Software SubMat

Calculation of binary substructure descriptors

Version 3.1, June 2004

K. Varmuza
Laboratory for Chemometrics
Institute of Chemical Engineering, Vienna University of Technology
Getreidemarkt 9/166-2, A-1060 Vienna, Austria
kvarmuza@email.tuwien.ac.at http://www.lcm.tuwien.ac.at

Demo Application

For the formula $C_7H_{14}O$ all isomers have been generated by software Molgen [1]. The resulting 596 molecular structures have been stored in Molfile format. A set of 135 substructures has been used for the calculation of substructure descriptors by SubMat. Result of the application of SubMat was a file containing a matrix of size 596x135. The subsequent data analysis has been performed by applying standard software from chemometrics.

A subset of 30 descriptors (exhibiting maximum variances) was selected for a principal component analysis (PCA). The scatter plot of the scores for the first and second principal component shows a clustering of substance classes (Figure 1).

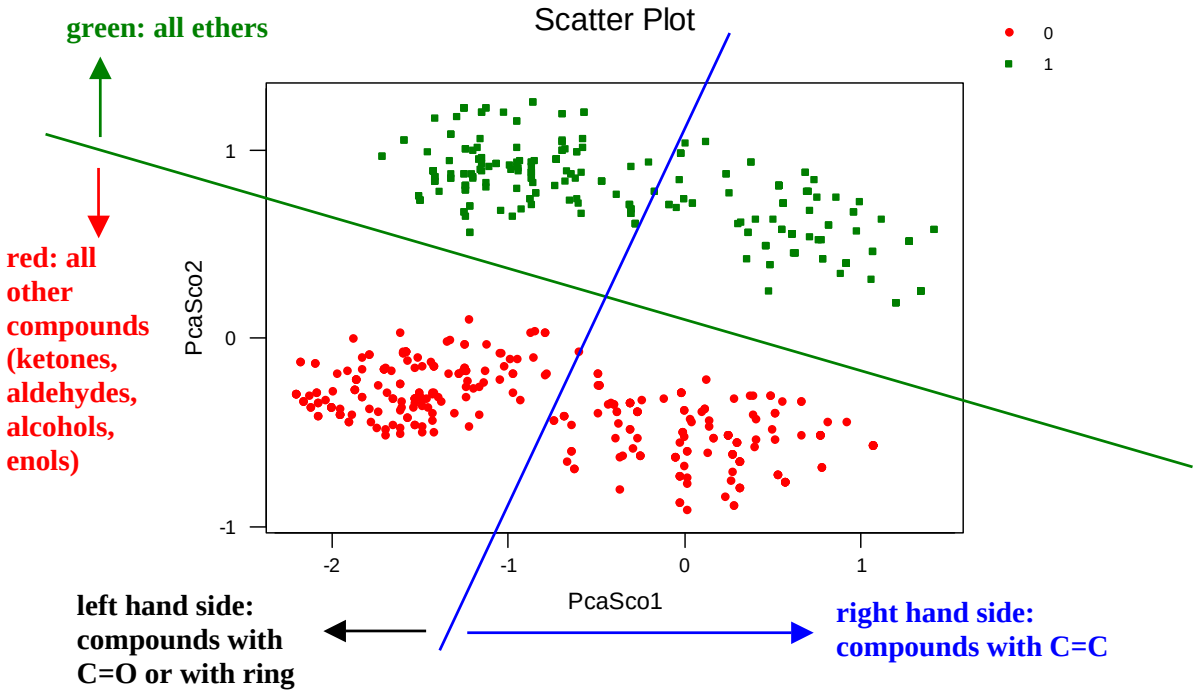

Figure 1. Cluster analysis of chemical structures by PCA using binary substructure descriptors generated by software SubMat. Chemical structures: all 596 isomers for $C_7H_{14}O$; 30 features (selected by maximum variance); coordinates are the scores of first and second principal component (with 25.0 and 10.5 % of total variance), respectively.

[1] Isomer generator software Molgen. Available from A. Kerber and R. Laue, Institute for Mathematics II, University of Bayreuth, Germany. Information: www.molgen.de.