Pak. J. Chem. 2(3): 112-116, 2012 Full Paper
ISSN (Print): 2220-2625
ISSN (Online): 2222-307X DOI: 10.15228/2012.v02.i03.p02

QSAR Prediction of Aqueous Solubility's of Some Pharmaceutical Compounds by Chemometrics Methods

[*]A. Azizi, [1]A. Niazi, [2]S. Mahmoudzadeh and [1]V. Najafi
[1]Department of Chemistry, Faculty of Science, Islamic Azad University, Arak Branch, Arak, Iran.
[2]Young Researchers Club, Islamic Azad University, Arak Branch, Arak, Iran.
[*]Department of Applied Chemistry, Bu-Ali Sina University, Hamedan, 65174 Iran.
E-mail: [*]zeaodin@yahoo.com

ABSTRACT
A quantitative structure–activity relationships (QSAR) study is suggested for the prediction of solubility of pharmaceutical compounds in aqueous solution by using chemometrics methods. Ab initio theory was used to calculate some quantum chemical descriptors including electrostatic potentials and local charges at each atom, HOMO and LUMO energies, etc. Also, Dragon software was used to calculate some descriptors such as WIHM and GETAWAY. QSAR studies are mathematical quantification of relations between structure and activity or property. These are extensively used in pharmaceutical and agricultural chemistry for screening potential compounds for specific biological activity. Computable molecular descriptors are preferred to experimental properties in QSAR analyses because require molecular structure as the only input and can be in expensively calculated for a chemical in less than a millisecond. By multivariate calibration methods such as partial least squares (PLS) regression and least squares support vector analysis (LS-SVM), it is possible to obtain a model adjusted to the concentration values of the mixtures used in the calibration range. Orthogonal signal/descriptor correction (OSC/ODC) is a preprocessing technique used for removing the information unrelated to the target variables based on constrained principal component analysis. OSC is a suitable preprocessing method for PLS calibration of mixtures without loss of prediction capacity using cited descriptors. The root mean square error of prediction (RMSEP) was also quite acceptable for OSC-PLS (0.0095) and LS-SVM (0.0023)

Keywords*:* Solubility, Pharmaceutical, PLS, OSC-PLS, LS-SVM, WHIM, GETAWAY.

1. INTRODUCTION
Aqueous solubility is one of the most important physicochemical properties that plays a significant role in various physical and biological processes and has a marked impact on the design and pharmaceutical formulation development. For weak electrolyte drugs, salt formation is a common approach to improve its solubility, since it is a much simpler method than complex molecular modifications. Using different counter-ions can result in salts with difference in physicochemical properties. Until now, various organic and inorganic salts of acidic and basic drugs have been prepared and their physicochemical properties were subsequently determined in order to aid the selection of the most suitable salt for drug development[1-2].

Among the investigation of QSAR, one of the most important factors affecting the quality of the model is the method to build the model. Many multivariate data analysis methods such as multiple linear regression (MLR)[3], partial least squares (PLS)[4] and artificial neural network (ANN)[5] have been used in QSAR studies. MLR, as most commonly used chemometrics method, has been extensively applied to QSAR investigations. However, the practical usefulness of MLR in QSAR studies is rather limited, as it provides relatively poor accuracy. ANN offers satisfactory accuracy in most cases but tends to over fit the training data. The support vector machine (SVM) is a popular algorithm developed from the machine learning community. Due to its advantages and remarkable generalization performance over other methods, SVM has attracted attention and gained extensive applications[6-7]. As a simplification of traditional of SVM, Suykens and Vandewalle[8-9] have proposed the use of least-squares SVM (LS-SVM). LS-SVM encompasses similar advantages as SVM, but its additional advantage is that it requires solving a set of only linear equations (linear programming), which is much easier and computationally more simple[10].

