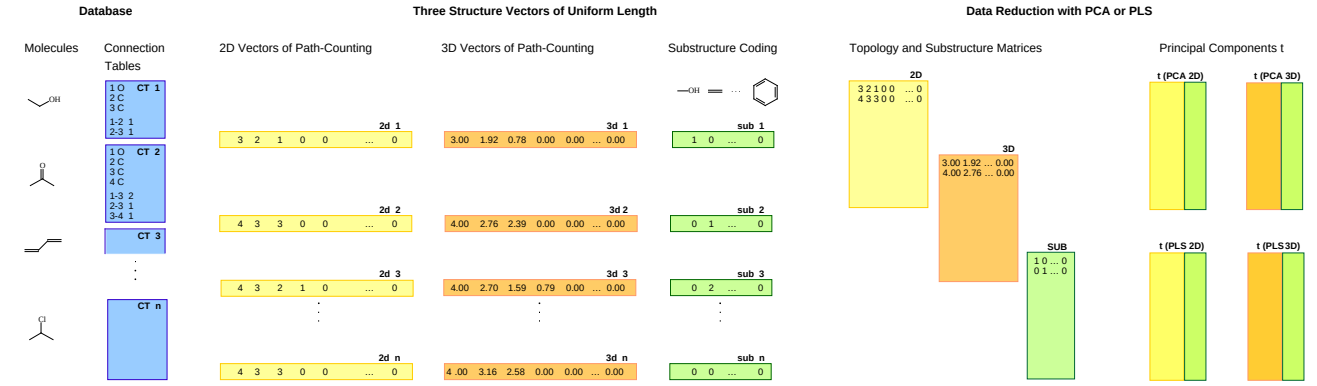
Uniform Topology-Based Structure Descriptor Combined with Substructure Coding for Estimating Partition Coefficients of Organic Compounds

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

Various structure descriptors of uniform length, based on path-counting of a node-colored molecular graph [1] or on interatomic distances in 3D structures, as well as substructure coding have been investigated in view of predicting partition coefficients, K_{ow}. The dimensions of the structure vectors generated in a first step have been reduced by PCA or PLS. A set of 245 molecules (from CH_3NH_2 to $C_{18}H_{12}$) with experimental log K_{ow} values were used for model building. With the best linear model, log K_{ow} was estimated with an RMS error of 0.153 and a maximal error of 0.67. Surprisingly, only minor improvements were achieved by including 3D information. A feed-forward back-propagation artificial neural network optimized for the above problem with 19 input and 20 hidden nodes and 1 output node was somewhat less powerful.
Cross validation using 123 randomly selected molecules as training set and the remaining 122 molecules as test set allowed predictions with an RMS error of 0.397 and a maximal error of 1.34 in log K_{ow} for the best linear model.

[1] J.-T. Clerc and A. L. Terkovics, Anal. Chem. Acta 253 (1990) 93-102
[2] J. Sadowski and J. Gasteiger, Chem. Rev. 93, (1993) 2567-2581



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.

2D Path-Counting

For each pair of atoms, the connecting bonds were counted, sorted according to the number of bonds and kind of atoms linked, and added to obtain the topological structure vector [1].

3D Path-Counting

Based on the connection tables and the 2D drawing coordinates of the molecules, low energy 3D conformations were calculated using the rule-based program ALCOGEN [2]. Path-counting was modified by adding the Euclidean distance between pairs of atoms divided by the number of bonds for each path.

Substructure Coding

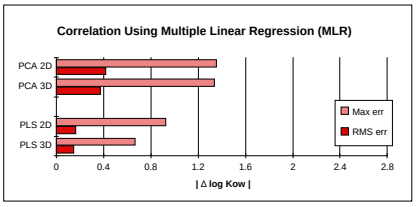
The set of potentially overlapping substructures used contained mainly functional groups. The number of substructures was determined for each molecule and added to give the substructure vector.

Collection of Information

2D, 3D and SUB matrices were written containing the structure and substructure vectors of all molecules. Columns without variance were removed.

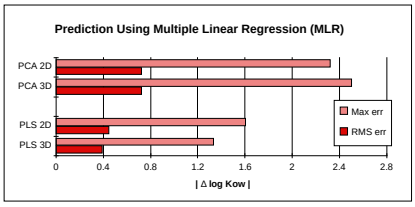
Reduction of Dimensionality

The 2D and 3D matrices were reduced to the first 32, the SUB matrix to the first 16 principal components by a separate PCA or PLS analysis. The latter also used the vector containing all experimentally determined partition coefficients.



