QSPRS ON AQUEOUS SOLUBILITY OF PAHS USING GA-SVM, GA-BPNN AND GA-PLS

Jun Qi, Junfeng Niu

State Key Laboratory of Water Environment Simulation, School of Environment, Beijing Normal University, Beijing 100875, P. R. China

Abstract

A modified method to develop quantitative structure-property relationship (QSPR) models of organic contaminants was proposed based on genetic algorithm (GA) and support vector machine (SVM) (GA-SVM). GA was used to perform the variable selection, and SVM was used to construct QSPR models. In this study, GA-SVM was applied to develop the QSPR models for aqueous solubility (S_w) of polycyclic aromatic hydrocarbons (PAHs). The R^2 (0.978), *SSE* (3.032), and *RMSE* (0.257) values of the model developed by GA-SVM indicated a good predictive capability for logS_w values of PAHs. Based on external validation, the results of GA-SVM were compared with those of genetic algorithm-Back-Propagation Neural Network (GA-BPNN) and genetic algorithm-Partial Least-Squares Regression (GA-PLS). The comparison showed that the R^2 (0.952) and *SSE* (0.380) values of GA-SVM were highest and lowest, respectively, which illustrated GA-SVM was the optimal to develop QSPR models for the logS_w values of PAHs among the three methods.

Introduction

Polycyclic aromatic hydrocarbons (PAHs) are organic contaminants which widely exist in the environment[1]. As their chemical stability, forceful carcinogenic, spermatogenetic and mutagenic effects, some PAHs are listed in *Water-Related Environmental Fate of 129 Priority Pollutants* by US EPA[2]. Aqueous solubility (S_w) of PAHs to a great extent can determine the distribution and accumulation in air, water, soil and living organisms, as well as migration velocity and degradation rates in the environment[3], Therefore, S_w is one of the most important parameters which need to be measured. However, because of large expenditures of money, time and equipment, measured S_w data for PAHs were rather scarce. Thus a great deal of effort had been put into attempting the estimation of S_w through statistical modeling, and a variety of Quantitative Structure-Property Relationship (QSPR) models were proposed[4-11].

Ordinary least squares regression (OLS), partial least-squares regression (PLS) and artificial neural network (ANN) were common methods for constructing QSPR models[7-9]. Recently, Support Vector Machine (SVM)[10,11] and some combination methods (e.g. GA-PLS, GA-ANN) were widely applied to develop QSPR models. As SVM did outstanding work in small sample, nonlinearity, high dimensional regression problems, it was successfully applied to develop QSPR models in many fields[10,11]. Moreover, GA-PLS and GA-ANN likewise showed fine performance in different QSPR cases[7,13,14]. Based on probabilistic choice, Genetic Algorithm (GA) has powerful global search capability, and was used for the variable selection[15,16]. However, the study on using GA to select variables for SVM, and using GA-SVM to develop QSPR models was scarce till now. In this work, GA-SVM, GA-BPNN and GA-PLS were applied to develop the QSPR models for logS_w of 47

PAHs respectively, and the results of the models constructed by three methods were compared based on Leave-n (about 10%)-out (LNO) cross validation.

Materials and Methods

47 PAHs with 2 to 7 rings were selected as samples, in which 16 PAHs have been listed as priority control pollutants by the Environmental Protection Agency of the USA[17]. The logS_w values of the 47 PAHs were quoted from literatures[3,18-21] and used as dependent variable matrix Y. Among most QSPR models for calculating S_w of PAHs, the model with molecular connectivity indices is applied widely. It is regarded as an accurate and advisable method[18]. Thus, the programs to calculate molecular connectivity indices were developed in MATLAB. 11 molecular connectivity indices (${}^0\chi^v$, ${}^1\chi^v$, ${}^2\chi^v$, ${}^3\chi_p^v$, ${}^4\chi_p^v$, ${}^5\chi_p^v$, ${}^6\chi_p^v$, ${}^3\chi_c^v$, ${}^4\chi_{pc}^v$, ${}^5\chi_{pc}^v$ and ${}^6\chi_{pc}^v$) of 47 PAHs were calculated and used as independent variable matrix X.

