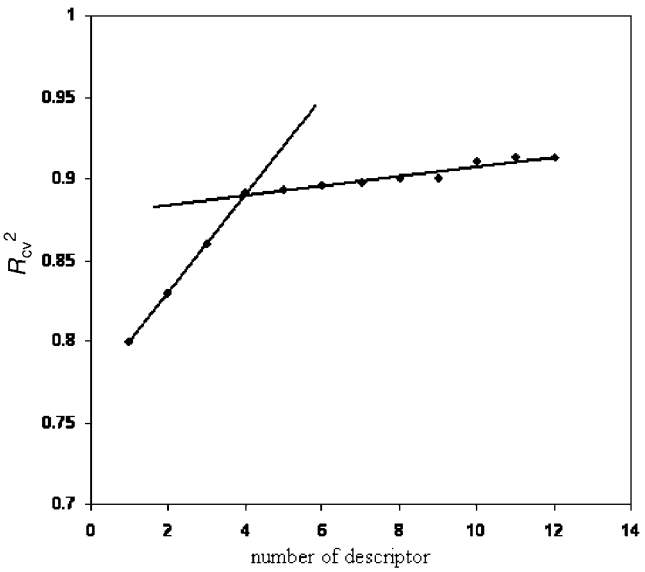


Figure 2. The plot of R_{cv}^2 against number of descriptors in the MLR model.

procedure). Figure 2 shows that addition of up to four descriptors to the model does not provide any significant improvement in the R_{cv}^2 values of obtained model. So we selected four descriptors as the optimum number of independent variables in developing quantitative structure–activity relationship (QSAR) models. These parameters are; the 1st-order solvation connectivity index (*XIsol*), average span *R* (*R.SPAM*), overall hydrogen bond basicity (*B*), and percent of hydrophilic surface area (p*HSA*). These descriptors would be used as inputs for developing of ANN, LS-SVM, and MLR models. Table 2 shows the correlation matrix among these four descriptors. As it can be seen from this table, there is no high correlation among the selected descriptors.

Table 2. Correlation Matrix for Descriptors Applying in This Work

	XIsol	*B*	*R.SPAM*	p*HSA*
XIsol	1	0.550	−0.565	0.279
B		1	−0.425	−0.355
R.SPAM			1	−0.245
p*HSA*				1

Artificial Neural Network. ANN represents a promising modeling technique especially in nonlinear modeling, which is frequently encountered in QSAR methods. Generally, each network is built from several layers: one input layer, one or more hidden layers, and one output layer. The node in each layer is connected to the nodes of the next layer by weights. During training these weights and biases are iteratively adjusted to minimize the network errors. Then the trained network is used as an analytical tool to predict the activity of a new set of input data. There are several training algorithms, one of which is Newton's method. The basic formula of Newton's algorithm is expressed as the following equation:

$$\boldsymbol{X}_{k+1} = \boldsymbol{X}_k - \boldsymbol{A}_k^{-1} g_k \quad (1)$$

where $\boldsymbol{X}_k$ is a vector of current weights and biases, g_k is the current gradient, and $\boldsymbol{A}$ is the Hessian matrix (second derivatives) of the performance function at the current values of the weights and biases. Newton's method often converges fast but unfortunately, it is complex and expensive to compute the Hessian matrix for feed-forward neural networks. Levenberg–Marquardt algorithm[27] is a fast algorithm which was designed to increase training speed without having to compute the Hessian matrix that uses standard numerical optimization techniques. When the performance function has the form of a sum of squares errors (as is typical in training feed forward networks), then the Hessian matrix can be approximated as:

$$\boldsymbol{H} = \boldsymbol{J}^T \boldsymbol{J} \quad (2)$$

and the gradient can be computed as:

$$g = \boldsymbol{J}_e^T \quad (3)$$

where $\boldsymbol{J}$ is the Jacobian matrix that contains first derivatives of the network errors with respect to the weights and biases and $\boldsymbol{J}_e$ indicates the network errors. The Jacobian matrix can be computed through a standard back propagation technique.