The basic principle of the multivariate calibration is the simultaneous utilization of many independent variables, x_1, x_2, . . . , x_n, to quantify one or more dependent variables of interest, y. The partial least squares (PLS) regression analysis is the most widely used method for this purpose, and it is based on the latent variable decomposition relating two blocks of variables, matrices X and Y, which may contain spectral and concentration data, respectively. These matrices can be simultaneously decomposed into a sum of f latent variables, as follows:

$$Y = TP^T + E = \sum t_f p'_f + E \qquad (1)$$

$$Y = UQ^T + E = \sum u_f q'_f + E \qquad (2)$$

in which T and U are the score matrices for X and Y, respectively; P and Q are the loadings matrices for X and Y, respectively, E and F are the residual matrices. The two matrices are correlated by the scores T and U, for each latent variable, as follows:

*Corresponding Author Received 22nd March 2012, Accepted 23rd May 2012

$$u_f = b_f t_f \qquad (3)$$

in which b_f is the regression coefficient for the f latent variable. The matrix Y can be calculated from u_f, as Eq. (4), and the concentration of the new samples can be estimated from the new scores T^*, which are substituted in Eq. (4), leading to Eq. (5)

$$Y = TBQ^T + F \qquad (4)$$

$$Y_{new} = T^* BQ^T \qquad (5)$$

In this procedure, it is necessary to find the best number of latent variables, which normally is performed by using cross-validation, based on determination of minimum prediction error. Several determinations based on the application of this method to spectrophotometric and QSAR data have been reported by several workers[11-15].

Orthogonal signal correction (OSC) was introduced by Wold et al. to remove systematic variation from the response matrix X that is unrelated, or orthogonal, to the property matrix Y. Therefore, one can be certain that important information regarding the analyte is retained. Since then, several groups have published various OSC algorithms in an attempt to reduce model complexity by removing orthogonal components from the signal[16-17].

Theory of LS-SVM has also been described clearly by Suykens et al. and application of LS-SVM in quantification, classification and QSAR reported by some of the workers[18-19]. So, we will only briefly describe the theory of LS-SVM. The LS-SVM is capable of dealing with linear and nonlinear multivariate calibration and resolves multivariate calibration problems in a relatively a fast way. In LS-SVM a linear estimation is done in kernel-induced feature space $(y = w^T \phi(x) + b)$. In the present paper, the PLS, OSC-PLS and LS-SVM methods were applied in QSAR/QSSR for modeling the relationship between the solubility of 148 pharmaceutical compounds by using ab initio and structural molecular descriptors.

2. COMPUTATIONAL METHODS

2.1 *Hardware and software*

The computations were made with an AMD 2000 XP (1 Gb RAM) microcomputer with the Windows XP operating system and with Matlab (version 7.0, Mathwork, Inc.). The PLS evaluations were carried out by using the PLS program from PLS-Toolbox Version 2.0 for use with Matlab from Eigenvector Research Inc. The LS-SVM optimization and model results were obtained using the LS-SVM lab toolbox (Matlab/C Toolbox for Least-Squares Support Vector Machines). ChemDraw Ultra version 9.0 (ChemOffice 2005, CambridgeSoft Corporation) software was used to draw the molecular structures and optimization by the AM1. Descriptors were calculated utilizing Dragon software[20]. These descriptors are calculated using two-dimensional representation of the molecules and therefore geometry optimization is not essential for calculating these types of descriptors.

2.2 *Data set*

The QSSR model for the estimation of the solubility's of pharmaceutical compounds is established in the following steps: the molecular structure input and generation of the files containing the chemical structures is stored in a computer-readable format; quantum mechanics geometry is optimized with a semi-empirical (AM1) method; structural descriptors are computed; and the structural-solubility model is generated by the chemometrics methods and statistical analysis. The solubility's of 148 pharmaceutical compounds was collected from[21].

3. RESULTS AND DISCUSSION

Solubility's of 148 of pharmaceutical compounds taken from the literature[21], and was studied in this study. A major step in constructing QSAR/QSRR models is finding one or more molecular descriptors that represent variation in the structural property of the molecules by a number. A wide variety of descriptors have been reported to be used in QSAR/QSRR analysis[20].

