Innovations in Pharmaceuticals and Pharmacotherapy

www.innpharmacotherapy.com

eISSN: 2321–323X

Research Article

Prediction of solubility of some chemicals in water, polyethylene glycol and their binary mixtures

Mohammad Hossein Fatemi*, Fatemeh Bagheri

Chemometrics Laboratory, Faculty of Chemistry, University of Mazandaran, Babolsar, Iran

Abstract

With the aim of solubility $A=\pi r^2$ estimation in water, polyethylene glycol 400 (PEG) and their binary mixtures, quantitative structure–property relationships (QSPR) were used to relate the solubility of a large number of chemicals to their molecular descriptors. Descriptors that were used by can encode features of molecules which are affected on dispersion, hydrophobic and steric interactions between solute and solvent molecules. To develop QSPR models, the methods of multiple linear regressions (MLR), least-squares support vector machine (LS-SVM), and artificial neural network (ANN) were used. The obtained statistical parameters of these models revealed that LS-SVM model was superior to the others. The standard error (SE), for LS-SVM model is: 0.270 and 0.697 for training and test set respectively. The leave-one-out cross validation lead to R_2^{cv}= 0.881 and SPRESS = 0.405 for LS-SVM model. These values and other statistics of this model indicate the robustness and credibility of developed LS-SVM model.

Keywords: Solubility, polyethylene glycol; quantitative structure–property relationships; least-squares support vector machine;molecular descriptor.

***Corresponding author: Mohammad Hossein Fatemi *** Chemometrics Laboratory, Faculty of Chemistry, University of Mazandaran, Babolsar, Iran. E- mail: mhfatemi@umz.ac.ir

1. Introduction

The solubility of a drug candidate depends on their physical and chemical properties and also to the solvent properties such as; polarity, dielectric constant, autoprotolysis constant of the solvent and also temperature and pH of the solution. Water is the main solvent and the aqueous solubility is one of the most important properties of a drug molecule. Aqueous solubility of a drug candidate influences absorption, distribution, metabolism, and excretion (ADME) properties. 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 [1]. Moreover aqueous solubility influence on metabolism and excretion 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 [2]. Many of drugs have low solubility in water. There are several methods to enhance the solubility of drugs such as: cosolvency, complexion, and ionization. Mixing a permissible nontoxic organic solvent with water, (cosolvency) is the most common technique to increase the aqueous solubility of drugs. Effects of volume fractions of a co-solvent in the binary mixtures of water can model theoretically. One of these methods was quantitative structure–property relationships (QSPR) approaches. In this method the chemical properties or (activities) of chemicals were mathematically related to the structural features (molecular descriptors) of molecules. There are some reports about QSPR modeling of aqueous solubility of chemicals. One earlier model that was developed by Paruta et al. described the solubility behavior of chemicals by using the dielectric constant of the mixed solvents [3]. 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 R2 = 0.87 and standard error of SE= 0.51[4].Also a QSPR model for prediction of solubility of 122 drugs in 0%, 25%, 50%, and 75%of PEG (v/v in water) was developed by Rytting et al [5]. In this work the solubility data of 84 drugs were modeled by linear regression using the following molecular descriptors: molecular weight, volume, radius of gyration, density, number of rotatable bonds, hydrogen-bond donors, and hydrogen-bond acceptors. QSPR-based models developed at each volume fraction with the training set compounds showed a reasonable correlation coefficient R of ~0.9 and a root mean square error (RMSE) of 0.5 in log unit. In the present work we try to improve this model by using non-linear feature mapping techniques such as; artificial neural network (ANN) and least squares- support vector Machine (LS-SVM).

2. Material and methods

Data set

The solubility data that was reported by Rytting et al., was used in this study. The data set consisted of equilibrium solubility of 122 compounds in 0%, 25%, 50%, and 75% (V/V) of PEG in water [6]. The compounds represent a broad range of log P values (−2.4 to 7.5), molecular weights (111–614 Da) and melting points (53.5–360 ◦C). 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 consists of 98 molecules, and the internal and external test sequally have 12 members. In developing the ANN model, training set was used in training and optimization of model parameters, during these processes the internal test set was used to monitor the extent of model development and prevention of over training. Prediction power of model was evaluated on independent data external test set that was not used during the training step.

In the case of MLR and SVM models internal and external test sets were considered as test set.

