

Structural and Spectral Similarity

K. Varmuza*, W. Demuth and M. Karlovits

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>



Manuscript for plenary lecture 6 March 2002

CHEM 02
Biannual Conference on Chemistry
Cairo University - Chemistry Department

4 - 7 March 2002, Cairo, Egypt

Additional references

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

Chemom. Intell. Lab. Syst., **67**, 95-108 (2003)

Clustering and similarity of chemical structures represented by binary substructure descriptors.

Varmuza K., Karlovits M., Demuth W.:

Anal. Chim. Acta., **490**, 313-324 (2003)

Spectral similarity versus structural similarity: infrared spectroscopy.

Manuscript version: 28 November 2003

Copyright: K. Varmuza, Vienna (Austria) 2003.

* Corresponding author, email: kvarmuza@email.tuwien.ac.at, <http://www.lcm.tuwien.ac.at>

Structural and Spectral Similarity

K. Varmuza*, W. Demuth and M. Karlovits

Laboratory for Chemometrics
Vienna University of Technology, Institute of Chemical Engineering
Getreidemarkt 9/166, A-1060 Vienna, Austria

Abstract

A method has been developed for the evaluation of spectral library search methods to measure their capability in giving useful results if the unknown is not contained in the database. The similarities of the chemical structures in the hitlist and the structure of a test compound are measured by the Tanimoto index (calculated for a set of binary substructure descriptors). Results for IR and MS databases show that in general the first hits have highest structural similarity with the unknown. A new type of spectral similarity for MS data is described that improves significantly the structural similarity between the found hits and the unknown.

Keywords: spectral library search, infrared spectra, mass spectra, Tanimoto index

Varmuza-Chem02-1b.doc 2002-02-28

1. Introduction

The chemical structure information contained in infrared spectra (IR) or mass spectra (MS) is difficult to extract because of the complicated and widely unknown relationships between spectral data and chemical structures. For instance, the fragmentation processes which result in the measured data characterize MS as a chemical method. Chemical effects are, in general, more difficult to describe and to predict than physical ones.

The aim of spectra evaluation can be either the *identification* of a compound (assuming the spectrum is already known and available) or the *interpretation* of spectral data in terms of the unknown chemical structure (with the spectrum of the unknown usually not available) [1-3].

Identification is performed best by library search methods based on spectral similarities; a number of spectral databases and powerful software products are offered for this purpose and are routinely used.

The more challenging problem is the interpretation of spectra which still is a topic of research projects in chemometrics and computer chemistry. No comprehensive solutions are available and these methods are not used in routine work.

Although spectral library search methods are successfully applied in many laboratories, their benefits are sometimes doubted when used for the interpretation of spectra from unknowns that are

not present in the library. An *interpretative power* is claimed by some spectroscopic database systems, however, systematic investigations are rare that quantify the similarity between the chemical structures of hitlist compounds and the chemical structure of test compounds. In this work preliminary results from a systematic investigation of this subject with MS and with IR data are presented.

2. Theory

The interpretative power of a spectral library search system is the ability to produce hitlists (containing the most similar spectra from a database) which provide molecular structures similar to that of the query compound. For a systematic evaluation of library search systems it is necessary to define similarity criteria for spectra as well as for chemical structures. Similarity concepts are always relative and refer to some specific context [4]; nevertheless they are necessary and useful if relationships between available data (spectra) and desired data (chemical structures) are not known sufficiently. All results, however, refer to the applied mathematical definitions for similarity and the used data sets.

2.1. Similarity of spectra

An infrared spectrum is characterized by a vector \mathbf{a} with the component a_i being the averaged absorbance in wave number interval i . The range used was 500 to 3700 cm^{-1} with 801 intervals,

* Corresponding author, email: kvarmuza@email.tuwien.ac.at, <http://www.lcm.tuwien.ac.at>

each 4 cm⁻¹ wide; a_i was scaled to a mean of zero for each spectrum. The similarity $s(IR)$ of two IR spectra A and B was defined as [5]

$$s(IR) = 999(r + 1) / 2$$

with

$$r = \Sigma(a_{iA} a_{iB}) / [\Sigma(a_{iA})^2 \Sigma(a_{iB})^2]^{0.5}$$

for $i = 1 \dots 801$

r corresponds to the correlation coefficient, and $s(IR)$ ranges between 0 and 999.