A pool containing molecular descriptors is derived to property characterize the chemical structure of the these compounds, involving variables of the type Constitutional, Topological Geometrical, Charge, GETAWAY (Geometry, Topological, Atoms-Weighted Assembly), WHIM (Weighted Holistic Invariant Molecular descriptors), 3D-MoRSE (3D-Molecular Representation of Structure based on Electron diffraction), Molecular Walk Counts, BCUT descriptors, 2D-Autocorrealtions, Aromaticity Indices, Randic molecular profiles, Radial Distribution Functions, Functional Groups and Atom-Centered Fragments. These variables are calculated by means of the software Dragon. For the evaluation of the predictive ability of a different model, the root mean square error of prediction (RMSEP) and relative standard error of prediction (RSEP) can be used.

$$RMSEP = \sqrt{\frac{\sum_{i=1}^{n}(y_{pred} - y_{obs})^2}{n}} \qquad (6)$$

$$RSEP(\%) = 100 \times \sqrt{\frac{\sum_{i=1}^{n}(y_{pred} - y_{obs})^2}{\sum(y_{obs})^2}} \qquad (7)$$

where y_{pred} is the predicted concentration in the sample, y_{obs} is the observed value of the concentration in the sample and n is the number of samples in the validation set.

In order to detect the homogeneities in the data set and identify possible outliers and clusters, PCA was performed within the calculated structure descriptors space for the whole data set. PCA is a useful multivariate statistical technique in which new variables (called principal components, PCs) are calculated as linear combinations of the old ones. These PCs are sorted by decreasing information content (i.e. decreasing variance) so that most of the information is preserved in the first few PCs. An important feature is that the obtained PCs are uncorrelated, and they can be used to derive scores which can be used to display most of the original variations in a smaller number of dimensions. These scores can also allow us to recognize groups of samples with similar behavior (Fig-1).

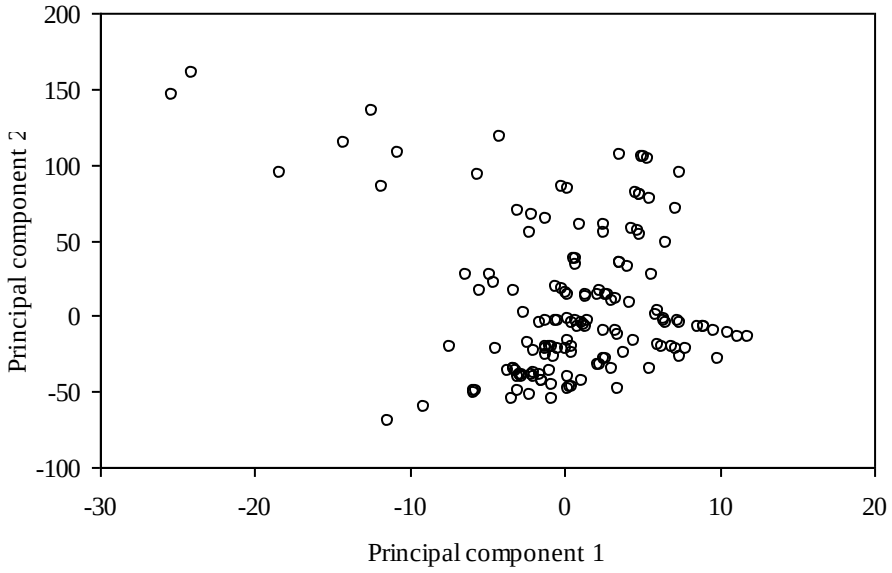

Fig-1: Principal components analysis of the structural descriptors for the data set.

3.1 *PLS analysis*

The factor-analytical multivariate calibration method is a powerful tool for modeling, because it extracts more information from the data and allows building more robust models. According to solubility's of pharmaceutical data[21], data classified to training and prediction sets by descriptor compound selection (MDC). The optimum number of factors to be included in the calibration model was determined by computing the prediction error sum of squares (PRESS) fro cross-validated models using a high number of factors. The cross-validation method employed was to eliminate only one compound at a time and then PLS calibrated the remaining of training set. The solubility of the left-out sample was predicted by using this calibration. This process was repeated until each compound in the training set had been left out once. According to Haaland suggestion[22], the optimum number of factor was selected (Fig-2).

3.2 *Preprocessing by orthogonal signal correction*

For calibration set six OSC components were used for filtering. Evaluation of the prediction errors for the validation set reveals that the OSC treated data give substantially lower RMSEP values than original data. Also, the OSC-filtered data give much simpler calibration models with fewer components than the ones based on original data. The results imply that the OSC method indeed removes information from descriptor data that is not necessary for fitting of the Y-variables. In some cases the OSC method also removes non-linear relationships between X and Y. The OSC-PLS score plots depicted in a more clear way the location of the solutions in the scores map which are the same as square experimental design was used in preparation of calibration set.