In this study, GA was used to select features for SVM, BPNN and PLS. 300 individuals were contained in the population; every individual contained 11 Binary codes; and every code represented a feature. The feature would be selected if the corresponding binary code was 1. Contrarily, the feature would not be selected if the corresponding binary code was 0. All selected features by an individual were retained to establish a new matrix, using it, a QSPR model could be developed by SVM (BPNN or PLS). The reciprocal of correlation coefficient between measured values and predicted values of the model was the fitness value of this individual. After 100 generation, the optimal QSPR model was constructed. This is main idea of GA-SVM, GA-BPNN and GA-PLS. The flow diagram of GA-SVM, GA-BPNN and GA-PLS was shown as Fig. 1, where *Gen* is the the generational counter. The Libsvm toolbox[22], Genetic Algorithm Toolbox[23] and ANN Toolbox[24] of MATLAB were applied in this study, and the parameters were listed in Table 1.

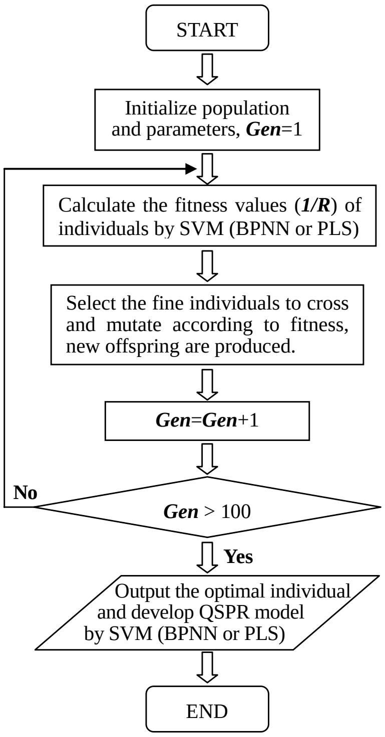

START
Initialize population and parameters, Gen=1
Calculate the fitness values (1/R) of individuals by SVM (BPNN or PLS)
Select the fine individuals to cross and mutate according to fitness, new offspring are produced.
Gen=Gen+1
No
Gen > 100
Yes
Output the optimal individual and develop QSPR model by SVM (BPNN or PLS)
END

Figure 1 Flow diagram of GA-SVM.

Table 1 The parameters of GA-SVM, GA-BPNN and GA-PLS

generation gap	crossover rate	mutation probability	kernel function	termination criterion tolerance	C	ε in loss function	input layer Nodes	hidden layer Nodes	output layer Nodes	error goal	epochs for stopping	learning rate	principal components number
0.9	0.5	0.3	RBF	0.1	100	0.001	m (the number of selected features)	m+6	1	10^{-5}	5000	0.05	k increased from 1 until the cumulative contribution rate achieve above 90%

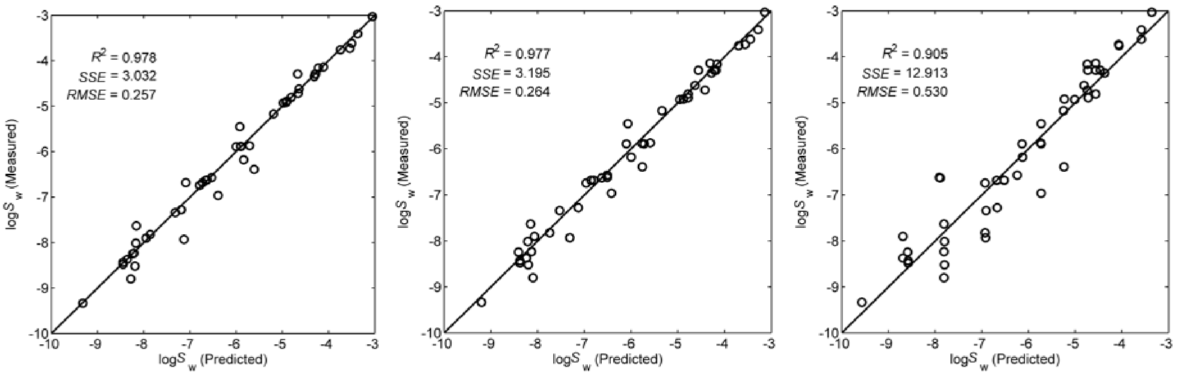
Results and discussion

Table 2 Optimal individuals and errors of 3 QSPR models constructed by GA-SVM, GA-BPNN and GA-PLS respectively