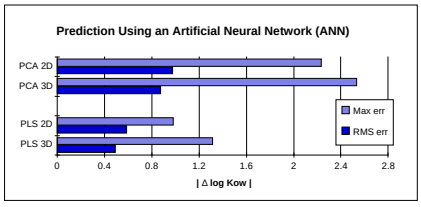
Correlation Using Multiple Linear Regression (MLR)

The principal components leading to the best fit were selected. An F test with a 95% statistical reliability was used as a criterion to avoid overestimation of the linear model.



Prediction Using Multiple Linear Regression (MLR)

The data set was split randomly into a training set of 123 and a test set of 122 molecules.
For estimation purposes, the training set was treated exactly the same as the complete set, with comparable RMS and maximal errors for each method. The test set was used to predict the partition coefficients of molecules unknown to the regression model.



Prediction Using an Artificial Neural Network (ANN)

Different topologies of a feed-forward back-propagation ANN were tested. The best results were found with a 19-20-1 network, the input neurons using nine principal components of 2D or 3D structure descriptors and ten principal components of the substructure coding.
With the same training set as for MLR, 32 000 training cycles were found to give the best prediction results for the test set.

PCA vs. PLS

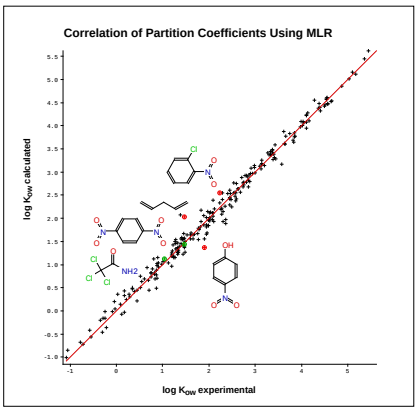
For the data set investigated, PLS proved to be the method of choice because RMS and maximal errors for correlated and predicted values were about halved as compared with PCA. Moreover, computation was as effective as for PCA.

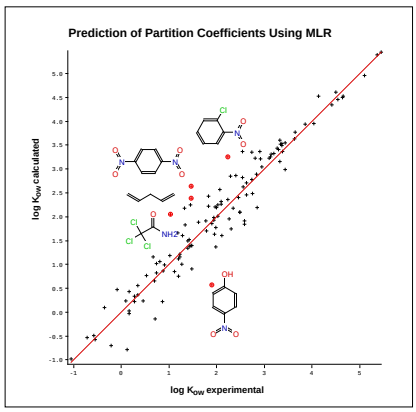
2D vs. 3D Description

The use of 3D information to calculate the structure descriptors yielded only slight improvements. This can be explained as follows: The data set did not contain any isomers (2D description is unable to distinguish between them) and the 3D information was not used to calculate additional values, e.g. volumes or surface properties of the molecules.

ANN vs. MLR

The maximal errors for predicting partition coefficients were about the same with ANN as with MLR, whereas RMS errors were ca. 20% smaller with MLR. When comparing the effects of 2D vs. 3D and of PCA vs. PLS, they were almost the same for both ANN and MLR.
The main disadvantage of ANN lay not so much in the training time (the network topology being very small) but in finding a good network topology, which was very time-consuming.





Examination of the Regression Plots

The two plots (left) show the regression between experimental and calculated values for correlation and prediction with MLR. Obviously, the deviation is bigger for prediction but the distribution is symmetrical and no outliers are found in either case.
Three of the five molecules with worst predictions are nitrobenzenes. While correlations for 1–chloro- and 3-hydroxynitrobenzene with the experimental values are poor, that for p–dinitrobenzene is found to be good. The poor prediction value for the latter can be explained by it being the only dinitro molecule of the whole data set, hence, prediction had to be made without a corresponding molecule in the training set. The same holds for trichloroamide. For 1,4-pentadiene, both correlation and prediction are poor.

Conclusions

For predicting the partition coefficients, the description used seems to be a good possibility to represent small and medium size organic molecules containing heteroatoms.
It was found that PLS gives much better results than PCA. For a data set without isomers, the use of 3D information, as mentioned above, is of only little importance but, in general, it slightly improves the results.
The method does not provide information as to whether the predicted value is based on a huge set of similar reference molecules or extrapolated.

Outlook

In order to achieve more reliable predictions of partition coefficients, extrapolated values must be detected. This could be achieved by clustering similar molecules of the training set and then investigating the position of those with unknown values relative to these clusters. To this purpose, local models for each cluster would have to be developed.