Molecular Descriptor

The molecular descriptor is the final result of a logic and mathematical procedure, which transforms chemical information encoded within a symbolic representation of a molecule in to the useful numbers [7]. Obviously, it follows that the information content of a molecular descriptor depends on the kind of a molecular representation and algorithm used for its calculation In order to calculate molecular descriptors in this work the chemical structures of molecules were drawn by using Hyperchem package (Ver. 7) [8] and optimized by the AM1 semi-empirical method. Then the package of Dragon (Ver.3) [9] was used to calculate molecular descriptors from the Hyperchem output files. This package can calculate various types of descriptors such as constitutional, topological, geometrical and charge descriptors [10]. After calculation of descriptors near-constant variable, would be excluded. Then pair of variables with a correlation coefficient greater than 0.90 were classified as inter-correlated, and only one of them was considered in developing of the QSPR model. In order to show the solvent composition the following parameter (SCD) was calculated and considered as solution composition descriptor:

$$SCD = \Phi . \varepsilon_0 H_2O + (1-\Phi)\, \varepsilon_{0\,PEG} \qquad (1)$$

In the above equation Φ is the volume fraction of water and $\varepsilon_0 H_2O$ and ε_{0PEG} is the dielectric constant for water and polyethylene glycol, respectively. In order to select the most important descriptors from remaining 578 descriptors, the value of adjusted squared of correlation coefficient (R^2u) was calculated and were plotted versus the number of descriptors in the models for the 1-15 parameter models that obtained by stepwise multiple linear regression**.** Consequently, the model corresponding to the break point is considered as the best/optimum model. As can be seen in Figure1, the application of the "break point" algorithm led to conclusion that the best model had seven parameters. These seven descriptors were used as independent variables in developing linear and non-linear quantitative structure activity relationship models. The name of selected descriptors are "n" umber of fragments of type

X-C(=X)-X (C-041), H attached to C1(sp3)/C0(sp2) (H-047), Moriguchioctanol-water partition coefficient (MLOGP), relative negative charge (RNCG), (3D-MoRSE) signal 21/weighted by mass (Mor21m), molecular multiple path count of order 5 (piPC05) and the (SCD) term. These descriptors would be used as inputs for developing of ANN, LS-SVM, and MLR models. Table 1 shows the correlation matrix among these seven descriptors. As it can be seen from this table, there is no high correlation among the selected descriptors.

Non-linear modeling

Today's artificial neural network represents a promising modeling technique especially in nonlinear modeling, which is frequently encountered in QSPR studies [11]. An artificial neural network is a biologically inspired computer program designed to simulate the way in which the human brain processes information. A detailed descriptions of the theory behind the ANN have has been adequately described elsewhere [12,13]. 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 [11]. In the present work, the STATISTICA package (ver.7) [14] was used for developing the ANN model. Another nonlinear feature mapping technique is support vector machine. This method algorithm has been introduced for solving classification and regression problems [15-17] and then applied successfully to many areas [18, 19]. Based on the statistical learning theory and the structural risk minimization principle, SVMs obtain the solution by solving the quadratic programming problem while avoiding the local minima, which provides an advantage over other regression techniques. The least squares version of the SVM algorithms (LS-SVM) [20,21] finds the solution by solving a set of linear equations. The motivation for choosing LS-SVMs as the approximation tool is their higher generalization capability, as well as the achievement of an almost global solution within a reasonably short training time. The LS-SVM model can be expressed as:

$$y_{\Sigma}= \sum_{i=1}^{N} \alpha_i k(x_{i.}) + \mathrm{b} \qquad (2)$$

$$\alpha_i=2\gamma e_i \qquad (3)$$

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 bi as term and the γ 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 [23]. Themost general kernel function is radial basis function (RBF):

$$K(x_i,x)=\exp(-||x_i-x_j||^2/2\sigma^2) \ldots\ldots(4)$$

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 [22]. In this work, we established LS-SVM by RBF kernel function to estimate the aqueous solubility of 122 drug compounds by using the STATISTICA package.

Results and discussion

In this work the method of stepwise multiple linear regression was used for the selection of the most relevant descriptors, and MLR, LS-SVM, and ANN methods were used as feature mapping techniques to build linear and nonlinear QSPR models. The data set and corresponding observed and predicted values of the –log (sol) of all molecules studied in this work are shown in Table2.