The Levenberg–Marquardt algorithm uses this approximation to the Hessian matrix in the following Newton-like update:

$$\boldsymbol{X}_{k+1} = \boldsymbol{X}_k - [\boldsymbol{J}^T \boldsymbol{J} + \mu]^{-1} \boldsymbol{J}_e^T \quad (4)$$

The parameter μ multiplied by μ_dec whenever the performance function is reduced by a step. It is multiplied by μ_inc whenever a step would increase the performance function. The value of μ_dec was varied between 0–1 while μ_inc would be greater than 1. Therefore μ is decreased after each successful step (reduction in performance function) and is increased only when a tentative step would increase the performance function. In this way, the performance function is always reduced at each training iteration.[28]

Least-Squares Support Vector Machine. Support vector machine, developed by Vapnik and Cortes[29] as a novel type of machine learning method, is gaining popularity due to many attractive features and promising empirical performance. Later, computational calculations of SVM were simplified by Suykens et al.[30] with the implementation of a least-squares version for support vector machine (LS-SVM). LS-SVM has the capability of dealing with linear and nonlinear multivariate calibration and resolving these problems in a relatively rapid way. This requires solving a set of linear equations, instead of the quadratic programming used in classical SVM. The details of LS-SVM algorithm could be found in the references of.[31–34] The LS-SVM model can be expressed as

$$y = \sum_{i=1}^{N} \alpha_i K(x_i, x) + b \tag{5}$$

$$\alpha_i = 2\gamma e_i \tag{6}$$

In the above equations, $K(x_i, x)$ is the kernel function, x_i is the input vector, α_i is the Lagrange multipliers called support value, b is the bias term and the γ (gam) parameter is the regularization parameter for determining the trade-off between the fitting error minimization and smoothness of the estimated function which has to be optimized by the user. A kernel function (in the form of a polynomial, Gaussian, or sigmoidal function) is used to map the input vectors into a higher dimensional feature space.[35] The most general kernel function is radial basis function (RBF):

$$K(x_i, x) = \exp(-\|x_i - x_j\|^2/2\sigma^2) \tag{7}$$

where σ^2 is the width of the RBF function. Generalization capability of SVM depends on the proper selection of parameters. The selection of the kernel function and corresponding parameters is very important because they define the distribution of the training set samples in the high dimensional feature space.[35] In this work, we established LS-SVM by RBF kernel function to estimate the aqueous solubilities of 145 drug-like compounds. All calculation in this work was carried out by using Matlab (v.7.6.) and the calculation was performed on a 1.8 GHz Intel Pentium (IV) with 0.5 GB RAM under windows XP.

Results and Discussion

In this work the method of stepwise multiple linear regression was used to the selection of the most relevant descriptors, then MLR, LS-SVM, and ANN methods were used as feature mapping techniques to build linear and nonlinear QSAR models to predict the aqueous solubility of drug-like compounds. The data set and corresponding observed and predicted values of the log(*sol*) of all molecules studied in this work are shown in Table 1. Rational division of the experimental data set into training and test sets are an important part in the development and validation of reliable QSAR model. In this study, diversity analysis was performed to make sure that the structures of the training and test cases can represent those of the whole ones. In this way, the mean distances of one sample to the remaining ones ($\bar{d}_i$) were computed from descriptor space matrix as follows:

$$\bar{d}_i = \frac{\sum_{j=1}^{n} d_{ij}}{n-1} \quad i = 1, 2, \ldots, n \tag{8}$$

