Similarity between Chemical Structures of Query Compounds and found Hitlist Compounds in Mass Spectral Similarity Searches

K. Varmuza*[1], W. Demuth[1], M. Karlovits[1]
W. Werther[2], E.R. Schmid#[2]

[1] **Vienna University of Technology, Austria**
Inst. of Chemical Engineering, Laboratory for Chemometrics

[2] **University of Vienna, Austria**
Inst. of Analytical Chem., Laboratory for Mass Spectrometry

*	Corresponding author	Kurt Varmuza	kvarmuza@email.tuwien.ac.at www.lcm.tuwien.ac.at
#	Presenting author	Erich R. Schmid	erich.schmid@univie.ac.at www.anc.univie.ac.at/massspec.html
	Acknowledgment	*Austrian Science Fund, project P14792-CHE*	

Poster Presentation: **16th International Mass Spectrometry Conference**
31 August - 5 September 2003, Edinburgh, Scotland

Introduction

Spectra similarity searches
(*library searches*)
are routinely used in mass spectrometry.

This method is often successful in identifying compounds, provided that the unknown is contained in the spectral library.

What happens
if the unknown is NOT in the library ?!

(1) Some MS database systems claim an *interpretive power*.

(2) **We present a study:**

How good is the similarity between the chemical structures of the hits and the chemical structure of the unknown ?

How to select a spectra similarity algorithm that gives hitlists with good structure information about the unknown ?

Similarity of Mass Spectra

The measure for the similarity of two mass spectra is based on the widely used correlation coefficient concept:

$$s_{ab} = 100\ \Sigma(x_{ia}\ x_{ib}) / [\Sigma(x_{ia})^2\ \Sigma(x_{ib})^2]^{0.5}$$

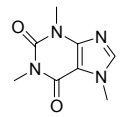
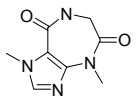
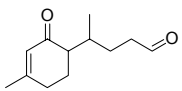
x_{ia}	variable *i* in spectrum ***a*** (for instance peak intensity at mass *i*)
x_{ib}	variable *i* in spectrum ***b*** (for instance peak intensity at mass *i*)
s_{ab}	spectral similarity between ***a*** and ***b***, range 0 - 100

Variables used for the similarity measure

- **peak intensities** $x_i = I_m$ (% B)
- **weighted peak intensities** $x_i = m^{\alpha}\ I_m^{\beta}$ [1,2]
- **spectral features** $x_i = f(I_m)$, $m = m_1 \ldots m_2$ [3,4]

  A spectral feature is a linear or non linear function of selected or all peak intensities. In this work 862 spectral features have been used.

  Examples are: modulo-14 summation, logarithm of intensity ratios, autocorrelation, peak series.

Example for a query (caffein)	Hit 1 and hit 2 from a search in the NIST MS Database (= compounds with reference spectra most similar to query spectrum)
	  ☹

[1] S. E. Stein, D. R. Scott, J. Am. Soc. Mass Spectrom. **5** (1994) 859-866.
[2] S.E. Stein, J. Am. Soc. Mass Spectrom. **6** (1995) 644-655.
[3] K. Varmuza, in: J. C. Lindon, G. E. Tranter, J. L. Holmes (Eds.), Encyclopedia of spectroscopy and spectrometry, Academic Press, London, 2000, p. 232-243.
[4] W. Werther, W. Demuth, F.R. Krueger, J. Kissel, E.R. Schmid, K. Varmuza, J. Chemometrics **16** (2002) 99-110.

Similarity of Chemical Structures

The similarity between two chemical structures is measured by the Tanimoto index:

$$t_{ab} = \Sigma\ \mathrm{AND}(d_{ja}, d_{jb}) / \Sigma\ \mathrm{OR}(d_{ja}, d_{jb})$$

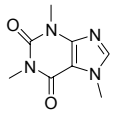
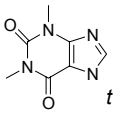
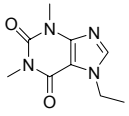
d_{ja}	substructure descriptor *j* in structure ***a***
d_{jb}	substructure descriptor *j* in structure ***b***
t_{ab}	Tanimoto similarity between structures ***a*** and ***b***, range 0 - 1 [5]

The performance (interpretive power) of a library search method is measured by a "grand mean":

$$T_h = (1/nh) \sum_q \sum_k t_{qk} \qquad k = 1 \ldots h \text{ (hits)} \qquad q = 1 \ldots n \text{ (queries)}$$

h	number of hits considered (1 - 30)
n	number of query compounds tested (200 - 1000)
t_{qk}	Tanimoto index for query structure *q* and hit *k*
T_h	grand mean of Tanimoto indices for *h* hits, range 0 - 1

A set of **1365 "general purpose" substructures** has been defined [6].
A molecular structure is characterized by a bit string (fingerprint) of length 1365.
Software **SubMat** is used to calculate these bit strings [7].