A mass spectrum is characterized by a vector y with the components y_i being numerical features derived from a low resolution mass spectrum. A basic approach is to use the peak intensities (in % of the base peak intensity) at masses i as features y_i . Another approach is described in the Result's section. The similarity $s(MS)$ of two mass spectra A and B was defined as [6]

$$s(MS) = \Sigma(y_{iA} y_{iB}) / [\Sigma(y_{iA})^2 \Sigma(y_{iB})^2]^{0.5}$$

for $i = 1 \dots p$

p is the number of features; $s(MS)$ corresponds to the correlation coefficient and is in the range 0 to 1 because y_i was always positive.

2.2. Similarity of chemical structures

A chemical structure has been characterized by a vector d with the components d_i being binary substructure descriptors. A set of 135 substructures was defined - with the aim to cover a broad range of chemistry; d_i is 1 if substructure i is present in the molecule and 0 otherwise. The similarity of two structures A and B was defined by the Tanimoto index t [7]

$$t = \Sigma \text{AND}(d_{iA}, d_{iB}) / \Sigma \text{OR}(d_{iA}, d_{iB})$$

AND denotes the logical "and" operation, and OR the logical "or" operation; summation is calculated over all 135 descriptors. t is in the range 0 to 1; the value 1 is obtained if all descriptors are pairwise equal.

2.3. Similarity between the structure of a test compound and the hitlist structures

The quality of a hitlist - in terms of high structural similarity to the structure of the test compound - was defined by an averaged Tanimoto index, t_h

$$t_h = (1/h) \Sigma t_j \quad \text{with } j = 1 \dots h$$

h is the number of hits considered, and t_j is the

Tanimoto index calculated from the structure of the test compound and hit j .

A spectral similarity search method was characterized by averaging the results for n randomly selected test compounds. T_h is the averaged structural similarity between the n test compounds and the corresponding first h hits. T_h ranges between 0 and 1. If the spectral similarity actually reflects the structural similarity then maximum values for T_h are expected for a low number of h . Different spectral similarity search methods were compared by comparing the corresponding results for T_h as a function of h .

3. Results

3.1. Databases and software

The IR database used consists of 13484 compounds and is part of the SpecInfo system [8]. The MS database used consists of 60909 compounds and is part of the NIST mass spectral database 98 [9]. Chemical structures for both spectral databases were available as connection tables (Molfile format).

Substructure searches for the calculation of substructure descriptors were performed by software SubMat [10]. Most other software used was developed using Matlab 6 (The Mathworks, Inc.).

3.2. Infrared spectra

A random sample with 200 test compounds was selected from the IR database. For each test compound a hitlist containing the compounds with the most similar spectra was determined; the compound identical to the test compound was excluded. Figure 1 shows the averaged structural similarities T between the hitlist compounds and the test compounds. From this result is evident that in general the structure(s) of the first hit(s) has (have) highest similarity with the structure of the test compound.

The results obtained from hitlists can be compared with two extreme situations. First is the generation of pseudo hitlists by selecting the compounds randomly from the database. In this case the structural similarity of hitlist compounds corresponds to the average Tanimoto index within the database, which is 0.23 (with a standard deviation of 0.13). The averaged structural similarities using hitlists are in the range 0.5 to 0.7 and thereby significantly higher than the mean within the database.