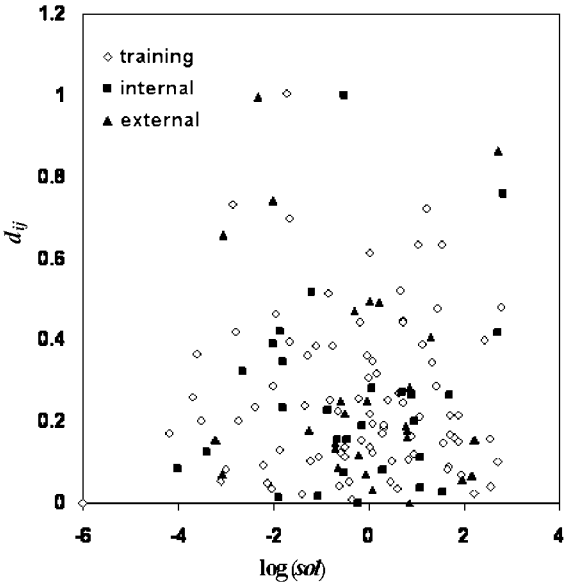


Figure 3. The results of diversity test.

where d_{ij} is a distance score for two different compounds, which can be measured by the Euclidean distance norm based on the compound's descriptors (x_{ik} and x_{jk}):

$$d_{ij} = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2} \tag{9}$$

Then the mean distances were normalized within the interval of zero to one and the resulting values were plotted against log(*sol*) (Figure 3). As can be seen from this figure, the structures of the compounds are diverse in all sets and the training set with a broad representation of the chemistry space was adequate to ensure the model's stability and the diversity of test sets can prove the predictive capability of the model.

Linear Modeling. The stepwise MLR technique was performed on the molecules of the training set by SPSS (V.13). By using break point procedure 4-parameter MLR model can be considered as the best linear model (Figure 2). The obtained MLR model has the following specifications:

$$\begin{aligned} \log(sol) = {} & 7.251 - 0.548(\pm 0.053) XIsol \\ & + 0.957(\pm 0.315) B - 7.951(\pm 2.34) R.SPAM \\ & - 29.86(\pm 9.97)(\mathrm{p}HSA)^{-1} \end{aligned} \tag{10}$$

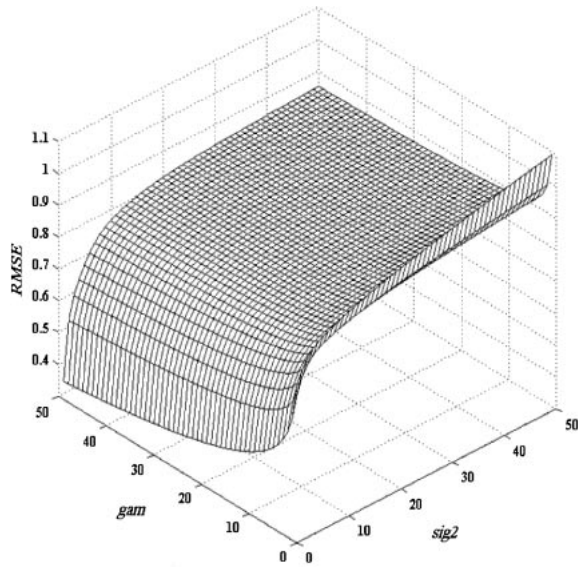
$$n = 87,\ R = 0.874,\ SE = 0.88,\ F = 276.25$$

The calculated value of log(*sol*) for training and test sets by this model is shown in Table 1. The standard error (*SE*), average error (*AE*), and average absolute error (*AAE*) of this calculation for training set are: $SE = 0.874$, $AE = 0.003$, and $AAE = 0.724$, while these values for test set are: $SE = 0.936$, $AE = 0.036$, and $AAE = 0.756$. Other statistical parameters of this calculation were shown in Table 3.

