International Journal of Chemistry and Pharmaceutical Sciences

Journal Home Page: www.pharmaresearchlibrary.com/ijcps

Research Article **Open Access**

Inslico-Prediction of Drug Solubility in Aquous Media by Theoritical Descriptors and QSAR Method

Elham Baher*, Poneh Ebrahimi, and Naser Darzi Naftchali

Faculty of science, Department of chemistry, Golestan University, Gorgan, Iran.

A B S T R A C T

The most key factor in transition of drugs across the biological membrane is their solubility in water. In this study, a novel theoretical model was proposed for the prediction of drug solubility in the aqueous media (log 1/S) by using employing the molecular structure descriptors. The data set consists of 58 different drugs. Molecular descriptors were calculated and selected by genetic algorithm (GA) and stepwise-multiple linear regression (SMLR) methods. The selected descriptors with GA were: momentum of inertia, total molecular surface area, difference in CPSA, PPSA-3 atomic charge weighted PPSA, FPSA-3 fractional PPSA, count of H-donors sites, HA dependent HDSA-2/SQRT (TMSA) and kier shape index. The selected descriptors with SMLR were: FPSA-3, Kier shape index, randic index, XY shadow and count of H-donors sites. Then prediction of log 1/S as a criterion of drug delivery was accomplished by support vector machine (SVM) by using GA and SMLR selected descriptors, separately. By comparison of results obtained, it was concluded that the GA-SVM model was superior over other models. The statistical result presented a good generality and ability of SVM model in drug delivery property. The simplicity, reliability and high speed calculation are some advantages of the present work.

Keywords: log 1/*S*, support vector machine, genetic algorithm, quantitative structure-activity relationship

A R T I C L E I N F O

CONTENTS

Article History: Received 05 April 2016, **Accepted** 09 May 2016, **Available Online** 27 June 2016

***Corresponding Author**
Elham Baher
Faculty of science,
Department of Chemistry,
Golestan University, Gorgan, Iran.
Manuscript ID: IJCPS3011

PAPER-QR CODE

Citation: Elham Baher, *et al. In-slico*-Prediction of Drug Solubility in Aquous Media by Theoritical Descriptors and QSAR Method. *Int. J. Chem, Pharm, Sci*., 2016, 4(6): 307-314.

1. Introduction

Solubility of a drug is still a challenging area in pharmaceutical industry. With increasing pressure to identify high-quality drug candidates, it is critical to assess the absorption, distribution, metabolism, excretion (ADME) attributes of compounds early on during drug discovery. These instance properties are such as aqueous solubility, permeability, metabolic stability and in vivo pharmacokinetics. One of the crucial properties to candidate screening is the solubility of the compound. When an aqueous solvent is inadequate for a candidate drug, co-solvents are often employed to improve solubility [1]. When a drug is insoluble in water, polyethylene glycol (PEG) can be used as co-solvent due to its good solubilization properties [2].

The literatures presented several models for determining solubilization, however these methods require the collection of many experimental data which often not afforded in the drug discovery process specially where the drug compound has a short supply. Hence these approaches for determining solubilization of drugs are expensive, time consuming and without guarantee of success [3-9]. Therefore, it is necessary to solve this problem; development of models to predict the aqueous solubility of drug compounds using their chemical structures has attracted the attention of many pharmaceutical scientists. The predictive models based on molecular descriptors also help understanding what feature(s) limits solubility and can thus provide useful information to medicinal chemists. Several reports have been presented for predicting the solubility of drugs, but most of them are required to collect experimental data and large quantities of drugs. For example Breitkreutz et al. attempted to make model projections for 32 drug solubility using their solubility parameters [10]. Schroeter et al. considered the scope of application methods such as Gaussian process, accumulation regression, a support vector machine and random forest, for examining the ability to estimate the solubility of many drugs [11]. Jouyban et al. predicted the solubility of 26 drugs in mixtures of water and methanol using Jouyban model based on compound solubility at different temperatures. The obtained absolute mean error of the model forecasted 0.19 [12]. Duchowicz et al. made several linear models to examine the solubility of 166 drugs that the best model had R = 0.87 and standard deviation of 0.903 [13]. Faller et al. presented numerous models based on log P_{ow} and Abraham solubility parameter for a some drugs which provided the best model correlation coefficient 0.85 and 0.58 and the model error 0.80 and 0.69 for the training and test sets, respectively [9].