The other extreme is to search for the compounds with highest structural similarity to the test compound. Such pseudo hitlists exhibit the maximum structural similarity that can be obtained with the

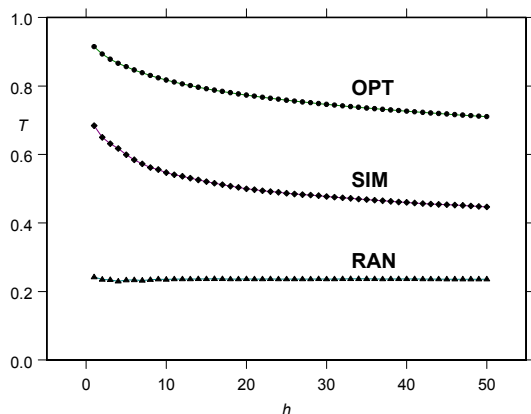


Fig. 1. IR data: Averaged structural similarity T between test compounds and h hitlist compounds. OPT, pseudo hitlist containing database structures with maximum similarity to test compound; SIM, hitlist by spectral similarity; RAN, pseudo hitlist containing a random selection of the database.

used database. As shown in Figure 1 these maximum values are considerable higher than those obtained from hitlists based on spectral similarity.

Further experiments indicate that a representation of IR spectra by only 100 data points (instead of 801) does not reduce the structural similarity of the hitlist compounds.

3.2. Mass spectra

A random sample with 200 test compounds was selected from the MS database. For each test compound a hitlist containing the compounds with the most similar spectra was determined; the compound identical to the test compound was excluded.

Figure 2 shows two curves of the averaged structural similarity, T , versus the number of hits, h . One curve is for using peak intensities in the similarity criterion, the other for a new approach, namely using *spectral features*. As with IR data the structure(s) of the first hit(s) has (have) highest similarity with the structure of the test compounds. The new approach with spectral features yields significantly better results than obtained with peak intensities. Generation of spectral features from low resolution MS data has been described elsewhere [6,11,12]; a summary is given in Table 1.

In Figure 3 the first three hitlist structures obtained by these two methods are given for a selected test compound. For comparison also the results are given as obtained by the software dis-

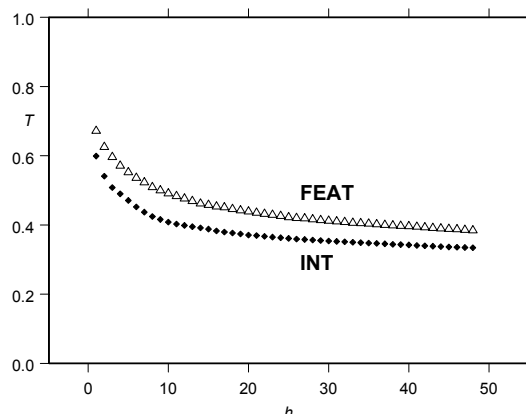


Fig. 2. MS data: Averaged structural similarity T between test compounds and h hitlist compounds. FEAT, 658 features used for spectral similarity; INT, peak intensities (% base peak) used for spectral similarity.

Table 1. Mass spectral features

group	feature description
1	intensities (% base peak) at single masses (m/z 25-150)
2	intensities at single masses normalized to local ion current (window of 7 masses, m/z 25-150)
3	averaged intensities of mass ranges m/z 25-39, 40-70, 71-100, 101-150
4	logarithmic intensity ratios of peaks with mass differences 1 and 2, m/z 25-150
5	modulo-14 summation for mass ranges 25-120, 25-250, and 121-250
6	autocorrelation for mass differences of 1, 2, and 14-60 in mass ranges 25-120, 25-250, and 100-250
7	spectra type features describing the distribution of peaks
8	peak series features that characterize the joint presence of peaks at given masses

tributed with the NIST mass spectral database 98. This system uses a library with 108000 compounds; the applied spectral similarity criterion is not published in detail. Best results were obtained with the new approach using spectral features in the calculation of spectral similarities; average Tanimoto index T_3 of the first three hits is 0.783. Only poor results were obtained using peak intensities (T_3 is 0.146). Result from the NIST system with $T_3 = 0.612$ is considerably lower than that obtained with the new approach.