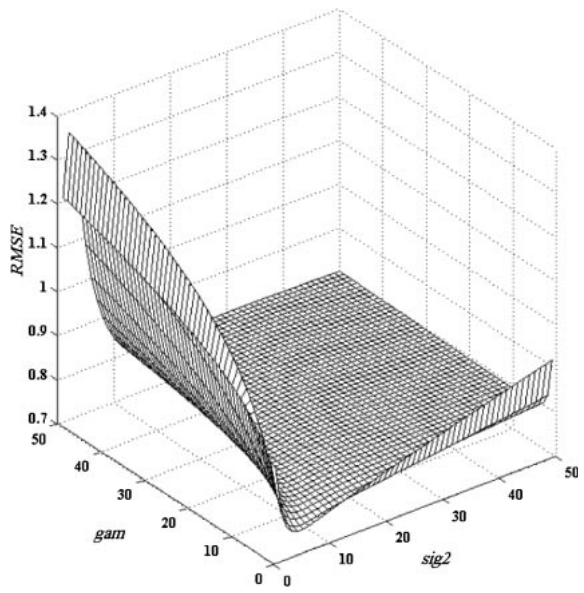
Nonlinear Models. LS-SVM: In the model development using LS-SVM technique with RBF kernel function, the values of γ and σ^2 parameters were a manageable task as a two-dimensional problem. We used training and internal test sets to obtain best values of γ and σ^2 parameters by considering the minimum value of *RMSE*. The obtained response surface for LS-SVM by utilization of different γ and σ^2 values are

Table 3. Statistical Parameters Obtained Using MLR, LS-SVM, and ANN Models

Model	Training		Internal test		External test		R_{cv}^2	*SPRESS*
	R	*RMSE*	*R*	*RMSE*	*R*	*RMSE*		
MLR	0.874	0.869	—	—	0.844	0.919	0.695	0.538
LS-SVM	0.920	0.702	0.864	0.883	0.920	0.726	0.698	0.532
ANN	0.925	0.709	0.890	0.815	0.923	0.748	0.816	0.320
Previous work	0.871	0.898			0.848	0.889	0.720	



(a)



(b)

Figure 4. Response surface for LS-SVM model for training set (a) and internal test set (b).

presented in Figures 4a and 4b for training and internal test set, respectively. By inspection to these figures, it was concluded that the optimum values of γ and σ^2 were 6 and 3, respectively. Then the developed LS-SVM model was examined by external test set to evaluate its predictability. The calculated values of log(*sol*) for training, external and internal test sets were shown in Table 1. The statistics of this calculation for training set are: $SE = 0.707$, $AE = 0.000$, and $AAE = 0.569$, while these values for internal test set are: $SE = 0899$, $AE = 0.205$, and $AAE = 0.739$ and for external test set are: $SE = 0.709$, $AE = -0.214$, and $AAE = 0.547$. Other statistical parameters of these calculations were shown in Table 3.

ANN: An artificial neural network with Levenberg–Marquardt training algorithm was used to construct the other QSAR model. A three-layer network with a sigmoid transfer function was designed which selected four descriptors were used as its inputs and log(*sol*) values as outputs. The ANN's input and targets were centered to their means and scaled by dividing to their standard deviation. Then the network parameters including μ, μ_dec, μ_inc, and number of neurons in the hidden layer were optimized which were found to be 0.001, 0.1, 10, and 6, respectively. Then the obtained 4-6-1 network was used to predict the log(*sol*) for external test set as well as training and internal test sets. The predicted values of log(*sol*) for training, internal and external test sets were shown in Table 1. The statistics of this calculation for training set are: $SE = 0.714$, $AE = -0.178$, and $AAE = 0.546$, while these values for internal test set are: $SE = 0.830$, $AE = -0.056$, and $AAE = 0.630$ and for external test set are: $SE = 0.762$, $AE = -0.431$, and $AAE = 0.626$. Other statistical parameters of these calculations were shown in Table 3.

Model Validation. Leave one out cross validation (LOO) and leave many out cross validation (LMO) tests are two methods which frequently used to validate QSPR models. The outcomes from these procedures are a cross validated correlation coefficient (R_{cv}^2) and standardized predicted error sum of squares (*SPRESS*), which are calculated according to the following equations:

$$R_{cv}^2 = 1 - \frac{\sum (y_0 - y_i)^2}{\sum (y_i - \bar{y})^2} \tag{11}$$

$$SPRESS = \sqrt{\frac{\sum (y_0 - y_i)^2}{n - k - 1}} \tag{12}$$

In the above expression, $\bar{y}$ is the mean of the experimental values, n is number of observations, and k is the number of descriptors in the model. The R_{cv}^2 values are the proportion of variability in data set that is accounted by a statistical model and *SPRESS* is criteria of deviation from observed data. We employed leave-one-out cross validation for MLR model and leave-ten-out cross validation for MLR, LS-SVM, and ANN models. The obtained R_{L00}^2, R_{L10O}^2, and *SPRESS* values of these tests were shown in Table 3. The values of R_{L10O}^2 and *SPRESS* for ANN model are 0.816 and 0.32, respectively, while these values are $R_{L10O}^2 = 0.698$ and $SPRESS = 0.532$ for