The main objective of this work was to investigate computer-based model derived from calculated molecular descriptors which can be used to predicted aqueous drug solubility as an important property influencing the absorption process. For this propose, the support vector machine (SVM) was used for modeling log1/S (S is solubility) of some drugs. Several descriptors were calculated. Then genetic algorithm (GA) and stepwise multiple linear regression (S-MLR) as nonlinear and linear techniques were employed for the variable selection, and their result was compared together. Also, the generality and reliability of model was validate by cross-validation and external test. Finally, the accuracy of the proposed method is compared with previous work of Faller and Ertl on this data set [9].

2. Experimental

The data set consists of 58 collective drugs from varies references [9]. Aqueous solubility (log 1/S) of these compounds was investigated. This data was presented in Table 1 and splitted into training (46 drugs) and external test (12 drugs) sets. Training set was used for constructing model by SVM and multiple linear regressions (MLR). Then the reliability of this model was investigated by model statistical parameters.

Descriptor generation and screening

An alternative strategy is to calculate firstly a large number of descriptors and then to remove those having a correlation coefficient superior to a defined value. Different strategies are used for descriptor selection. One of them is to select a particular set of descriptors with good performance on a given problem [14, 15]. Another different approach is to select the optimal combination of descriptors by the computer. Machine learning techniques can be also another alternative strategy to solve this kind of problem in this field. At the present work, CODESSA software was employed for calculating difference molecular descriptors [16] after optimzation on the basis of the minimum molecular energy that optimized by AM1 semi empirical method in the Hyperchem package (Ver. 7.0) [17]. After calculation of the molecular descriptors, those that are constant for all molecules were eliminated and the pairs of variables with a correlation coefficient greater than 0.90 were classified as inter-correlated, and one of them in each correlated pair was deleted. Then, stepwise multiple linear regression (SMLR) and genetic algorithm (GA) variable subset selection methods [18] were used for the selection of the most relevant descriptors from the pool of remaining 129 descriptors. These descriptors would be used as inputs for construction of SVM and MLR models.

SVM generation

Support vector machine (SVM), developed by Cortes and Vapnik [19] as a novel type of machine learning method, is gaining popularity due to many attractive features and promising empirical performance. A detailed description of SVM theory can be found in several excellent books and tutorials [20–23]. SVM was originally developed for classification problems, but they can also be extended to solve non-linear regression problems by the introduction of ε-insensitive loss function. SVM approach has been proposed to minimize the structural risk rather than the empirical risk; that is to preserve good generalization ability rather than optimizing the agreement with a given (limited) training set, and so constitutes a trade-off between the complexity of the model and its capability to reproduce experimental observations. In support vector regression, the input **x** is first mapped into a higher dimensional feature space by the use of a kernel function, and then a linear

model is constructed in this feature space. The kernel functions often used in SVM include linear or polynomial functions, radial basis functions and sigmoid functions. The linear model $f(\mathbf{x}, \omega)$ in the feature space is given by:

$$f(X,\mu) = \sum \mu_j g_j(X) + b \quad (1)$$

Where $g_j(\mathbf{x})$, j=1, m represents a set of non-linear transformations, ω_j and b are coefficients.

The quality of estimation is measured by the loss function $L(y, f(\mathbf{x}, \omega))$. SVM regression uses a new type of loss function called ε-insensitive loss function proposed by Vapnik [24]:

$$L_{\pounds}(y, f(X,\mu)) = \begin{cases} 0, & if\ |y - f(X,\mu)| \le \pounds \\ |y - f(X,\mu)| - \pounds & otherwise \end{cases} \quad (2)$$

The empirical risk is:

$$R_{emp}(\mu) = \frac{1}{n}\sum_{i=1}^{n} L_{\pounds}(y_i, f(X_i,\mu)) \quad (3)$$