Diversity analysis

Rational division of the experimental data set into training and test sets are an important part in the development and validation of reliable QSPR model. In this study, diversity analysis was performed to make sure that the structures of thetraining and test cases can represent those of the whole ones[11]. In this way, the mean distances of one sample to the remaining ones (di) was computed from descriptor space matrix as follows:

d_i= (∑_(𝑖 = 1)^𝑛▒𝑑_𝑖𝑗)/ i=1,2...,n (5)

whered_{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}= $\sqrt{\sum_{k=1}^{m}(x_{ik}{}^{2}}$ (6)

Then the mean distances were normalized within the interval of zero to one and the resulting values were plotted against –log (sol) (Figure.2). 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.

Modeling

In order to selecting the most relevant descriptors, the stepwise MLR technique was performed on the training set by SPSS package (V.20). By using break point procedure 7-parameter MLR model can be considered as the best linear model (Figure1). The obtained MLR model has the following specifications:

-Log(sol) =1.154 (±0.230) + 0.724 (±0.092 × C041 - 0.083 (±0.009) × H047 - 6.740 (±0.885) × RNCG+0.134 (±0.024) × MLOGP + 0.129 (±0.0230)× Mor21m + 0.005 (±0.001) x piPC05 + 0.029 (±0.001) × SCD…………………..(7)

n=583 R=0.758 F=110.842 SE=0.8238

where R is the correlation coefficient, SE is the standard error and F is the F-test of significance. In order to examine any nonlinear relationship between solubility and selected molecular descriptors artificial neural network and support vector machine were used. In the first step a three-layer network with a sigmoid transfer function was designed, which selected seven descriptors were used as its inputs and -log (sol) values as outputs. After optimization and training of this network, it was used to calculate the –log (sol) for test sets as well as training set. These calculated values are indicated in Table 2. The statistical parameters of these calculations were shown in Table 3. Another nonlinear feature mapping method is support vector machine. Parameters of SVM including γ for RBF kernel function; epsilon (ε) and capacity (c) must be optimized to achieving the optimum performance. The optimization of LS- SVM parameters was performed by systemically changing their values and calculating the RMSE of the model. The optimal value for ε depends on the type of noise present in the data, which is usually unknown. Even if enough knowledge of the noise is available to select an optimal value for ε, there will be some practical consideration of the number of resulting support vectors. The parameter of ε-insensitivity prevents the entire training set to meeting boundary conditions and allows the possibility of sparsity in the dual formulations solution. So, choosing the appropriate value of ε is a critical step. To find an optimal value for ε, the RMSE of SVM models with different ε values was calculated. The variation of RMSE versus the epsilon values is plotted in Figure3.which indicates that, the optimal value of ε was 0.05. To find an optimal value of C, the RMSE of LS-SVM models with different C values is calculated plotted in Figure 4, which indicates that the optimum value of C was 5. Then the developed LS-SVM model was used to calculation the solubility of all molecules in data set. These values were shown in Table1. Statistical parameters of these calculations were shown in Table 3. Figure 5 indicate the plot of LS-SVM calculated versus the experimental values of log(sol) for whole molecules in the data set, which shows the good correlation between them (R_{train}= 0.965 , R_{test}= 0.832). The residual of the LS-SVM calculated values of the -log (sol) are plotted against their experimental values in Figure6. The propagation of the residuals on both sides of zero line shows that no systematic error exists in the developed LS-SVM model.

Descriptors

In order to determine the relative importance of each variable in the SVM model, the sensitivity analysis approach was applied. This method is performed based on the sequential removal of variables by zeroing the specific descriptor. For each sequentially zeroed input variable, root-mean-square error of prediction (RMSEP) as the prediction error was calculated. Generally RMSEP value increases in this way. Then, differences between RMSEP and root-mean-square error of established SVM was calculated and shown as DRMSE. Each variable which causes greater value of DRMSE is more important. The result of these calculations indicated that the order of importance of input descriptors in SVM model is: SCD>H047>MLOGP>piPC05>MOR21M>C041>RNCG