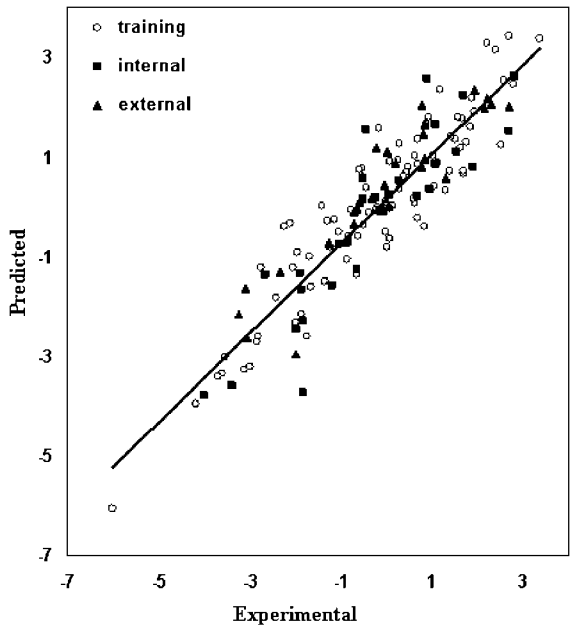


Figure 5. The plot of the ANN calculated log(*sol*) against the experimental values.

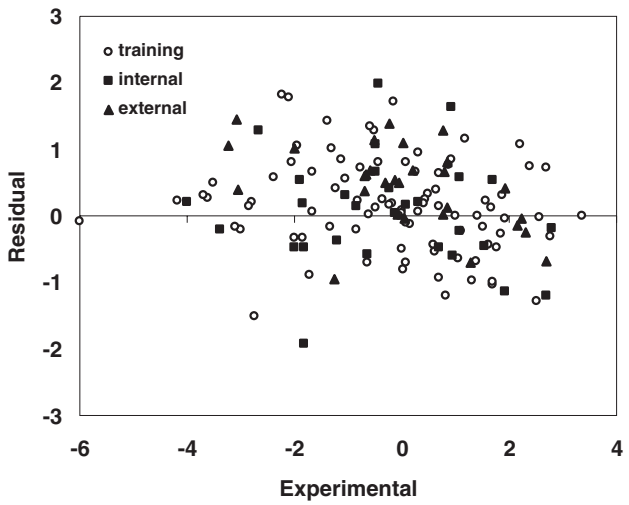


Figure 6. Plot of the ANN residuals against experimental values of log(*sol*).

LS-SVM model and $R_{L10O}^2 = 0.695$ and $SPRESS = 0.538$ for MLR model, respectively. Inspection to these values and other statistical parameters in Table 3, show the superiority of ANN over MLR and LS-SVM models. It was worth nothing that the statistical parameters of developed nonlinear models are better than those obtained by Duchowicz et al.[6]

Plot of the ANN calculated versus the experimental values of log(*sol*) for whole molecules in the data set are shown in Figure 5 which shows the good correlation between values. The residual of the ANN calculated values of the log(*sol*) are plotted against their experimental values in Figure 6. The propagation of the residuals on both sides of zero line shows that no systematic error exists in the development of the neural network.

