Exploration of Chemical Structure Sets: Binary Molecular Descriptors and Mapping Methods

Karlovits M., Demuth W., Scsibrany H., Müller F., Varmuza K.*

Vienna University of Technology
Institute of Chemical Engineering, Austria

* Presenting and corresponding author: Kurt VARMUZA
Laboratory for **C**hemo**M**etrics, Institute of Chemical Engineering,
Vienna University of Technology, Getreidemarkt 9/166, A-1060 Vienna, Austria

WWW.LCM.TUWIEN.AC.AT

kvarmuza@email.tuwien.ac.at

Poster Presentation: 11th German-Japanese Workshop on Chemical Information
12 - 13 June 2003, Kyoto, Japan

Overview / Software *SubMat*

A set of 1365 substructures (2-dimensional) has been defined for the representation of organic chemical structures by binary substructure descriptors.

Software *SubMat*

Calculates binary substructure descriptors for an input file with molecular structures and an input file with substructures.

SubMat runs under Microsoft Windows operating systems.

Computing time for100 molecular structures, 1000 substructures is 3 s.

Operating modes of SubMat

Interactive.
Remote (call and control for instance from a Matlab program).

Free download and information *at* http://www.lcm.tuwien.ac.at

Free demo version of software, demo structure files, User Guide.
Full version with 500 substructures is available.

Test of the 1365 substructures with two spectral databases

Databases: IR with 13484 compounds, MS with 106955 compounds.

No. of substructures present in the databases: 1265 (92.7%).

No. of substructures per database structure: 0 - 287 (median 76).

Structure similarity searches were performed, using the ***Tanimoto index***, *t,* as similarity criterion of two binary vectors ***x*** and ***y***.

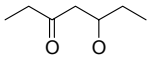
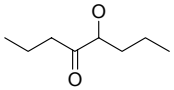
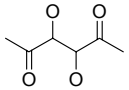
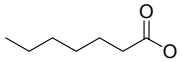
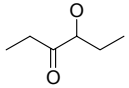
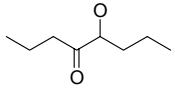
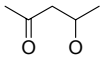
$$t = \boldsymbol{x}^T\boldsymbol{y} / [\boldsymbol{x}^T\boldsymbol{1} + \boldsymbol{y}^T\boldsymbol{1} - \boldsymbol{x}^T\boldsymbol{y}] = \Sigma(\text{AND}) / \Sigma(\text{OR}) \qquad [t:\ 0 \ldots 1]$$

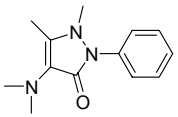
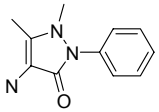
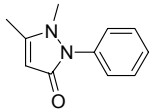
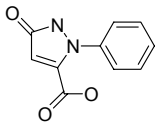
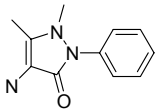
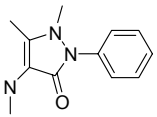
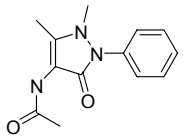
Aim of the work was the application of binary substructure descriptors, together with PCA and PLS, for explorations of spectra similarity hitlists.

📖 Varmuza K., Karlovits M., Demuth W.: *Anal. Chim. Acta* **490**, 95-108 (2003)

Structure similarity search

Examples

query structure	data base	hit 1	hit 2	hit 3
	IR	 $t = 0.86$	 $t = 0.71$	 $t = 0.71$
	MS	 $t = 0.88$	 $t = 0.86$	 $t = 0.85$

query structure	data base	hit 1	hit 2	hit 3
	IR	 $t = 0.98$	 $t = 0.90$	 $t = 0.76$
	MS	 $t = 0.98$	 $t = 0.98$	 $t = 0.94$