Methods	Optimal individual											R^2	*SSE*	*RMSE*
	$^0\chi^v$	$^1\chi^v$	$^2\chi^v$	$^3\chi_p^v$	$^4\chi_p^v$	$^5\chi_p^v$	$^6\chi_p^v$	$^3\chi_c^v$	$^4\chi_{pc}^v$	$^5\chi_{pc}^v$	$^6\chi_{pc}^v$			
GA-SVM	0	1	0	0	1	1	1	0	1	1	1	*0.978*	*3.032*	*0.257*
GA-BPNN	0	1	0	1	0	1	1	1	1	0	1	0.977	3.195	0.264
GA-PLS	0	1	0	0	0	0	0	0	0	0	0	0.905	12.919	0.530

Using logS_w values and molecular connectivity indices of 47 PAHs as dependent variable and independent variables, 3 QSPR Models were developed by GV-SVM, GA-PLS and GA-BPNN, respectively. The optimal individuals and the errors between predicted values and measured values were listed in Table 2. Furthermore, the correlated plots of predicted and measured logS_w values of PAHs by GA-SVM, GA-BPNN and GA-PLS were shown in Fig. 2.

Seven molecular connectivity indices (i.e. $^1\chi^v$, $^4\chi_p^v$, $^5\chi_p^v$, $^6\chi_p^v$, $^4\chi_{pc}^v$, $^5\chi_{pc}^v$ and $^6\chi_{pc}^v$) were selected by the model constructed by GA-SVM. As the R^2 values between measured and predicted logS_w values of every groups were higher than 0.97, indicating that the correlation was significant. Moreover, the *SEE* and *RMSE* values were lower than 3.10 and 0.26, respectively. Therefore, the QSPR model developed by GA-SVM has good predictive capability.



(a) (b) (c)

Figure 2 The correlated plots of predicted and measured logS_w values of PAHs: (a) The correlated plots of the QSPR model developed by GA-SVM. (b) The correlated plots of the QSPR model developed by GA-BPNN. (c) The correlated plots of the QSPR model developed by GA-PLS.

A linear QSPR model was developed by GA-PLS as Eq. (1).

$$\log S_w = -1.6178\,^1\chi^v - 6.1249 \quad (1)$$

As the model contained only one variable, it was simpler than the models constructed by GA-SVM and GA-BPNN. However, the R^2 values between measured and predicted logS_w values was 0.91, the *SSE* and *RMSE* values were 12.92 and 0.53 respectively. Hence, the predictive capability of the model was not satisfactory.

According to Table 2, seven molecular connectivity indices ($^1\chi^v$, $^6\chi_{pc}^v$, $^5\chi_p^v$, $^6\chi_p^v$, $^3\chi_c^v$, $^4\chi_{pc}^v$ and $^3\chi_p^v$) were selected in the GA-BPNN QSPR model. As R^2 value was 0.98, the correlation between measured and predicted logS_w values was very significant. In addition, *SSE* (3.20) and *RMSE* (0.26) values showed that the predictions of QSPR models constructed by GA-BPNN were reasonable, and GA-BPNN was a suitable method to develop QSPR models for logS_w of PAHs.

Leave-n (about 10%)-out (LNO) cross validation was employed to compare the predictive capability and stability of the 3 methods. 9 groups of trainings and forecasts were carried on respectively. 47 PAHs are divided into 9 groups randomly, each of the first 8 groups contained 5 PAHs and the 9th group contained 7 PAHs. The steps of validation were briefly described as follows:

Step 1: Select group 1 as predicted data, the others are training data. Using GA-SVM, a QSPR model can be developed and validated.

Step 2: Choose another group as predicted data and repeat the calculations of Step 1. 9 QSPR models can be developed by GA-SVM finally.

Step 3: Use GA-PLS and GA-BPNN to develop QSPR models, repeat the processes of Steps 1 and 2.

The LNO cross validation discussed above was carried out, and the errors of three methods were calculated and vividly shown by Fig 3. As the correlation of GA-SVM results was more significant than those of GA-BPNN and GA-PLS (R^2>0.94, *SSE*<6.65, *RMSE*<0.39), the model developed by GA-SVM has highest predictive ability and robust stability. Consequently, GA-SVM is the optimal method among three methods in this study.