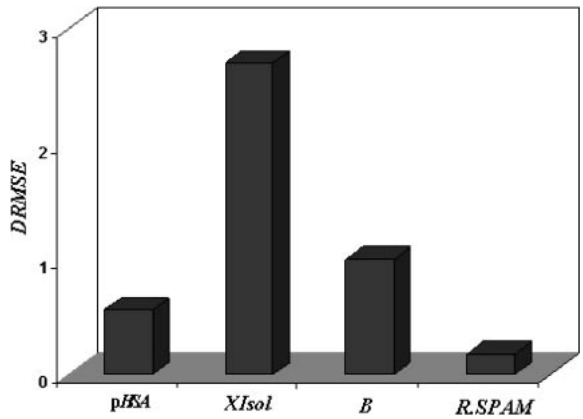


Figure 7. The results of sensitivity analysis on the ANN model.

Descriptors Interpretation. In order to determine the relative importance of each variable in the ANN model, the sensitivity analysis approach was applied. This method is performed based on the sequential removal of variables by zeroing the specific connections weight for that specific input variable in the first layer of the ANN. For each sequentially zeroed input variable, root-mean-square error of prediction (*RMSEP*) as the prediction error of network was calculated. Generally *RMSEP* value increases in this way. Then, differences between *RMSEP* and root-mean-square error of established ANN was calculated and shown as *DRMSE*. Each variable which causes greater value of *DRMSE* is more important. The calculated values of *DRMSE* are plotted in Figure 7. As it can be seen from this figure, the order of importance of input descriptors in ANN model is; *XIsol* > *B* > p*HSA* > *R.SPAM*. Based on this result, it was concluded that the most important descriptor is the 1st-order solvation connectivity index which was a topological descriptor and calculated by Dragon package. This descriptor was calculated as follow:[36]

$$ {}^1\chi^s = \frac{1}{4} \cdot \sum_{b=1}^{B} \frac{(L_i \cdot L_j)_b}{(\delta_i \cdot \delta_j)_b^{1/2}} \tag{13} $$

where b runs over all the B bonds; L_i and L_j are the principal quantum numbers of the two vertices incident to the considered bond and characterize the atom size, and the parameters of δ_i and δ_j are the corresponding vertex degree. This descriptor was used to account the solvation entropy and can describe the dispersion interactions between solute and solvent.

Next descriptor is overall hydrogen bond basicity (*B*), which is a LFER descriptor and encodes the strength and number of H-bonds formed by lone pairs of solute acceptor groups with solvent donor groups. By increasing the values of this descriptor, the aqueous solubility of organic compounds increase due to increasing of hydrogen bonding. Third descriptor is percent of hydrophilic surface area that was calculated by molecular modeling Pro Plus software. The appearance of this descriptor in the model represents the role of solute hydrophobicity in its aqueous solubility. The last descriptor is the average span of *R* (*R.SPAM*), which was calculated by Dragon package. This geometric descriptor can account for the size of molecules which affect steric interaction

between solute and solvent. Therefore it was concluded that descriptors that appeared in this work can account dispersion, hydrophobic and steric interactions between solute and solvents, which were affected on aqueous solubilities of drug-like compounds.

Conclusion

In the present study, a linear (MLR) and two nonlinear feature mapping method (LS-SVM and ANN) were used to develop some QSAR models for prediction of aqueous solubilities some of 145 drug-like organic compounds. The obtained statistical parameters of these models revealed that ANN model was superior over other models, which showed that nonlinear modeling technique can successfully used to predict the aqueous solubilities of drug-like compounds. Descriptors that appeared in these models were electronic, geometrical, and topological descriptors that can encode features of solutes which were affected on their aqueous solubilities, including steric, dispersion, and hydrophobic interactions.

References

1 J. Wang, G. Krudy, T. Hou, W. Zhang, G. Holland, X. Xu, *J. Chem. Inf. Model.* **2007**, *47*, 1395.

2 J. I. Boullata, *Am. J. Nurs.* **2009**, *109*, Issue 10, 34.

3 J. Ghasemi, S. Saaidpour, *Chem. Pharm. Bull.* **2007**, *55*, 669.

4 D. S. Wishart, *Drugs R&D* **2007**, *8*, 349.

5 B. Faller, P. Ertl, *Adv. Drug Deliv. Rev.* **2007**, *59*, 533.

6 P. R. Duchowicz, A. Talevi, L. E. Bruno-Blanch, E. A. Castro, *Bioorg. Med. Chem.* **2008**, *16*, 7944.

7 J. Wang, T. Hou, X. Xu, *J. Chem. Inf. Model.* **2009**, *49*, 571.

8 S. H. Yalkowsky, S. Banerjee, *Aqueous Solubility Methods of Estimation for Organic Compounds*, Marcel Dekker, New York, **1992**, pp. 149–154.

