APPLICATION OF MULTIVARIATE IMAGE ANALYSIS IN QSPR STUDY OF pK_a OF VARIOUS ACIDS BY PRINCIPAL COMPONENTS-LEAST SQUARES SUPPORT VECTOR MACHINE

SOMAYEH VEYSEH ^a, HAMIDEH HAMZEHALI ^a, ALI NIAZI ^{a,*}, JAHAN B. GHASEMI ^b

^a Department of Chemistry, Faculty of Science, Arak Branch, Islamic Azad University, Arak, Iran

^b Department of Chemistry, Faculty of Science, K.N. Toosi University of Technology, Tehran, Iran

ABSTRACT

A new implemented quantitative structure-property relationships (QSPR) method, whose descriptors achieved from bidimensional images, was suggested for the predicting of acidity constant (pK_a) of various acid. The resulted descriptors were subjected to principal component analysis (PCA) and the most significant principal components (PCs) were extracted. Multivariate image analysis applied to QSPR modeling was done by means of principal component-least squares support vector machine (PC-LSSVM) methods. The resulted model showed high prediction ability with root mean square error of prediction of 0.0195 for PC-LSSVM.

Keywords: QSPR; PC-LSSVM; Multivariate image analysis; Acidity constant.

INTRODUCTION

Acidity constants are important parameters indicating the extent of ionization of molecules in solution at different pH values. The acidity constants of organic reagents play a fundamental role in many analytical procedures such as acid-base titration, solvent extraction, complex formation, and ion transport. It has been shown that acid-base properties affect the toxicity, chromatographic retention behavior, and pharmaceutical properties of organic acids and bases. Much of the theoretical foundation of modern organic chemistry is based on the observation of the effects on acid-base equilibrium of changing molecular structure¹.

Quantitative structure-activity/property relationship (QSAR/QSPR) studies, as one of the most important areas in chemometrics, give information that is useful for molecular design and medicinal chemistry²⁻⁵. QSAR/QSPR models are mathematical equations relating chemical structure to a wide variety of physical, chemical, biological and technological properties. Model development in QSAR/QSPR studies comprises different critical steps as (1) descriptor generation, (2) data splitting to calibration (or training) and prediction (or validation) sets, (3) variable selection, (4) finding appropriate model between selected variables and activity/property and (5) model validation.

Among the investigation of QSAR/QSPR, 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 regressions (MLR) and artificial neural network (ANN) have been used in QSAR/QSPR studies. However, the practical usefulness of MLR in QSAR/QSPR 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 applications⁶. Also, for simplification of traditional of SVM, Suykens and Vandewalle⁷ 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. Some application of LS-SVM in quantification and QSAR reported by different researchers in literature⁸⁻¹¹.

A major step in constructing the QSAR/QSPR models is finding one or more molecular descriptors that represent variation in the structural property of the molecules by a number. Different descriptors have been studied to be used in QSAR analysis. Nowadays, image analysis is becoming more important because of its ability to perform fast and non-invasive low-cost analysis on different processes in chemistry. Image analysis is a wide denomination that encloses classical studies on gray scale or (red-green-blue) RGB images¹². Esbensen and Geladi¹³ have demonstrated that image analysis may provide useful information in chemistry, through the descriptors do not have a direct physicochemical meaning, since they are binaries and also, Goudarzi and Goodarzi¹⁴ have studied the prediction of the acidic dissociation constant of some organic compounds using linear and nonlinear QSPR methods. In QSAR/QSPR, images (2D chemical structure) have shown to contain chemical information¹⁵⁻¹⁸, allowing the correlation between chemical structures and

properties.

The present paper is focused on the application of 2D images, which are the proper structures of the compounds that can be drawn with aid of any appropriate program, as descriptors in QSPR. Then, multivariate image analysis-quantitative structure property relationship study (MIA-QSPR) is proposed to model and predict the pK_a of a series of acids by principal component-least squares support vector analysis (PC-LSSVM) modeling method.

EXPERIMENTAL

Data set

The pK_a data of 28 acids were obtained from the literature¹⁹. The chemical name of these compounds and their pK_a has been listed in Table 1. In order to guarantee that training and prediction sets cover the total space occupied by the original data set, it was divided into parts of training and prediction set according the Kennard-Stones algorithm²⁰. The Kennard-Stones algorithm is known as one of the best ways of building training and prediction sets and it has been used in many QSAR/QSPR studies.

Hardware and software

The HP Personal Computer (2 GB RAM) equipped with the Windows Vista operating system and Matlab (Version 7.13, Mathwork Inc.) was used. Hyper Chem (version 7) and paint packages were used to draw the molecular structure. The LS-SVM optimization and model results were obtained using the LS-SVM lab toolbox (Matlab/ C Toolbox for Least-Squares Support Vector Machines).