SVM regression performs linear regression in the high dimension feature space using ε-insensitive loss and, at the same time, tries to reduce model complexity by minimizing ||ω||2. This can be described by introducing (non-negative) slack variables $\P_i$, $\P_i^*$ i= 1, . . ., n, to measure the deviation of training samples outside ε-insensitive zone. Thus SVM regression is formulated as minimization of the following functional:

$$\min \frac{1}{2}\|\mu\|^2 + Cn\sum_{i=1}(\P_i + \P_i^*) \quad (4)$$

This optimization problem can be transformed into a quadratic programming problem [25], by introduction of Lagrange multipliers α and its solution is given by:

$$f(X) = \sum_{i=1}^{n_{SV}} (\sim_i - \sim_i^*) K(X_i, X) \qquad 0 \le \sim_i^* \le C, \quad 0 \le \sim_i \le C \quad (5)$$

Where n_{SV} is the number of support vectors (SVs) and the kernel function:

$$K(X, X_i) = \sum_{j=1}^{m} g_i(X) g_i(X_i) \quad (6)$$

Parameter C is a regularization constant which determines the trade-off between the model complexity and the degree to which deviations larger than ε are tolerated in optimization formula [26]. The generalization performance of SVR depends on a good setting of parameters: C, ε and the kernel type and corresponding kernel 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. All SVM models in our present study were implemented using the SVM developed by Gunn [27]. The difference kernel functions like linear, polynomial and RBF were used in this work. For RBF kernel, the most important parameter is the width ($\pm$) of the radial basis function and for polynomial is degree of polynomial (d). All calculation programs implementing SVM were written in M-file based on MATLAB script. The overall performance of SVM was evaluated in terms of root-mean-square error (RMSE) which was defined in the next section. The calculations were performed on a 3.4 GHz Intel Pentium IV with 1GB RAM under Windows XP.

Model assessment method

Several parameters were applied to evaluating the performance of the developed QSPR models. Correlation coefficient (Q^2) was calculated to measure the correlation of the experimental and calculated values. Root mean square error (RMSE) was used to assess the prediction or generalization accuracy in the training, test and validation sets. Relative standard error (RSE) is utilized to estimate the relative error of the predictors. The representations of these parameters are defined as below:

$$Q^2 = 1 - \frac{\sum(\hat{y}_i - y_i)^2}{\sum(y_i - y_m)^2} \quad (8)$$

$$RMSE = \sqrt{\frac{\sum(\hat{y}_i - y_i)^2}{n-1}} \quad (9)$$

$$RSE = \sqrt{\frac{\sum(\hat{y}_i - y_i)^2}{\sum y_i^2}} \times 100 \quad (10)$$

Where y_i is the experimental value, $\hat{y}_i$ the calculated value, y_m the mean value, n the no. of the samples, and k is the no. of the used descriptors.

3. Results and Discussion

Descriptor selection

After the calculation of the descriptors, selection of best one that can predict and interpret is very important. There are different variable selection methods that are routinely applied in regression modeling to identify a small number of descriptors which "best" explain the variation in the response variable. In this work, stepwise-multiple linear regression and genetic algorithm were employed as linear and nonlinear variable selection methods, respectively. Stepwise-multiple linear regression (MLR) models are developed in a stepwise procedure. Thus, descriptors and correlations are ranked according to the values of the F-test and the correlation coefficient. In the following steps new descriptors are added/removed one-by-one. The descriptors selected by this method are: XY shadow (V_1), FPSA-3 (V_2), count of H-donors sites (v_3), Randic index (V_4) and Kier shape index (V_5).

Holland [28] introduced GA as problem-solving methods that simulate a natural evolution process. GA is well fitted to solve subset selection problems, such as variable selection. The flow chart of the genetic algorithm is shown in Fig. 1. The selected descriptors with GA are: 1)

Momentum of inertia, 2) Total molecular surface area, 3) Difference in CPSA, 4) PPSA-3 atomic charge weighed PPSA, 5) FPSA-3 fractional PPSA, 6) count of H-donors sites, 7) HA dependent HDSA-2/SQRT(TMSA) and 8) Kier shape index.