9 C. Hansch, J. E. Quinlan, G. L. Lawrence, *J. Org. Chem.* **1968**, *33*, 347.

10 P. Isnard, S. Lambert, *Chemosphere* **1989**, *18*, 1837.

11 Y. Ran, Y. He, G. Yang, J. L. H. Johnson, S. H. Yalkowsky, *Chemosphere* **2002**, *48*, 487.

12 M. H. Abraham, J. Le, *J. Pharm. Sci.* **1999**, *88*, 868.

13 R. Kühne, R.-U. Ebert, F. Kleint, G. Schmidt, G. Schüürmann, *Chemosphere* **1995**, *30*, 2061.

14 G. Klopman, H. Zhu, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 439.

15 I. V. Tetko, V. Y. Tanchuk, T. N. Kasheva, A. E. P. Villa, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1488.

16 J. Huuskonen, D. J. Livingstone, D. T. Manallack, *SAR QSAR Environ. Res.* **2008**, *19*, 191.

17 A. Cheng, K. M. Merz, Jr., *J. Med. Chem.* **2003**, *46*, 3572.

18 C. A. S. Bergström, C. M. Wassvik, U. Norinder, K. Luthman, P. Artursson, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1477.

19 C. A. Lipinski, F. Lombardo, B. W. Dominy, P. J. Feeney, *Adv. Drug Deliv. Rev.* **2001**, *46*, 3.

20 D. F. Veber, S. R. Johnson, H.-Y. Cheng, B. R. Smith, K. W. Ward, K. D. Kopple, *J. Med. Chem.* **2002**, *45*, 2615.

21 *HyperChem Release 7.0 for Windows*, Hypercube Inc., **2002**.

22 http://www.disat.unimib.it/chem.

23 A. R. Katritzky, V. S. Lobanov, M. Karelson, Version 2.0, *Comprehensive Descriptors for Structural and Statistical Analysis*, Reference Manual, **1994**.

24 J. P. P. Stewart, Version 6.0, Quantum Chemistry Program Exchange, QCPE, No. 455, India University, **1989**.

25 http://www.chemsw.com/d_13071.htm.

26 http://www.ap-algorithms.com/absolv.htm.

27 D. W. Marquardt, *J. Soc. Ind. Appl. Math.* **1963**, *11*, 431.

28 MATLAB 7.4, http://www.mathworks.com/products/matlab/help.

29 C. Cortes, V. Vapnik, *Mach. Learn.* **1995**, *20*, 273.

30 J. A. K. Suykens, T. van Gestel, J. de Brabanter, B. de Moor, J. Vandewalle, *Least-Squares Support Vector Machines*, World Scientific, Singapore, **2002**.

31 M. F. Ferrão, S. C. Godoy, A. E. Gerbase, C. Mello, J. C. Furtado, C. L. Petzhold, R. J. Poppi, *Anal. Chim. Acta* **2007**, *595*, 114.

32 W. Cui, X. Yan, *Chemom. Intell. Lab. Syst.* **2009**, *98*, 130.

33 Q. Chen, J. Zhao, C. H. Fang, D. Wang, *Spectrochim. Acta, Part A* **2007**, *66*, 568.

34 J. A. K. Suykens, J. Vandewalle, *Neural Process. Lett.* **1999**, *9*, 293.

35 C. W. Yap, H. Li, Z. L. Ji, Y. Z. Chen, *Mini-Rev. Med. Chem.* **2007**, *7*, 1097.

36 R. Todeschini, V. Consonni, *Handbook of Molecular Descriptors*, Wiley-VCH, Weinheim, **2000**, pp. 88–89.