Local Modeling for Estimating Partition Coefficients of Organic Compounds

Martin Junghans, Ernö Pretsch
Department of Organic Chemistry, Swiss Federal Institute of Technology (ETH), CH-8092 Zürich

A structure vector based on a combination of interatomic distances in 3D structures and substructure coding is calculated. Subsequent PLS analysis is applied to generate structure descriptors. These are then clustered with an agglomerative clustering algorithm with complete-linkage. For each cluster, distinct structure descriptors are generated by individual PLS analyses and used to build local linear models. From a library containing 245 molecules, 239 (98 %) pertained to 9 clusters with 5 to 104 members. The local linear models allowed the partition coefficient, log K_{OW}, to be estimated with RMS errors between 0.036 and 0.274 and maximal errors between 0.01 and 0.61. The overall RMS error for the 239 estimations was 0.095. Compared with the two global models based on structure descriptors and on parameters from AM1 calculations [1], the use of local models reduced the RMS error by a factor of 1.9 and 3.1, respectively.
For cross validation, a randomly selected training set of 123 molecules was used to build five linear models derived from its clusters. The test set with the remaining 122 molecules gave an RMS error of 0.225 and a maximal error of 0.84 for 67 % of predictable log K_{OW} values.

[1] N. Bodor and M.-J. Huang, J. Pharm. Sci. 81 (1992) 272-281

Representation in the Database

The 245 molecules were represented by connection tables containing all non-hydrogen atoms and their 2D drawing coordinates. Aromatic bonds were stored as alternating single and double bonds.

3D Path-Counting

For each pair of atoms, the connecting bonds were counted, sorted according to the number of bonds and kind of atoms linked. The Euclidean distance between them divided by the number of bonds was added to obtain the structure vector.

Substructure Coding

The set of potentially overlapping substructures used contained mainly functional groups. The number of substructures was determined for each molecule and added to give the substructure vector.

Collection of Information and PLS

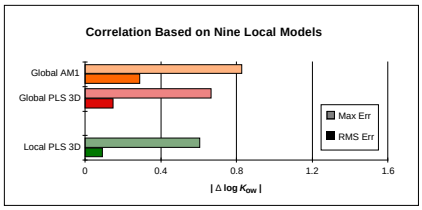
3D and SUB matrices were written containing the structure and substructure vectors of all molecules. Columns without variance were removed.
The 3D matrix was reduced to the first 32, the SUB matrix to the first 16 principal components by a separate PLS analysis.

Clustering of PLS Score Vectors

The score vectors were combined and a agglomerative clustering with complete-linkage was applied in a 48-dimensional structure space. Distance restrictions were chosen so as to maximize the number of molecules falling into a given cluster without losing quality because of clustering outliers.

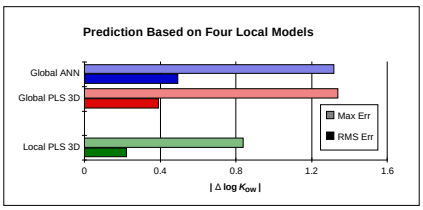
Reduction of Dimensionality

For each cluster, 3D and SUB matrices were concatenated and their principal components calculated each with a separate PLS analysis.



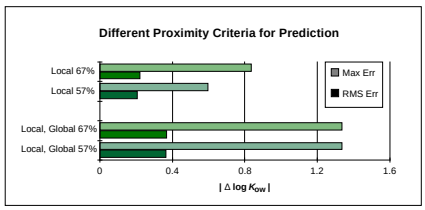
Correlation Based on Nine Local Models

For each cluster (shown in the Figure below), an MLR was performed and the principal components leading to the best fit were selected. An F test with a statistical reliability of 95% was used as a criterion to avoid overestimation of the linear model.
The results given in the diagram above refer to the partition coefficients of 98% of clustered molecules and were estimated using one of nine local models.