Multivariate image analysis descriptors

In the MIA-QSPR method, the descriptors are the pixels of images that can be two or three dimensional. These pixels are correlated with dependent variables for making QSPR models. The 2D structures of each compound of Table 1 were systematically drawn in the HyperChem program by same of font and size, and then, converted to bitmaps in 100×120 pixels using paint package. All the drawn molecular structures were systematically fixed in a given coordinate. In our dataset, the pixel located at the 100×120 coordinate, was used as reference in the alignment step, as illustrated in Fig. 1. Pixels of each 2D image were read as binaries in Matlab. Each image of dimension 100×120 pixels was unfolded to a 1×12000 row and then the 28 images were grouped to form a 28×12000 matrix. Columns with zero variance were removed to minimize memory, reducing the size of matrix to 28×1056.

RESULTS AND DISCUSSION

Principal component analysis of data set

To reduce the dimensionality of the independent variable space, a limited number of principal components (PCs) are used. Also, PCA was performed on the bidimensional images descriptors to the whole data set of pK_a (Table 1), for investigation the distribution in the chemical space, which shows the spatial location of samples to assist the separation of data into training and prediction sets. The PCA results show that two PCs (PC1 and PC2) describe 71.25% of

the overall variances: PC1=52.17% and PC2=18.54% (Fig. 2). Since almost all variables can be accounted for the first two PCs, their score plot is a reliable presentation of the spatial distribution of the points for the data set. The data separation is very important in the development of reliable and robust QSPR models. The quality of the prediction depends on the data set used to develop the model. For regression analysis, data set was separated into two groups, a training set (22 data) and a prediction set (6 data) according to Kennard-Stones algorithm. As shown in Fig. 2, the distribution of the compounds in each subset seems to be relatively well-balanced over the space of the principal components.

Fig. 1. 2D images and unfolding step of the 28 chemical structures to give the X-matrix. The arrow in structure indicates the coordinate of a pixel in common among the whole series of compounds, used in the 2D alignment step.

Table 1: pK_a of some acids from literature⁽¹⁹⁾.

Acids	pK_a	Acids	pK_a
Sulfanilic acid	3.21	Salicylic acid	2.97
Aniline hydrochloride	4.10	3-Nitrobenzoic acid	3.40
Pyridine hydrochloride	5.20	1-Naphthoic acid	3.70
Ammonium perchlorate	9.20	Benzoic acid	4.20
<i>n</i> -Butylammonium perchlorate	10.6	Acetic acid	4.80
Tri- <i>n</i> -butylammonium perchlorate	10.9	Picric acid	0.25
Piperidine hydrochloride	11.12	2,4-Dinitronaphthol	3.50
Guanidine hydrochloride	13.54	2,6-Dinitrophenol	3.71
Trichloroacetic acid	0.66	2,4-Dinitrophenol	4.10
Dichloroacetic acid	1.30	4-Nitro-3-CF ₃ -phenol	6.40
2,4-Dinitrobenzoic acid	1.43	4-Nitrophenol	7.15
2,6-Dichlorobenzoic acid	1.60	2,3-Dihydroxy naphthalene	7.62
2-Nitrobenzoic acid	2.18	3-Nitrophenol	8.40
2,6-Dihydroxybenzoic acid	2.94	Pyrogallol	8.94

Fig. 2. Principal components analysis of the 2D image descriptors for the data set.

LSSVM analysis

After achieving PCs, PCs were used as the input to develop nonlinear model by LS-SVM. The quality of LS-SVM for regression depends 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 1000, respectively, with increment steps of 1. Table 2 shows the optimum γ and σ^2 parameters for the LS-SVM and RBF kernel, using the calibration sets for 22 acidity constant data.

Model validation and prediction of pK_a

The predictive ability of this method (PC-LSSVM) was determined using six pK_a data (Table 1). The results obtained by PC-LSSVM methods are listed in Tables 2.

Different statistical parameters have been used for the evaluation of the suitability of the developed model for prediction of the pK_a of the studied compounds this include the root mean square error of prediction (RMSEP) and relative standard error of prediction (RSEP) can be used:

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

$$RSEP(\%) = 100 \times \left(\frac{RMSEP}{y_{mean}} \right)$$

where $y_{i,pred}$ is the predicted pK_a using different model, $y_{i,obs}$ is the observed value of the pK_a , y_{mean} is the mean experimental pK_a values n is the number of molecule in the prediction set. Table 2 also shows RMSEP and RSEP for prediction of pK_a .

Table 2 also shows RMSEP, RSEP and the percentage error for prediction of pK_a . As can be seen, the percentage error was also quite acceptable for PC-LSSVM. Good results were achieved in PC-LSSVM model with percentage error ranges from -0.36 to 3.03 for pK_a . Other statistical parameters have been used for the evaluation of the suitability of the developed models for prediction of the pK_a of the studied compounds this include cross validation coefficient (Q^2 and R^2). These parameters are listed in Table 3 and show the good statistical qualities.