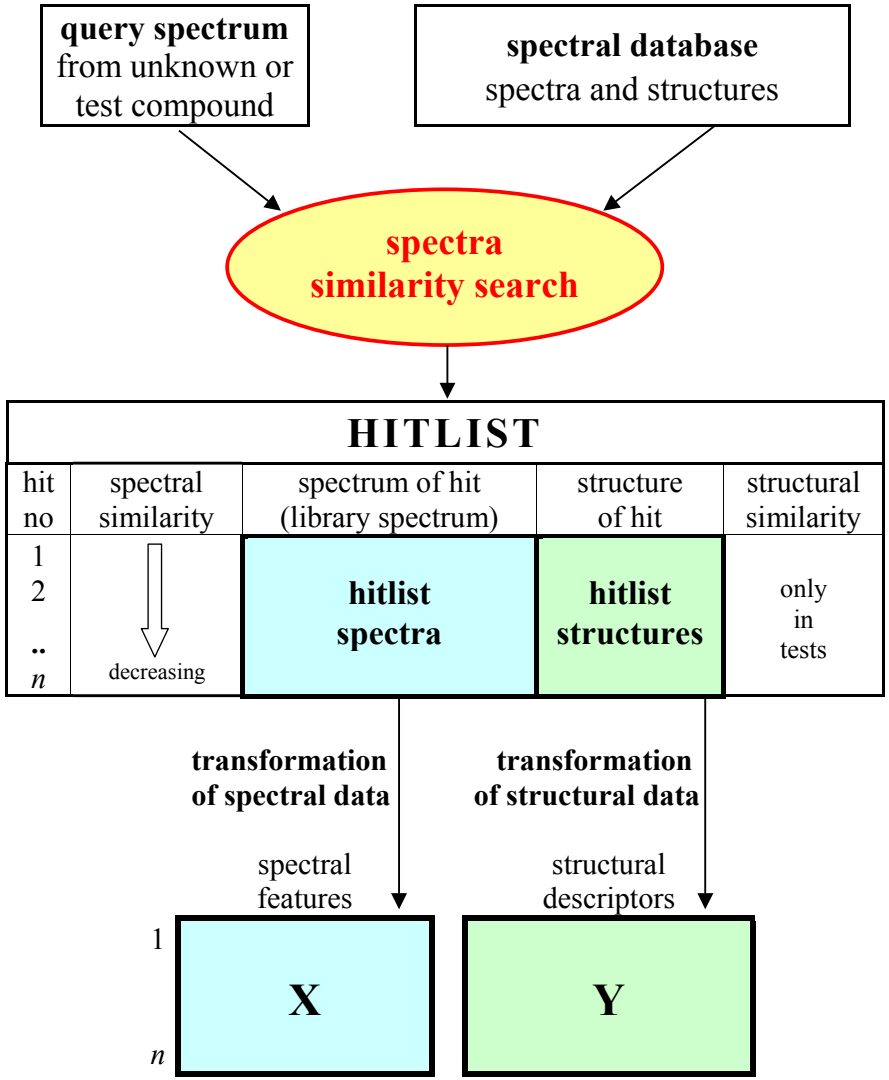
t Tanimoto index

IR SpecInfo Database (13484 compounds)

MS NIST Mass Spectral Database (106955 compounds)

Structures in the database that are identical to the query structure have been excluded.

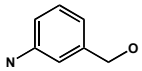
Spectra similarity search



MULTIVARIATE DATA ANALYSIS

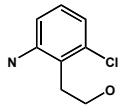
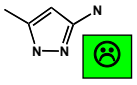
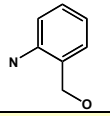
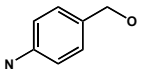
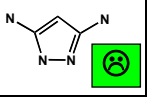
PCA, PLS, Kohonen mapping, KNN, LDA

Example: IR spectrum similarity search (1)

Query compound	 **3-amino-benzylalcohol**
Database	13484 compounds (IR spectra and structures, SpecInfo)
Spectral similarity	correlation coefficient of absorbance units
Structural similarity	Tanimoto index (t) based on 1365 substructures

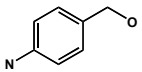
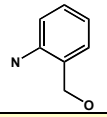
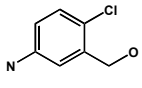
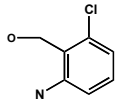
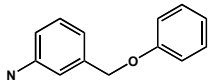
(A) Most similar spectra in database