Example for a query structure (caffein)	Two most similar structures (measured by Tanimoto index) in the MS library containing 106,955 compounds (NIST)
	 *t* = 0.99  *t* = 0.99 ☺

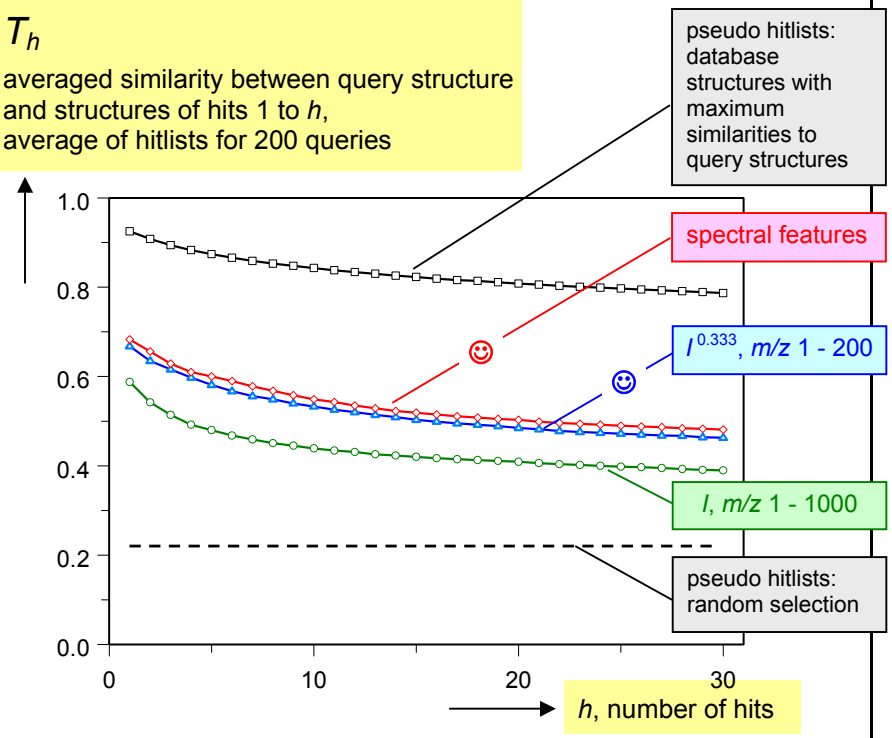
[5] P. Willet, Similarity and clustering in chemical information systems, Research Studies Press, Letchworth (UK), 1987.
[6] H. Scsibrany, M. Karlovits, W. Demuth, F. Müller, K. Varmuza, Chemom. Intell. Lab. Syst. **67** (2003) 95-108.
[7] Software **SubMat** is available from ***Laboratory for ChemoMetrics***. Information at **www.lcm.tuwien.ac.at** (software); demo version and User Guide for free download.

Results

Mass spectral library 106,955 compounds from NIST MS Database *

Queries n = 200 spectra randomly selected

T_h
averaged similarity between query structure and structures of hits 1 to h,
average of hitlists for 200 queries



Conclusions

- First hit has highest structural similarity with query structure.
- **Spectral features** give best results.
- **Cubic root of peak intensities up to m/z 200** give good results.

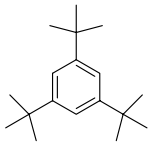
An analogous study has been made with infrared spectra [8].

* Thanks to S.E. Stein for providing this database.

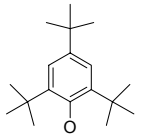
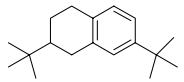
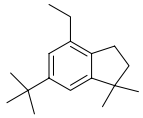
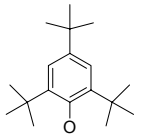
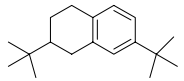
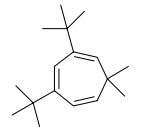
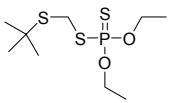
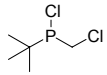
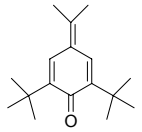
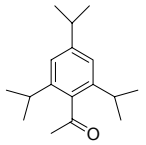
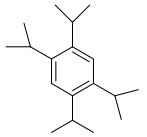
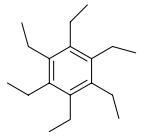
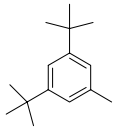
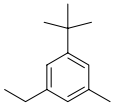
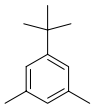
[8] K. Varmuza, M. Karlovits, W. Demuth, Anal. Chim. Acta **490** (2003) 313-324.

Example

Query



s spectral similarity (0 ... 100)
s_{NIST} spectral similarity in NIST MS Database (0 ... 999)
t structural similarity (0 ... 1)
$t_{1\text{-}3}$ averaged t for hit 1 to 3
Hits identical with query have been excluded.

Method	Hit 1	Hit 2	Hit 3	
862 spectral features				
$t_{1\text{-}3}$ = 0.65	s = 94.6 t = 0.76	s = 94.5 t = 0.60	s = 94.4 t = 0.60	☺
Cubic root of peak intensities m/z 1 - 200				
$t_{1\text{-}3}$ = 0.58	s = 89.8 t = 0.76	s = 89.3 t = 0.60	s = 89.0 t = 0.39	☺
Peak intensities m/z 1 - 1000				
$t_{1\text{-}3}$ = 0.22	s = 90.3 t = 0.11	s = 73.5 t = 0.16	s = 75.2 t = 0.39	☹
NIST 98 Similarity search for "user spectrum"				
$t_{1\text{-}3}$ = 0.66	s_{NIST} = 775 t = 0.56	s_{NIST} = 769 t = 0.72	s_{NIST} = 715 t = 0.71	☺
Maximum structural similarity				
$t_{1\text{-}3}$ = 0.99	t = 1.00	t = 0.98	t = 0.98	