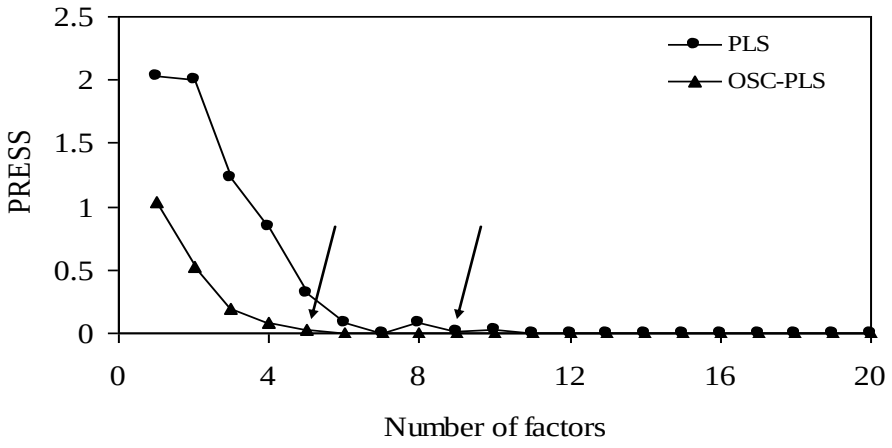

Fig-2: Plots of PRESS versus number of factors by PLS and OSC-PLS.

3.3 *LS-SVM analysis*

The all descriptors were used as the input to develop nonlinear model by LS-SVM. The quality of LS-SVM for regression depend on γ and σ^2 parameters. In this work, LS-SVM was performed with radial basis function (RBF) as a kernel function. To determine the optimal parameters, a grid search was performed based on leave-one-out cross-validation on the original training set for all parameter combinations of γ and σ^2 from 1 to 500 and 1 to 500, respectively, with increment steps of 1. Table 1 shows the optimum γ and σ^2 parameters for the LS-SVM and RBF kernel, using the calibration sets.

3.4 *Prediction of solubility of pharmaceutical compounds*

The predictive ability of these methods (PLS, OSC-PLS and LS-SVM) were determined using 8 solubility (their structure are given in Table-1). The results obtained by PLS, OSC-PLS and LS-SVM methods are listed in Table-1. Table-1 also shows RMSEP, RSEP and the percentage error for prediction of solubility of these compounds. As can be seen, the percentage error was also quite acceptable only for OSC-PLS and LS-SVM.

Table-1: Actual and predicted values of solubility of complex petroleum compounds using PLS, OSC-PLS and LS-SVM models.

Substance	Experimental Solubility	Predicted solubility					
		PLS	Error (%)	OSC-PLS	Error (%)	LSSVM	Error (%)
Aspirin	0.663	0.689	3.92	0.672	1.36	0.661	-0.30
Histidine	1.659	1.682	1.39	1.668	0.54	1.662	0.18
Picric Acid	1.104	1.058	-4.17	1.088	-1.45	1.105	0.09
Adenine	0.013	0.018	38.46	0.016	23.08	0.013	0.00
Diazinon	-1.398	-1.392	-0.43	-1.387	-0.79	-1.397	-0.07
Biotin	-0.658	-0.666	1.22	-0.663	0.76	-0.655	-0.46
Isoniazid	2.146	2.162	0.75	2.153	0.33	2.148	0.09
Diazepam	-1.301	-1.316	1.15	-1.311	0.77	-1.305	0.31
N.F. [a]		9		5			
PRESS		0.0224		0.0202			
γ						100	
σ^2						20	
RMSEP		0.0221		0.0095		0.0023	
RSEP (%)		1.7316		0.7425		0.1833	

[a]Number of factor

Good results were achieved in LS-SVM model with percentage error ranges from -0.46 to 0.31 for solubility of pharmaceuticals. Also, it is possible to see that LS-SVM presents excellent prediction abilities when compared with other regression. According to the results, structural descriptors are suitable descriptors for describing the solubility of these compounds. When LS-SVM method with all descriptors is used, prediction of solubility in test step, with a small error is possible; this is improved in comparison with other method (PLS and OSC-PLS). This shows that by using all structural descriptors and also LS-SVM method, the solubility pf pharmaceutical compounds is predicted with satisfactory results.