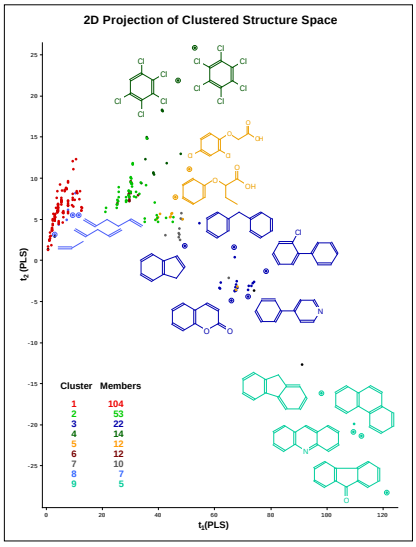
Prediction Based on Four Local Models

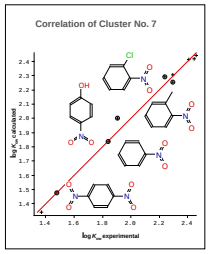
The data set was split randomly into a training set of 123 molecules and a test set of 122 molecules.
The training set was clustered using the same methods and distance restrictions as with the complete data set. Thus, 95% of the molecules fell into five clusters. For each cluster, a linear model was built leading to estimations comparable with those of the complete data set.
When projecting the test set into the PLS space, 67% of the molecules were found to be near four out of the five clusters. The resulting four regression models were then used to predict the partition coefficients.

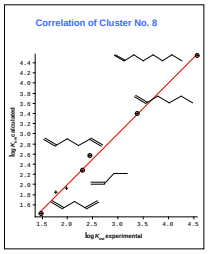


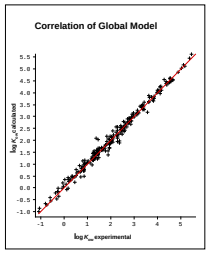
Different Proximity Criteria for Prediction

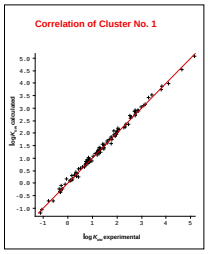
The decision on the proximity of a projected molecule to a cluster was based on the average distance of molecules in the cluster of interest. A less restrictive decision leads to higher RMS and maximal errors but, at the same time, to a high percentage of predicted molecules. On the other hand, by deciding more restrictively lower errors are found but also a low percentage of prediction. Comparison of the results given above for 67% of prediction with those for 57% shows that the RMS error was reduced by 7% only, whereas the maximal error dropped by ca. 20% (log K_{ow} from 0.84 to 0.60).
When completing the missing values with predictions from the global model, comparable results were found for both levels of decision.











Topological vs. AM1-Based Description

By correlating log K_{ow} with topological descriptors, the RMS error was reduced by 48% and the maximal error by 19% relative to the AM1-based values [1]. In addition, no complicated quantum mechanical calculations were necessary.

Global vs. Local Models

For correlation and prediction, the RMS errors decreased by about 40% with local models. While the local method could be used to fit the log K_{ow} values of 98% of the molecules, only those of 67% were predictable with good results. When predicting these values for the 122 chosen molecules, using the global model for the missing ones, the RMS error is reduced by only 6%, but it is thus possible to define two thirds of the values as reliable or interpolated and one third as extrapolated.

Examination of the Regression Plots

The upper figures on the left show the regression plots for the clusters with nitroaromatics and with alkenes. Both have low RMS and maximal errors. Although log K_{ow} values for 1–chloro- and 3-hydroxynitrobenzene as well as for 1,4-pentadiene are poorly estimated in the global model (errors from 0.4 to 0.5), their estimations in the local models are excellent (errors from 0.04 to 0.09).
The lower plots on the left show the correlations of log K_{ow} values for the complete data set with an RMS error of 0.154 as well as for cluster no. 1 with an RMS error of 0.082. Both linear models cover a range of over 6 orders of magnitude in log K_{ow}. This indicates that not only a reduction in range (as with the upper plots) but actually the distinction between different chemical classes leads to a reduction in RMS and maximal errors.

Conclusions

The use of local models not only reduces the RMS error and the maximal errors of correlated and predicted values but also makes it possible to decide whether a predicted value is extrapolated or not. This last point seems to be the strongest argument in favor of local-model building because it allows to classify any molecule in terms of its similarity to, or difference from, those of the database used to build the prediction model.
In combination with a global model, less reliable approximations can be obtained for the values of molecules not predictable from local models.