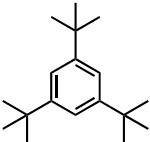
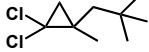
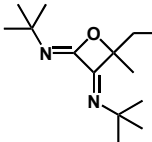
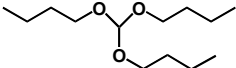
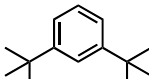
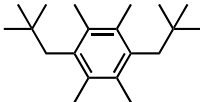
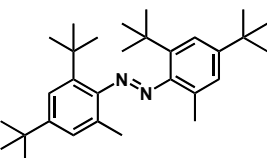
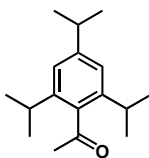
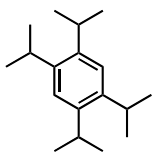
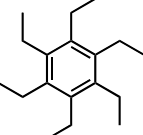
<div>test compound</div> <div>  </div>				
method	hit 1	hit 2	hit 3	T_3
A peak intensities	 0.235	 0.143	 0.059	0.146
B spectral features	 0.889	 0.889	 0.571	0.783
C NIST	 0.462	 0.750	 0.625	0.612

Fig. 3. MS data: Hitlists obtained by three different methods of spectra similarity search. A, this work, using peak intensities; B, this work, using spectral features; C, NIST MS database system. Numbers near the structures are Tanimoto indices calculated from the structure of the test compound and the structures of the hitlist compounds; T_3 , mean of Tanimoto indices for the first three hits.

4. Conclusions

Hitlists from spectral library searches in IR and MS databases have been investigated in terms of the structural similarity between hitlist structures and structures of query compounds. The first hits (corresponding to the most similar spectra) exhibit highest structural similarity. For MS the use of spectral features (instead of peak intensities) for calculating the spectral similarity yields a significant improvement of the structural similarities. The presented method provides an objective approach for an evaluation of spectra similarity searches for situations in which the unknown is not contained in the database.

Acknowledgments. We thank R. Neudert and E. Pretsch for providing the infrared spectra as well as H. Scsibrany and F. Müller for collaboration. The Austrian Science Fund (project P14792-CHE) supported this work.

References

- [1] T. L. Clerc, in: H. L. C. Meuzelaar, T. L. Isenhour (Eds.), *Computer-enhanced analytical spectroscopy*, Plenum Press, New York, 1987, p. 145-162.
- [2] H. J. Luinge, *Vib. Spectrosc.* 1 (1990) 3-18.
- [3] F. W. McLafferty, S. Y. Loh, D. B. Stauffer, in: H. L. C. Meuzelaar (Ed.), *Computer-enhanced analytical spectroscopy*, Plenum Press, New York, 1990, p. 163-181.
- [4] D. H. Rouvray, *J. Chem. Inf. Comput. Sci.* 34 (1994) 446-452.
- [5] K. Varmuza, P. N. Penchev, H. Scsibrany, *J. Chem. Inf. Comput. Sci.* 38 (1998) 420-427.
- [6] 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.
- [7] P. Willet, *Similarity and clustering in chemical information systems*, Research Studies Press, Letchworth (UK), 1987.
- [8] SpecInfo: Spectroscopic Information System. 3.1, Chemical Concepts: PO Box 100202, D-6944 Weinheim, Germany; 1996.
- [9] NIST '98 Mass spectral database. National Institute of Standards and Technology, Gaithersburg, MD 20899, 1998.
- [10] K. Varmuza, H. Scsibrany, *J. Chem. Inf. Comput. Sci.* 40 (2000) 308-313.
- [11] K. Varmuza, W. Werther, *J. Chem. Inf. Comput. Sci.* 36 (1996) 323-333.
- [12] K. Varmuza, J. Kissel, F. R. Krueger, E. R. Schmid, in: E. Gelpi (Ed.), *Advances in mass spectrometry*, Wiley, Chichester, 2001, p. 229-246.