4. CONCLUSION

A least squares-support vector machine (LS-SVM) model was established to predict the solubility of some pharmaceutical compounds. A proper model with high statistical quality and low prediction errors was obtained. The model could predict the solubility of drug compounds not existed in the modeling procedure accurately. The structural and topological descriptors concerning to the whole molecular properties and those of individual atoms in the molecule were found to be important factors controlling the solubility behavior.

5. REFERENCES

1. Niazi, A., Jameh-Bozorghi, S., Nori-Shargh, D. Chin. Chem. Lett. **(2007)**, 18, 621, http://dx.doi.org/10.1016/j.cclet.2007.02.017.
2. Forbes, R. T., York, O., Davidson, J. R., Int. J. Pharm. **(1995)**, 126, 199, http://dx.doi.org/10.1016/0378-5173(95)04126-5.
3. Niazi, A., Jameh-Bozorghi, S. Nori-Shargh, D. J., Hazard, Mat. **(2008)**, 151, 603, http://dx.doi.org/10.1016/j.jhazmat.2007.06.030.
4. Niazi, A., Jameh-Bozorghi, S., Nori-Shargh, D., Turk. J. Chem. **(2006)**, 30, 619.
5. Buyukbingol, E., Sisman, A., Akyildiz, M., Alparslan, D. N., Adejare, A., Bioorg. Med., Chem. **(2007)**, 15, 4265, http://dx.doi.org/10.1016/j.bmc.2007.03.065.
6. Vapnik, V., Statistical Learning Theory, John Wiley, New York, **(1998)**.
7. Cortes, C., Vapnik, V., Mach. Learn. **(1995)**, 20, 273.
8. Suykens, J. A. K., Vandewalle, J., Neural Process. Lett. **(1999)**, 9, 293, http://dx.doi.org/10.1023/A:1018628609742.
9. Suykens, J. A. K., Gestel, T., Brabanter, J., Moora, B., Vandewalle, J., Least-Squares Support Vector Machines, World Scientifics, Singapore, **(2002)**.
10. Suykens, J. A. K., Eur. J. Control **(2001)**, 7, 311, http://dx.doi.org/10.3166/ejc.7.311-327.
11. Niazi, A., Azizi, A., Turk. J. Chem. **(2008)**, 32, 217.
12. Niazi, A., Ghasemi, J., Yazdanipour, A., Anal. Lett. **(2005)**, 38, 2377, http://dx.doi.org/10.1080/00032710500317975.
13. Niazi, A., Soufi, A., Mobarakabadi, M. Anal. Lett. **(2006)**, 39, 2359, http://dx.doi.org/10.1080/00032710600755868.
14. Niazi, A., Braz, J. Chem. Soc. **(2006)**, 17, 1020.
15. Niazi, A., Croa. Chem. Acta **(2006)**, 79, 573.
16. Niazi, A., Yazdanipour, A., Hazard, J., Mat. **(2007)**, 146, 421.
17. Niazi, A., Ghasemi, J., Zendehdel, M., Talanta **(2007)**, 74, 247, http://dx.doi.org/10.1016/j.talanta.2007.06.005.
18. Niazi, A., Ghasemi, J., Yazdanipour, A., Spectrochim. Acta Part A **(2007)**, 68, 523, http://dx.doi.org/10.1016/j.saa.2006.12.022.
19. Niazi, A., Sharifi, S., Amjadi, E. J. Electroanal. Chem. **(2008)**, 623, 86, http://dx.doi.org/10.1016/j.jelechem.2008.06.021.
20. Todeschini, R., Consonni, V., Handbook of Molecular Descriptors, Wiley-VCH, Weinheim, **(2000)**, http://dx.doi.org/10.1002/9783527613106.
21. Duchowicz, P. R., Talevi, A., Bellera, C., Bruno-Blanch L. E., Castro E. A., Bioorg. Med. Chem. **(2007)**, 15, 3711, http://dx.doi.org/10.1016/j.bmc.2007.03.044.
22. Haaland, D. M., Thomas, E. V., Anal. Chem. **(1988)**, 60, 1193, http://dx.doi.org/10.1021/ac00162a020.