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[1]

[1]NCCT/ORD/USEPA, RTP, NC 27711

Octanol/water partition coefficient (logP) and aqueous solubility (logS) are two important parameters in pharmacology and toxicology studies, and experimental measurements are usually time-consuming and expensive. In the present research, novel methods are presented for the estimation of both logP and logS 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 for obtaining reliable and robust regression models with high prediction performance. 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 logP and logS, respectively. The inclusion of logP and molecular weights (MW) as two descriptors into logS models significantly improved the prediction accuracy, especially for RF modeling. 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 logP and logS of environmental chemicals. *This abstract does not necessarily reflect U.S. EPA policy.*