Training and test set selection

The division of a dataset into the training and test sets can be performed using various techniques. We used Y-ranking procedure for this object. This procedure is following that the dataset is sorted as ascending or descending according to experimental value of object. Then test set is selected from whole of sorted dataset. Moreover, to be sure uniformly distribution test set, the diversity analysis was used to generate training and test subsets. Molecular diversity analysis explores the way of molecules to cover a determined structural space and underlies many approaches for compound selection and design of combinatorial libraries. In this study, diversity analysis was performed for the data set based on decrypted algorithm by Luan and et. al. [19]. In this way, the mean distances of one sample (đ) to the remaining ones were computed from descriptor space matrix as follow:

$$\overline{d}_i = \frac{\sum_{j=}^{m} d_{i,j}}{n-1} \qquad i = 1,2,...,\ n \quad (11)$$

Where $d_{i,j}$ is A distance score for two different compounds X_i and X_j, X_i and X_j denote the collective database involving the descriptors. The closer to one the distance is the more diverse to each other the compound is. For the whole of data set, the mean distances of samples were calculated and plotted vs. log 1/S for variables selected by stepwise-MLR in Fig. 2-a and variables selected by GA methods in Fig. 2-b. As can be seen from these figures, the structures of the compounds are diverse according to variable spaces which were selected by stepwise-MLR and GA models. The training set with a broad representation of the chemistry space was adequate to ensure models' stability and the diversity of test set can prove the predictive capability of the model.

MLR modeling

Among all possible multiple combinations of describing parameters obtained for modeling log 1/S of drugs, the best statistical model was developed using five SMLR selected descriptors. The best penta-parametric model generated for the series is summarized below, along with the statistical parameters:

Los1/S=13.692(±1.567)-0.524(±0.214) V_1-2.226(±0.488) V_2+0.178(±0.063) V_3-0.458(±0.110) V_4 0.367(±0.075) V_5 n=46, R=0.82, SE=0.95, F=17 (12)

Then, test set was used for assessing the prediction power of constructed model. The statistical parameters R, RMSE and RSE were calculated for the MLR models. These results were presented in Table 2. MLR method was employed for investigation relationship between log 1/S and the selected GA descriptors. The octa-parametric constructed model was summarized below, along with the statistical parameters:

Los1/S=4.519(±1.750)+0.117(±0.065) W_1+0.504(±0.157) W_2-0.082(±0.045) W_3+0.022(±0.072) W_4 -1.784(±0.642) W_5 + 0.224(±0.061) W_6-0.076(±0.056) W_7-0.199(±0.068) W_8 n=46, R=0.85, SE=0.86, F=12 (13)

Besides deriving quantitative statistical significance models, an important aspect of QSAR modeling is validating the model since a good statistical fit does not guarantee the predictive ability of the model. In the view of above, the external consistency of the selected models was assessed by test set. The prediction ability of this model was obtained by external validation test in terms of parameters R, RMSE and RSE. These results were presented in Table 2.

SVM modeling

The selected descriptors with SMLR and GA were used as inputs in SVM, separately. In order to avoid overestimation in SVM, these models were assessed by leave-one-out cross-validation method. In this method, one data point is systematically deleted from the dataset and a QSPR model is constructed on the basis of reduced data set. The obtained model is subsequently used to predict the removed data point. The procedure was repeated until a complete set of predicted values is generated. The validation parameters calculated are Q^2 and RMSE based on equation 8 and 9, respectively. Different kernel functions were used for SVM modeling that linear and polynomial function had good results for SMLR and GA selected descriptors, respectively. The optimum conditions of these models have been presented in Table 3. After optimizing the parameters of the SVM model, the external test set was used for investigation of predictive ability of these models, and the statistical parameters of them were calculated. These results were presented in Table 2. According to this Table, it can be concluded that the statistical results of GA-SVM model is better than the other models. The plots of predicted log 1/S versus experimental log 1/S and the related residuals versus experimental log 1/S values, obtained by the GA-SVM modeling, were shown in Fig. 3 and 4, respectively. The good agreement observed between the predicted and experimental values (Fig. 3) and the random distribution of residuals about zero (Fig. 4) confirms the good predictive ability of MLR modeling.

