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

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), 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.

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).

Results

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

<table>
<thead>
<tr>
<th>Method</th>
<th>Log P</th>
<th>Log S</th>
</tr>
</thead>
<tbody>
<tr>
<td>R²</td>
<td>0.936</td>
<td>0.927</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.492</td>
<td>0.588</td>
</tr>
<tr>
<td>SVM</td>
<td>0.955</td>
<td>0.839</td>
</tr>
<tr>
<td>RF</td>
<td>0.915</td>
<td>0.653</td>
</tr>
</tbody>
</table>

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.

Disclaimer: The views expressed in this presentation are those of the authors and do not necessarily reflect the views or policies of the U.S. EPA