Estimation of Octanol/Water Partition Coefficient and Aqueous Solubility of Environmental Chemicals Using Molecular Fingerprints and Machine Learning Methods

Qingda Zang (1), Kamel Mansouri (1), Richard S. Judson (2)

(1) ORISE Postdoctoral Fellow at the U.S. EPA, Research Triangle Park, NC, USA, (2) National Center for Computational Toxicology, U.S. EPA, Research Triangle Park, NC, USA

Richard Judson | judson.richard@epa.gov | 919-541-3085

www.epa.gov

Abstract

Novel methods are presented for the estimation values of the octanol/water partition coefficients (log P) and aqueous solubility (log S) of environmentally interesting chemicals solely based upon simple binary molecular fingerprints on a single data set which consists of 993 training samples and 251 test samples. A group of quantitative structure-property relationship (QSPR) models were developed using four approaches with different complexity: multiple linear regression (MLR), random forest (RF) regression, partial least squares regression (PLSR), and support vector regression (SVR). Genetic algorithms (GA) and RF method were employed to select the most information-rich subset of descriptors. It was found that MLR, PLSR and SVM exhibited satisfactory predictive results with low prediction errors and substantially outperformed RF. MLR coupled with GA for descriptor selection was clearly superior to all other approaches and achieved correlation coefficients of 0.936 and 0.927 between the calculated and experimental data on the validation set for log P and log S, respectively. The present study demonstrates that molecular fingerprints are very useful descriptors, GA is a very efficient feature selection tool and the selected descriptors can effectively model the two properties, and simple methods such as MLR give better results than more complicated methods. These models can be used for rapidly and accurately predicting log P and log S of environmental chemicals.

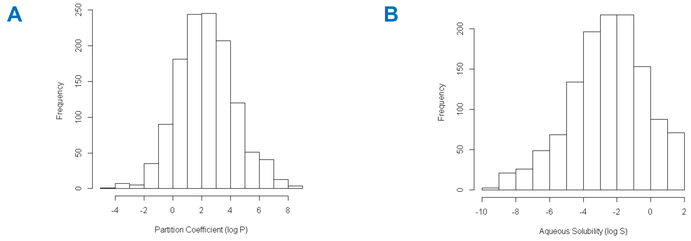

Figure 1. Data distribution of log P (A) and log S (B).

Table 1. Summary Statistics for Training (993 Samples) and Test (251 Samples) Sets

Property		Minimum	Maximum	Mean	Median	Standard Deviation
Log P						
	Training	-4.27	8.54	2.29	2.18	1.98
	Test	-3.89	8.39	2.39	2.29	2.03
Log S						
	Training	-9.70	1.58	-2.54	-2.38	2.24
	Test	-9.21	1.57	-2.58	-2.39	2.28

Methods

Mathematical processing for data standardization, multivariate regression analysis, and statistical model building were performed using the *R* statistical computing environment for Windows (version 2.15.1). Genetic algorithms, random forests, multiple linear regression, partial least squares regression and support vector regression were implemented by the packages *subselect*, *randomForest*, *stats*, *pls* and *e1071*, respectively.

Feature Selection – Random Forests

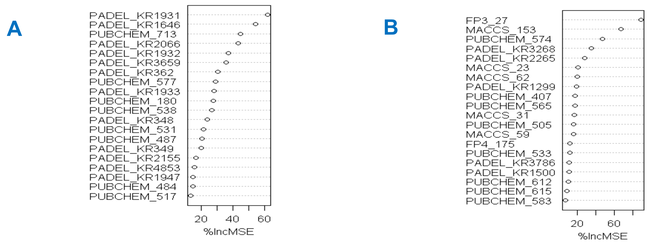

Figure 2. The top 20 fingerprints ranked by random forest (RF) feature selection for log P (A) and log S (B).

The Relationship of log S with log P and Mw

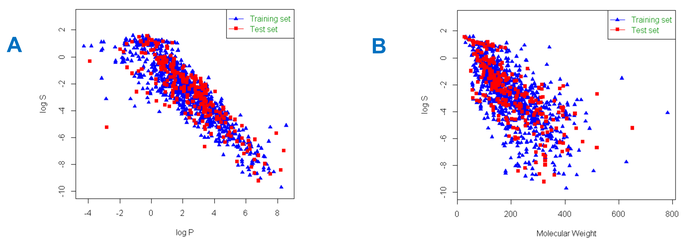

Figure 3. Aqueous solubility (log S) versus partition coefficient (log P) (A) and molecular weights (Mw) (B).

$$\log \text{Property} = \sum_{j=1}^{m} c_j f_j \qquad \log S = \sum_{j=1}^{m} c_j f_j + c_m Mw + c_p \log P$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (p_i - \hat{p}_i)^2}{\sum_{i=1}^{n} (p_i - \bar{p})^2} \qquad RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (p_i - \hat{p}_i)^2}$$

R^2: Correlation Coefficient; RMSE: Root Mean Squared Error.

Multiple Linear Regression

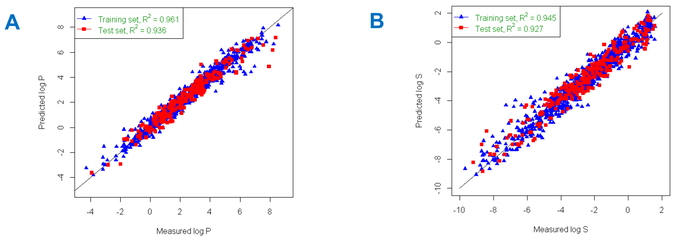

Figure 4. Plots of estimated values versus experimental values for the training and test sets of log P (A) and log S (B).

Partial Least Squares Regression

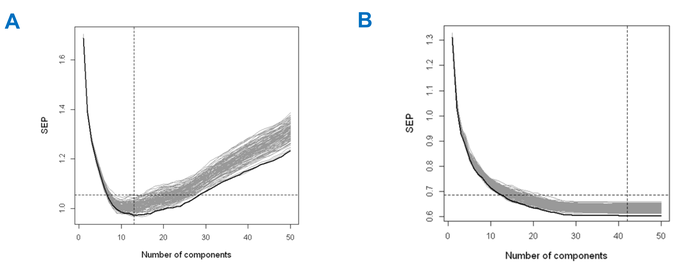

Figure 5. The relationship between the number of principal components (PCs) and the standard error of prediction (SEP) for the log P models of all fingerprints (A) and 250 fingerprint bits selected by GA. Black: a single of 10-fold CV; Gray: 100 repetitions of the 10-fold CV.

Results

Table 2. Comparison of the Best Models from the Four Methods for the Test Set

Method		MLR	PLSR	SVM	RF
Log P					
	R^2	0.936	0.936	0.915	0.835
	RMSE	0.492	0.495	0.535	0.666
Log S					
	R^2	0.927	0.924	0.901	0.839
	RMSE	0.588	0.597	0.653	0.777

Conclusions

The results demonstrated that excellent prediction performance was achieved under optimal conditions and the estimated values highly correlated with experimental values. Overall, there are multiple ways for deriving regression models with similar statistics.