Descriptor interpretation

Since GA-SVM model had the better results, it is feasible to discuss the effect of optimal subset of structural descriptors selected by GA on the aqueous solubility. These descriptors belong to three groups of electrical, topological and geometrical descriptors. The first one is momentum of inertia. Momentum of inertia is geometrical descriptor and was related to the rotational dynamics of a molecule [29]. It represents the potential energy of molecule for solubility. The next one is total molecular surface area which is a geometrical descriptor. The term of molecular surface usually refers to any surface surrounding some or all of the nuclei of the molecule and is very important to bind for other molecules [29]. The effective surface area of drug can be affected on solubility of drug in aqueous liquid of stomach. Difference in CPSA, PPSA-3 atomic charge weighed PPSA and FPSA-3 fractional PPSA, HA

dependent HDSA-2/SQRT (TMSA) are a set of charge descriptors which combine shape and electronic information to characterize molecules and therefore encode features responsible for polar interaction between molecules [29]. According to these descriptors, it can be concluded that charge surface are of drug affects to its salability.

The next descriptor is count of H-donors sites that is a topological descriptor. It is simply molecular descriptor based on counting the H-donor sites of a drug. Ability of H-donor bonding of drug with water can increase the solvation of drug in water. The last descriptor is kier shape index which is a topological descriptor. This descriptor depends on molecular graph and defined in terms of the number of graph vertex and the number of path. It characterizes the branching and rigidity of molecule and subsequently represents polarizabilaty of compound that can affect solubility values of drug in water solution.

The used descriptors by GA-SVM model in order to prediction of drug solubility contain parameters which have clear physical meaning and can be easily interpreted. These descriptors characterize the electrical and steric features of drug solubility in aqueous systems.

Table 1: Experimental and predicted values of aqueous solubility (log 1/S)

No.	Compound	Exp.	SMLR-MLR	GA-MLR	SMLR-SVM	GA-SVM
1	Acyclovir[t]	2.24	0.66	0.657	0.88	0.278
2	Amiloride	3.36	4.03	3.391	3.97	3.110
3	Amiodarone	8.10	6.58	6.327	6.64	6.073
4	Amitriptyline[t]	5.19	3.81	4.166	3.67	3.957
5	Amoxicilin	2.17	2.54	2.764	2.47	2.471
6	Ampicillin[t]	1.69	2.60	3.107	2.53	2.783
7	Atenolol	1.30	1.98	1.870	1.83	1.608
8	Atropine[tt]	1.61	3.27	3.070	3.15	2.821
9	Benzoic acid	1.59	1.94	2.221	2.12	1.840
10	Carbamazepine[tt]	3.40	4.88	3.178	4.81	2.810
11	Chlorpromazine	5.27	4.91	4.658	4.75	4.419
12	Cimetidine[t]	1.43	1.77	2.144	1.58	1.855
13	Ciprofloxacin	3.73	3.22	2.855	3.17	2.559
14	Clozapine[tt]	3.70	4.27	3.675	4.18	3.374
15	Corticosterone	3.20	3.34	3.294	3.22	3.095
16	Cortisone[t]	3.00	3.25	3.457	3.11	3.250
17	Desipramine	3.81	4.46	4.251	4.34	4.060
18	Diclofenac[tt]	5.59	3.21	3.555	3.02	3.349
19	Diltiazem[tt]	2.95	2.69	3.172	2.43	2.959
20	Dipyridamole	5.00	4.05	4.931	4.16	4.748
21	Doxycycline	2.35	1.67	2.645	1.52	2.413
22	Famotidine	2.48	1.82	1.416	1.78	1.075
23	Flurbiprofen[tt]	4.36	4.46	4.069	4.24	4.112
24	Furosemide	4.75	4.20	3.772	3.98	4.009