As mentioned earlier the SCD term relates to solvent composition and can be calculated from equation 1. Descriptor of H-047 represent the H attached to C1 (sp3)/C0 (sp2), which can encode some structural features of molecule that effects on solute-solvent interactions [23]. Moriguchioctanol-water partition coefficient (MLogP) is calculated from Moriguchilog P model consisting of a regression equation based on 13 structural parameters, which indicated the liphophilicity of solutes [24]. Molecular multiple path count of order 5(piPC05) is defined as the sum of the weights of the paths of length 5 in the molecule. This descriptor relate to hydrophilicity of molecule. Mor21m (3D MoRSE-signal 21/weighted by atomic masses) is belonged to 3D-MoRSE descriptors [25].These types of descriptors are calculated based on the idea of obtaining information from the 3D atomic coordinates by the transform used in electron diffraction studies for preparing theoretical scattering curves. 3D-MoRSE descriptors are important in QSOR studies because they take into account the 3D arrangement of the atoms without ambiguities (in contrast with those coming from chemical graphs). Since these descriptors do not depend on the molecular size, thus being applicable to a large number of molecules with great structural variance and being a characteristic common to all of them. Number of fragments of type X-C(=X)-X (C-041) is the next descriptor. This atom centered fragment descriptor is defined by looking at the first neighbors of carbon atoms and can encode some information about topological of a molecule. The last descriptor is RNCG that indicate the relative negative charge on carbon atoms and belong to the electronic descriptors [8]. More explanations about theory behinds these descriptors can be found in the book Hand book of Molecular Descriptors by Todechine [7].By inspection to these descriptors it was concluded that these descriptors can encode topological and electronic aspect of a molecules which are important on solute-solvent interactions.

Model evaluations

In spite of good accuracy and apparent mechanistic appeal, QSPR models should pass rigorous validation tests to be useful as reliable screening tools. Y-randomization test is a tool used in validation of QSAR models. In this test the performance of the original model in data description is compared to that of models built for permuted (randomly shuffled) response [26]. The Y-scrambling procedure was performed to ensure that there is not any chance correlation within the data matrix. The mean value of R^2 after 30 times Y-scrambling runs was0.129, which does not indicate the probability of a chance correlation. Leave one out cross validation (LOO) test are one method which frequently used to evaluate the robustness of QSPR models. The outcomes from these procedures are a cross validated correlation coefficient (R^2_{cv}) and standardized predicted error sum of squares (SPRESS), which are calculated according to the following equations:

$$\left(\sum (y_i-\hat{y}_i)^2\right)/\left(\sum (y_i-\bar{y})^2\right) \quad (8)$$

SPRESS = (9)

In the above expression, y is the mean of the experimental values, n is number of observations, and k is the number of descriptors in the model and $y_i$and $\hat{y}_i$ is experimental and predicted values of responses. The $\hat{y}$ 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. The leave-one-out crosses validation for MLR, LS-SVM, and ANN models. The values of R^2_{cv} and SPRESS for LS-SVM model are 0.881 and 0.405, respectively, while these values are R^2_{cv} =0.861 and SPRESS = 0.458 For ANN model and R^2_{cv} = 0.571 and SPRESS =0.824 test was performed on MLR model, respectively. These values indicate the reliability of obtained models. Also comparison between these values and also those other statistics table 3, indicate the superiority of LS-SVM model over other models. The value of RMSE for LS-SVM model was 0.4 which was lower than those obtained by Rytting et al. (RMSE≃0.5).

Conclusion

In the present study, a linear (MLR) and two nonlinear feature mapping method (LS-SVM and ANN) were used to develop some QSPR models for prediction of solubilities of 122 drug compounds in binary mixtures of water and PEG. The obtained statistical parameters of these models revealed that LS-SVM model was superior over other models, which showed that nonlinear modeling technique can successfully use to predict the solubilities of drug compounds. Descriptors that appeared in these models were electronic, geometrical, and topological descriptors that can encode features of solutes which were affected on their solubilities, including steric, dispersion, and hydrophobic interactions.