Table 3: Comparison of the statistical parameters by different QSAR models for the prediction of the pK_a .

Methods	Data set	R^2	Q^{2*}
LSSVM	Training	0.8926	0.8726
	Test	0.8624	0.8412
PC-LSSVM	Training Test	0.9912 0.9354	0.9872 0.9268

* Q^2 coefficient for the model validation by leave-one-out.

As the conclusion, the proposed PC-LSSVM algorithm produced perfect model for MIA-QSPR study of pK_a of some acids. Also, obtained results indicated that MIA descriptors are capable to recognize the physicochemical information and may be useful to predict pK_a . Also, it is possible to see that PC-LSSVM presents excellent prediction abilities when compared with LSSVM.

CONCLUSION

Using PC-LSSVM, a QSPR model for the prediction of pK_a for 28 compounds based multivariate image analysis alone have been developed. The results well illustrate the power of pixels in prediction of pK_a of various acids. The model could predict the pK_a of various acids not existed in the modeling procedure accurately. The results of this study clearly show the potential and versatility of PC-LSSVM modeling in QSPR study of pK_a of different acids using multivariate image analysis, which could be applied to prediction of pK_a .

ACKNOWLEDGMENTS

The authors are gratefully acknowledging the support of this work by the Islamic Azad University, Arak Branch.

Table 2: Experimental and predicted values of pK_a using LSSVM and PC-LSSVM models.

Acids	Actual pK_a	LSSVM		PC-LSSVM	
		Predicted pK_a	Error (%)	Predicted pK_a	Error (%)
Pyridine hydrochloride	5.20	5.03	-3.27	5.19	-0.19
Dichloroacetic acid	1.30	1.21	-6.92	1.32	1.54
2,6-Dihydroxybenzoic acid	2.94	2.79	-5.10	2.93	-0.34
Salicylic acid	2.97	2.81	-5.39	2.95	-0.67
Trichloroacetic acid	0.66	0.61	-7.58	0.68	3.03
3-Nitrophenol	8.40	8.11	-3.45	8.37	-0.36
γ		10		5	
σ^2		100		80	
RMSEP		0.1691		0.0195	
RSEP (%)		3.8277		0.4430	

REFERENCES

1. A. Niazi, A. Yazdanipour, J. Ghasemi, A. Amini, Sh. Bozorgzad, M. Kubista, Chem. Eng. Comm. 195, 1257 (2008).
2. A. Niazi, S. Jameh-Bozorgi, D. Nori-Shargh, Chin. Chem. Lett. 18, 621 (2007).
3. A. Niazi, S. Jameh-Bozorgi, D. Nori-Shargh, J. Hazard. Mater. 151, 603 (2008).
4. M. Sarkhosh, J. Ghasemi, M. Ayati, Chem. Central. J. 6, 1 (2012).
5. T. Asadollahi, Sh. Dadfarnia, A.M. Haji Shabani, J. Ghasemi, M. Sarkhosh, Molecules 16, 1928 (2011).
6. J.M. Prats-Montalban, A. de Juan, A. Ferrer, Chemometr. Intell. Lab. Syst. 107, 1 (2011).
7. J.A.K. Suykens, J. Vandewalle, Neural Process Lett. 9, 293 (1999).
8. A. Niazi, J. Zolgharnein, S. Afiuni-Zadeh, Anal. Sci. 23, 1311 (2007).
9. A. Niazi, J. Ghasemi, A. Yazdanipour, Spectrochim Acta Part A 68, 523 (2007).
10. J. Li, H. Liu, Z. Yao, M. Liu, Z. Hu, B. Fan, Anal. Chim. Acta 581, 333(2007).
11. Y. Xu, S. Zomer, R. Brereton, Crit. Rev. Anal. Chem. 36, 177 (2006).
12. J.M. Prats-Montalban, A. de Juan, A. Ferrer, Chemometr. Intell. Lab. Syst. 107, 1 (2011).
13. K. Esbensen, P. Geladi, J. Chemometr. 7, 67 (1989).
14. N. Goudarzi, M. Goodarzi, Molecular Phys. 107, 1495 (2009).
15. J.M. Prats-Montalban, A. de Juan, A. Ferrer, Chemometr. Intell. Lab. Syst. 107, 1 (2011).
16. M.P. Freitas, Org. Biomol. Chem. 4, 1154 (2006).
17. M.P. Freitas, S.D. Brown, J.A. Martins, J. Mol. Struct. 738, 149 (2005).
18. Z. Garkani-Nejad, M. Poshteh-Shirani, Talanta 83, 225 (2010).
19. H.S. Kim, T.D. Chung, H. Kim, J. Electroanal. Chem. 498, 209 (2001).
20. R.W. Kennard, L.A. Stones, Technometrics 11, 137 (1969).