25	Glyburide	5.90	6.03	6.084	6.13	5.767
26	Hydrochlorothiazide[t]	2.63	3.68	3.483	3.57	3.559
27	Ibuprofen	3.62	3.67	3.875	3.55	3.617
28	Indomethacin	5.20	4.76	4.701	4.43	4.950
29	Ketoprofen	3.33	4.87	4.222	4.65	4.244
30	Labetalol[t]	3.45	2.51	2.293	2.34	2.021
31	Mefenamic acid	6.60	4.99	5.598	4.68	5.824
32	Methotrexate[t]	4.29	3.36	4.773	3.64	4.540
33	Metolazone	4.10	3.44	3.220	3.33	2.886
34	Metoprolol	1.20	1.98	2.376	1.74	2.119
35	Nadolol[tt]	1.57	2.10	2.502	1.92	2.279
36	Naproxen	4.21	5.00	5.147	4.76	5.289
37	Norfloxacin	2.78	2.75	2.411	2.69	2.110
38	Nortryptiline	4.18	4.81	4.307	4.71	4.080
39	Phenazopyridine	4.24	4.17	3.569	4.07	3.469
40	Phenytoin[t]	4.13	2.26	4.807	2.35	4.386
41	Pindolol[t, tt]	3.70	3.20	2.603	3.13	2.337
42	Piroxicam	5.48	5.36	5.851	5.10	5.719
43	Primaquine	2.77	3.36	2.548	3.27	2.363
44	Probenecid[t]	5.68	5.08	3.784	4.92	3.375
45	Progesterone	4.40	4.45	4.453	4.27	4.260
46	Promethazine	4.39	4.75	4.572	4.63	4.398
47	Propoxyphene	5.01	3.36	3.734	3.08	3.604
48	Propranolol	3.62	3.57	3.034	3.45	2.738
49	Quinine[tt]	2.82	3.68	3.280	3.56	3.093
50	Rufinamide	3.50	3.60	3.367	3.47	3.291
51	Tamoxifen	7.55	6.37	5.491	6.53	5.169
52	Terfenadine	6.69	6.34	6.930	6.49	6.940
53	Testosterone[tt]	4.06	4.21	4.314	4.10	4.087
54	Theophylline	1.38	1.74	3.432	1.81	3.526
55	Trovafloxacin	4.53	2.79	2.603	2.73	2.341
56	Valsartan[tt]	4.20	5.59	5.079	5.71	4.796

57	Verapamil[t, tt]	4.67	5.56	4.640	5.71	4.420
58	Warfarin	4.74	5.70	4.772	5.43	4.990

t and tt, denotes the test set compounds for SMLR and GA variable selection methods, respectively

Table 2: The statistical parameters of the generated models

Variable selection methods	Regression methods	Training set		Test set	
		R	RMSE	R	RMSE
SMLR	MLR	0.82	0.91	0.71	1.00
	SVM	0.82	0.90	0.73	1.02
GA	MLR	0.81	0.89	0.59	1.41
	SVM	0.85	0.81	0.60	1.53
Previous work [9]	Based on salvation parameters	—	—	—	—
	Based on log P(1)	0.74	—	0.69	—
	Based on log P(2)	0.75	—	0.57	—
	Hybrid method	0.85	0.80	0.58	0.69

Table 3: The optimum conditions of SVM model

Models	Kernel function	RMSE	Q^2	C	£
GA - SVM	Polynominal-1 order	0/81	0/64	250	0/25
SMLR- SVM	Linear	0/90	0/05	44000	0/65

