In Silico Prediction of Physicochemical Properties of Environmental Chemicals in Combination of Molecular Fingerprints and Machine Learning Approaches

Q Zang[1], K Mansouri[2], D Allen[1], N Kleinstreuer[1], W Casey[3], R Judson[2]

[1]ILS/NICEATM, RTP, NC, USA; [2]EPA/ORD/NCCT, RTP, NC, USA; [3]NIH/NIEHS/DNTP/NICEATM, RTP, NC, USA

Quantitative structure-property relationship (QSPR) models were developed for the prediction of six physicochemical properties of environmental chemicals: octanol/water partition coefficient (log P), water solubility (log S), boiling point (BP), melting point (MP), vapor pressure (VP), and bioconcentration factor (BCF). Models were developed using simple binary molecular fingerprints and four approaches with differing complexity: multiple linear regression, random forest regression, partial least squares regression, and support vector regression (SVR). To obtain reliable and robust regression models with high prediction performance, genetic algorithms (GA) were employed to select the most information-rich subset of fingerprint bits. Predictions from the various models were tested against a validation set, and all four approaches exhibited satisfactory predictive results, with SVR outperforming the others. BP was the best-predicted property with a correlation coefficient (R^2) of 0.95 between the estimated values and experimental data on the validation set while MP was the most poorly predicted with an R^2 of 0.84. The statistics for other properties were intermediate between MP and BP with R^2 equal to 0.94, 0.93, 0.92 and 0.86 for log S, log P, VP and BCF, respectively. The prediction results for all properties were superior to those from Estimation Program Interface (EPI) Suite (R^2 values ranged from 0.63 to 0.94), a widely used tool for property prediction. This study demonstrates that (1) molecular fingerprints are useful descriptors, (2) GA is an efficient feature selection tool from which selected descriptors can effectively model these properties, and (3) simple methods give comparable results to more complicated methods. *This project was funded in whole or in part with Federal funds from the NIEHS, NIH under Contract No.HHSN27320140003C and does not represent EPA or NIEHS policy.*