Reference

1. Faller, B.; Ertl, P., (2007), Computational approaches to determine drug solubility. Adv. Dru Deliv.Rev.,59, 533-54.
2. Duchowicz, P.; Talevi, R. A. L.; Bruno-Blanch, E.; Castro, E. A., (2008), New QSPR study for the prediction of aqueous solubility of drug-like compounds. Bioorg. Med. Chem., 16, 7944–7955.
3. Paruta, A. N.; Sciarrone, B. J.; Lordi, N. G., (1964), Solubility of salicylic acid as a functionof dielectric constant. J. Pharm. Sci., 53, 1349–1353.
4.]Huuskonen, J.; Livingstone, D. J.; Manallack, D. T., (2008), Prediction of drug solubility from molecular structure using a drug-like training set. SAR QSAR Environ. Res., 19, 191.
5. Rytting, E.; Lentz, K. A.; Chen, X. Q.; Qian, F.; Venkatesh, S., (2004), Quantitative Structure–Property Relationship for Predicting Drug Solubility in PEG 400/Water cosolvent systems. Pharm. Res., 21, 237-244.
6. Rytting, E.; Lentz, K. A.; Chen, X. Q.; Qian, F.; Venkatesh, S., (2005), Aqueous and cosolvent solubility data for drug-like organic compounds. Am. Assoc. Pharm. Sci. J.,7, 78-105.
7. Todeschini, R.; Consonni, V., (2000), Handbook of Molecular Descriptors; Wiley-VCH: Weinheim.
8. Hyper Chem Release 7.0 for windows; (2002), Hypercube: Inc.
9. http://www.disat.unimib.it/chem.
10. Todeschini, R.; Consonni, V., (2009), Molecular Descriptors for Chemo informatics; Wiley-VCH: Weinheim, Germany.
11. Fatemi, M. H.; Heidari, A.; Ghorbanzade, M., (2010), Prediction of Aqueous Solubility of Drug-Like Compounds by Using an Artificial Neural Network and Least-Squares Support Vector Machine. Bull. Chem. Soc.,83, 1338–1345
12. Zupan, J.; Gasteiger, J., (1993), Neural Networks for Chemists: An Introduction; Wiley-VCH: New York.
13. Zupan, J., Gasteiger, J., (1999), Neural Networks in Chemistry and Drug Design; Wiley-VCH: Weinheim.
14. http://www.statsoft.com/
15. Vapnik, V., (1995), The Nature of Statistical Learning Theory; Springer-Verlag: New York.
16. Vapnik,V., (1998), Statistical Learning Theory; John Wiley: New York.
17. Vapnik,V., 1998, The support vector methodof function estimation. Nonlinear Modeling: Advanced Black-Box Techniques; Suykens, J.A.K.,Vandewalle, J., Eds.; Kluwer Academic Publishers: Boston, pp. 55–85.
18. Sch¨olkopf, B.; Burges, C.J.C.; Smola, A.J., (1999), Advances in Kernel Methods: Support Vector Learning; The MIT Press: Cambridge.
19. Cristianini, N.; Taylor, J. S., (2000), An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods; Cambridge University Press: New York.
20. Suykens, J. A. K.; Vandewalle, J., (1999), Least Squares Support Vector Machine Classifiers. Neural.Process.Lett.,9, 293-300.
21. Suykens, J. A. K.; Brabanter, J. D.; Lukas, L.; Vandewalle, J., (2002), Weighted least squares support vector machines: Robustness and sparse approximation. Neurocomput.,48, 85-102.
22. Yap, C.; Li, H.; Ji, Z.; Chen, Y., (2007), Regression methods for developing QSAR and QSPR models to predict compounds of specific pharmacodynamic, pharmacokinetic and toxicological properties. Mini-Rev. Med. Chem., 7, 1097-107.
23. Sharma, BK.; Singh, P.; Prabhaka, r. Y. S., (2013), QSAR Rationale of Matrix Metalloproteinase Inhibition Activity in a Class of Carboxylic Acid. Based Compounds. Br. J. Pharm. Res., 3, 697-721.
24. Moriguchi, I.; Hirono, S.; Liu, Q.; Nakagome, I.; Matsushita, Y., (1992), Simple method of calculating Octanol/Water Partition Coefficient. Chem. Pharm. Bull.,40, 127-130.
25. Saiz-Urra. L;Gonzalez. MP; Teijeira, M., (2006), QSAR studies about cytotoxicity of benzophenazines with dual inhibition toward both topoisomerases I and II: 3D-MoRSE descriptors and statistical considerations about variable selection. Bioorg. Med. Chem., 14, 7347–7358.
26. Izadiyan, P.; Fatemi, M.; Izadiyan, M., (2013), Elicitation of the most important structural properties of ionic liquids affecting ecotoxicity in limnic green algae; a QSAR approach. Ecotoxicol. Environ Saf., 87, 42-48.