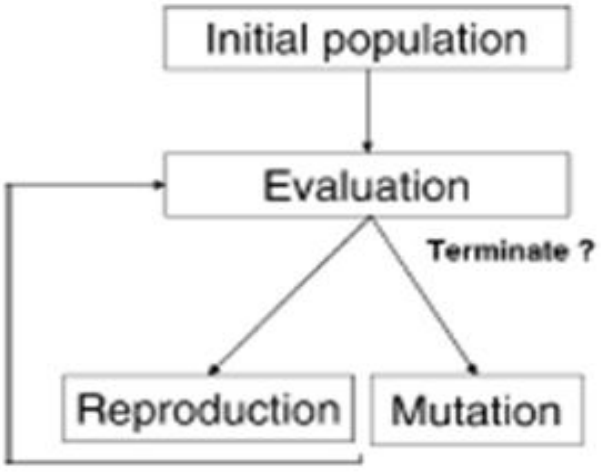


Figure 1: Flow chart representing the genetic algorithm

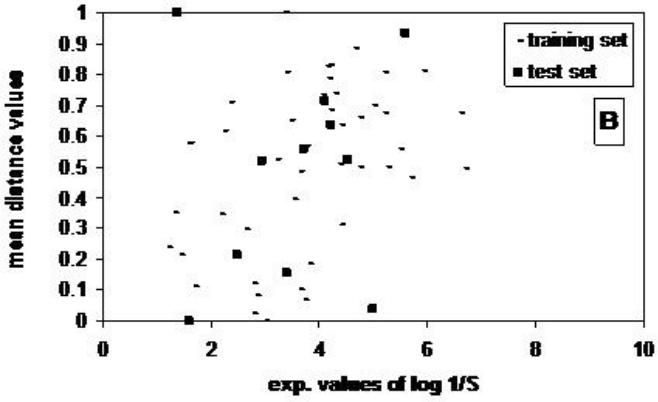
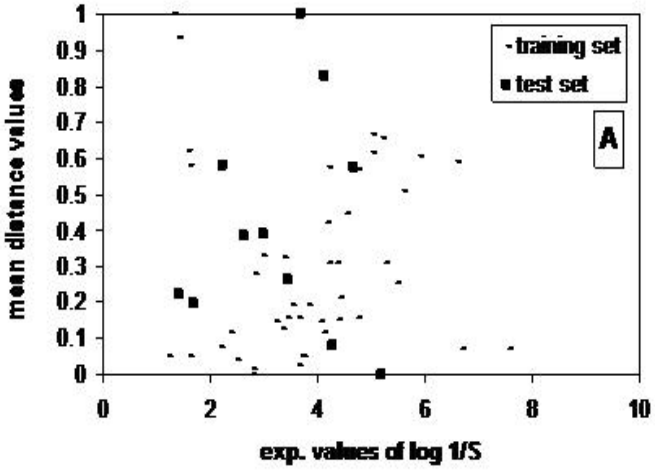


Figure 2: Diversity test results for splitting of data set: a) variable selected by stepwise-MLR, b) variable selected by GA



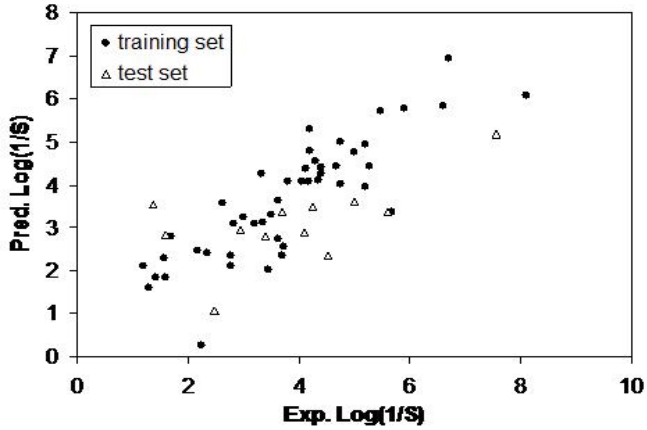


Figure 3: The plots of predicted log 1/S by the GA-SVM modeling versus experimental log 1/S

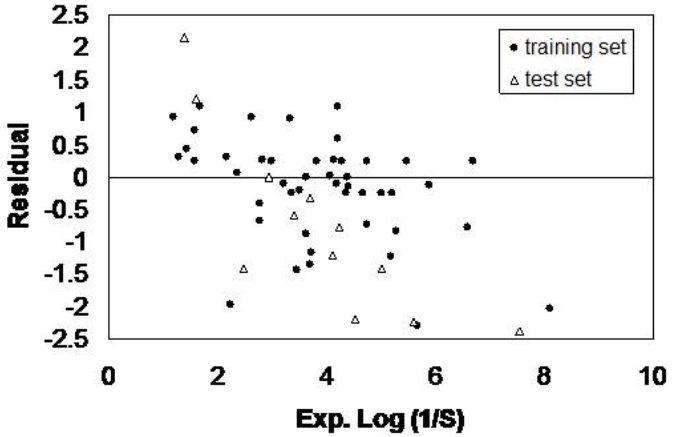


Figure 4: The plots of the residuals (predicted log1/S – experimental log1/S) versus experimental log 1/S values, obtained by the GA-SVM modeling

4. Conclusion

Summarizing the above discussion, the present study gives rise to QSPRs with good statistical significance and predictive capacity for log 1/S of varies drugs. Besides deriving quantitative models of statistical significance, an important aspect of QSAR modeling is validating the model since a good statistical fit does not guarantee the predictive ability of the model. In the view of above, the external consistency of the selected models was assessed by test set. The statistical result presents a good generality and ability of SVM model in drug delivery property. Some advantages of this method are simplicity, reliability and high speed calculation.