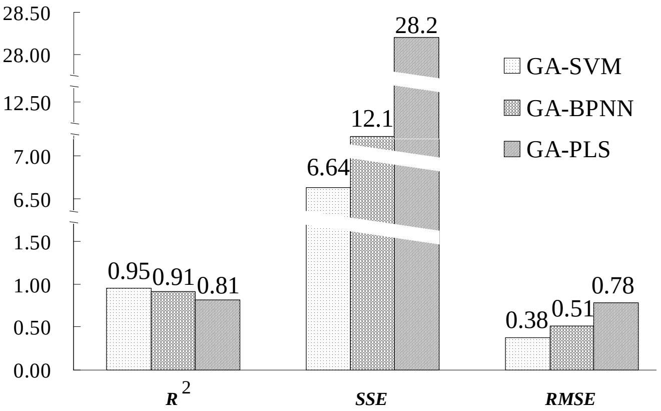


Figure 3 Comparison plots of mean R^2, *SSE*, and *RMSE* between predicted and measured logS_w values from QSPR models constructed by GA-SVM, GA-BPNN and GA-PLS respectively during LNO cross validations.

Acknowledgements

The research was supported by the National Basic Research Program of P.R. China (973 Project, 2003CB415204).

References

1. Nemr AE, Aly MA. *Chemosphere* 2003; 1711:1716.
2. Callahan MA, Slimak MW, Gabel NW, May IP, Fowler CF, Freed JR, Jennings P, Durfree RC, Whitmore FC, Maestri B, Mabley WR, Holt BR, Gould C. *Water Related Environmental Fate of 129 Priority Pollutants, Vol II. Halogenated Aliphatic Hydrocarbons. Halogenated Esters, Monocyclic Aromatics, Phthalate Esters, Polycyclic Aromatic Hydrocarbons, Nitrosamines and Miscellaneous Compounds*. EPA-440/4-79-029 b 1979.
3. Wang LS, Han SK. *Organics Quantitative Structure-Activity Relationships*, Chinese Environmental Science Press, Beijing, 1993; 429:454.
4. Niu JF, Huang LP, Chen JW, Yu G, Schramm KW. *Chemosphere* 2005; 917:924.
5. Chen JW, Xue XY, Schramm KW, Quan X, Yang FL, Kettrup A. *Comput. Biol. Chem.* 2003; 165:171.
6. Ding GH, Chen JW, Qiao XL, Huang LP, Lin J, Chen XY. *SAR QSAR Environ. Res.* 2005; 301:312.
7. Gramatica P, Giani E, Papa E. *J. Mol. Graphics Modell.* 2007; 755:766.
8. Niu JF, Yang ZF, Shen ZY, Wang LL. *SAR QSAR Environ. Res.* 2006; 173:182.
9. Devillers J, Domine D, Guillon C. *Eur. J. Med. Chem.* 1998; 659:664.
10. Liu HX, Yao XJ, Zhang RS, Liu MC, Hu ZD, Fan BT. *Chemosphere* 2006; 722:733.
11. Yao XJ, Panaye A, Doucet JP, Chen HF, Zhang RS, Fan BT, Liu MC, Hu ZD. *Anal. Chim. Acta*. 2005; 259:273.
12. Vapnik VN. *The nature of statistical learning theory*, Springer, New York, 1995; 181:190.
13. Fernández M, Caballero J. *J. Mol. Graph. Model.* 2006; 410:422.
14. Modarresi H, Dearden JC, Modarress H. *J. Chem. Inf. Model*. 2006; 930:936.
15. Nicolotti O, Carotti A. *J. Chem. Inf. Model*. 2006; 264:276.
16. Cho SJ, Hermsmeier MA. *J. Chem. Inf. Comput. Sci.* 2002; 927:936.
17. Manoli E, Samara C, Konstantinou I, Albanis T. *Chemosphere*. 2000; 1845:1855.
18. Yaws CL. *Chemical properties handbook*. McGraw2Hill Book Co, New York, 1999; 383:387.
19. Schirmer K, Chan AGJ, Greenberg BM, Dixon DG, Bols NC. *Toxicology* 1998; 143:155.
20. Ferreira MMC. *Chemosphere* 2001; 125:146.
21. SRC PhysProp Database. Available online at: http://esc.syrres.com.
22. Chang CC, Lin CJ. LIBSVM (Version 2.82) – A Library for Support Vector Machines. Available online at: http://www.csie.ntu.edu.tw/~cjlin/libsvm/, 2006.
23. Andrew C, Peter F, Hartmut P. *Genetic Algorithm Toolbox (Version 1.2) User's Guide*. Department of Automatic Control and Systems Engineering, University of Sheffield, 1994; 6:53.
24. Howard D, Mark B. *Neural-network toolbox. User's Guide*. 4th (ed.) The Mathworks Inc, 2001; 120:192.