Tanimoto mean 1 - 5: **0.66**

				
t = 0.74	0.48	0.96	0.96	0.14

(B) Most similar structures in database

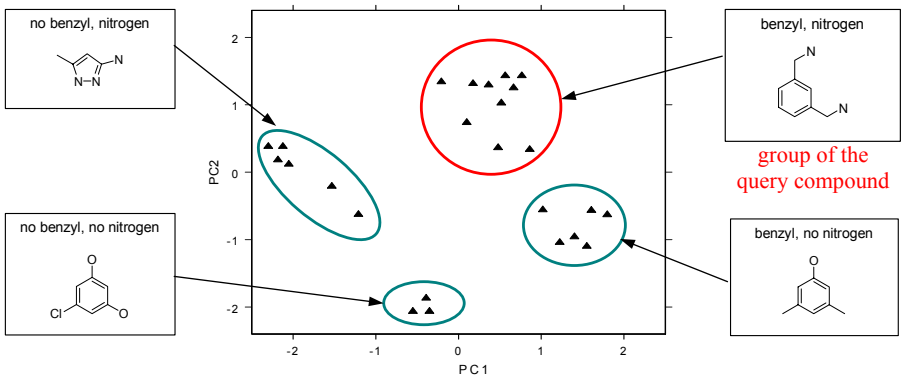
Tanimoto mean 1 - 5: **0.91**

				
t = 0.96	0.96	0.89	0.89	0.85

two best reference structures in yellow

(C) Cluster analysis of hitlist structures by PCA

18 binary substructure descriptors; variance retained in PC1, PC2: 36%, 28%



Four groups of chemical structures found in the hitlist (example structures shown)

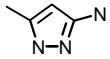
Example: IR spectrum similarity search (2)

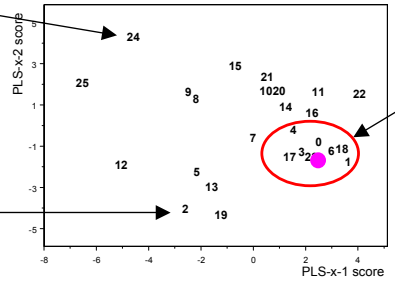
(D) PLS mapping of spectra (X) and structures (Y)

X: averaged absorbances (autoscaled) of 50 wavenumber intervals between 500 and 3700 cm^{-1}
Y: 18 binary substructure descriptors (autoscaled)
PLS-x components are defined by the first two eigenvectors of X^TYY^TX

Hitlist numbers 1-25: 1 = most similar spectrum.

Hit 2: Spectrum is similar to query spectrum, but structures are very different.





Unknown ● is surrounded by database compounds with similar structures; all contain the substructure

Distances in PLS-component space can be used for a new ranking of the hitlist compounds.

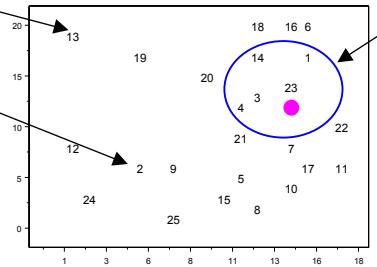
(E) Kohonen mapping of spectra (X)

X: averaged absorbances of 50 wavenumber intervals between 500 and 3700 cm^{-1}
Software SOMPAK (Helsinki University of Technology), map size 20*20

Hitlist numbers 1-25: 1 = most similar spectrum.

Hit 2: Same situation as with PLS.

Only X-data (spectra) have been used for this mapping !



Unknown ● is surrounded by database compounds with similar structures.

Distances in the Kohonen map can be used for a new ranking of the hitlist compounds. However, the result heavily depends on the parameters used.

PCA and PLS support the evaluation of hitlists

- by cluster analysis of chemical structures
- by selection of most relevant database structures

Acknowledgments

A. Kerber and R. Laue (Isomer generator MOLGEN); R. Neudert and E. Pretsch (IR SpecInfo Database)