5. References

[1] Sweetana, S. and Akers, M. J., *PDA J. Pharm. Sci. Technol*., 1996, vo. 50, p. 330-342.
[2] Pang, S. N. J., *J. Am. Coll. Toxicol*., 1993, vo. 12, p. 429-457.
[3] Stouch, T. R., Kenyon, J. R., Johnson, S. R., Chen, X. Q., Doweyko, A. and Li, Y. , *J. Comput. Aided Mol. De.s*, 2003, vo. 17, p. 83-92.
[4] Pan, L., Ho, Q., Tsutsui, K. and Takahashi, L., *J. Pharm. Sci*, 2001, vo.90, p. 521–529.
[5] Grant, D. J. W. and Higushi, T., *Solubility Behavior of Organic Compounds*, New York, John Wiley, 1990.
[6] Avdeef, A., *Pharm. Pharmacol. Commun*, 1998, vo. 4, p. 165–178.
[7] Avdeef, A. and Berger, C. M., Brownell, C., *Pharm. Res*, 2000, vo. 17, p. 85–89.
[8] Glomme, A. and Maerz, J., Dressman, J. B. *J. Pharm. Sci*, 2005**,** vo. 94, p. 1–16.
[9] Faller, B. and Ertl, P., *Adv Drug Deliv Rev*, 2007, Vo. 59, P. 533–545.
[10] Breitkreutz, J., *Pharm. Res.*, 1998, Vo. 15(9), P. 1370-1375.
[11] schroeter, T. S., Schwaighofer, A., Mika, S., Laak, A. T., Suelzle, D., Ganzer,U., Heinrich, N. and Muller, K. R., *J. Comput. Aided. Mol. Des.*, 2007, Vo. 21, P. 485- 498.
[12] Jouban, A. and Acree Jr., W. E., *J. Pharmaceut. Sci.*, 2006, Vo. 9(2), P. 262-269.
[13] Duchowicz, P. R., Talevi, A., Bruno-Blanch, L. E. and Castro, E. A., *Bioorg. Med. Chem.*, 2008, Vo. 16, P. 7944-7955.
[14] B¨ocker, A., Schneider, G. and Teckentrup, A., *QSAR Comb. Sci.*, 2004, Vo. 23, P. 207–213.
[15] Xue, L., Godden, J.W., Stahura, F.L. and Bajorath, J., *J. Chem. Inf. Comput. Sci.*, 2003, Vo. 43, P. 1151–1157.
[16] Katritzky, A. R., Lobanov, V. S. and *Karelson*, M., *CODESSA Comprehensive descriptors for structural and statistical analysis*, Reference Manual, Version 2.0 (1994).
[17] HyperChem, Hypercube Inc., Release 7.0 for windows. (2002).
[18] http://www.models.kvl.dk/source/GAPLS/index.asp
[19] Cortes, C. and Vapnik, V., *Mach. Learn*. 1995, Vo. 20, P. 273-297.
[20] Cristianini, N. and Shawe-Taylor, J., *An Introduction to Support Vector Machines*, UK, Cambridge University Press: Cambridge, 2000.
[21] Joachims, T., *Learning to Classify Text Using Support Vector Machines: Methods, Theory, and Algorithms*, Kluwer, Thorsten Joachims**,** 2002.
[22] Sch olkopf, B. and Smola, A., *Learning with Kernels*, Cambridge, MA, MIT Press, 2002.
[23] Herbrich, R., *Learning Kernel Classifiers*, Cambridge, MA, MIT Press, 2002.
[24] Vapnik, V., *Statistical Learning Theory*, New York, Wiley, 1998.
[25] Buydens, L., Massart, D. L. and Geerlings, P., *Anal. Chem*., 1983, Vo. 55, P. 738–744.
[26] Stanton, D. T. and Jurs, P. C., *Anal. Chem*., 1989, Vo. 61, P. 1328–1332.
[27] http://www.isis.ecs.soton.ac.uk/isystems/kernel/
[28] Holland, J. H., *Adaptation in Natural and Artificial Systems*, Ann Arbor, MI, University of Michigan Press, 1975 (Revised Print: MIT Press, Cambridge, MA, 1992)
[29] Todeschini, R. and Consonni, V., *Handbook of Molecular Descriptors*, Weinheim, USA, Wiley/VCH, 2000.