From MS Data via Chemometrics to Chemical Structure Information

K. Varmuza

Laboratory for Chemometrics
Vienna University of Technology, Getreidemarkt 9/160, A-1060 Vienna, Austria

kvarmuza@email.tuwien.ac.at, http://www.lcm.tuwien.ac.at

Manuscript for lecture   20 January 2001

13th Sanibel Conference on Mass Spectrometry
Informatics and Mass Spectrometry
American Society for Mass Spectrometry
19 - 22 January 2001, Sanibel Island, Florida, USA

Version: 1g-ref (3 Jan 2001), and 8 Jan 2001

From MS Data via Chemometrics to Chemical Structure Information

K. Varmuza

Vienna University of Technology, Laboratory for Chemometrics
Getreidemarkt 9/160, A-1060 Vienna, Austria
kvarmuza@email.tuwien.ac.at, http://www.lcm.tuwien.ac.at

1. Introduction

The chemical structure information contained in mass spectra is difficult to extract because of the complicated and widely unknown relationships between MS data and chemical structures. 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 spectra similarities; a number of MS databases and powerful software products are offered for this purpose and are routinely used [4,5].

The more challenging problem is the interpretation of mass 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.

Four groups of different strategies have been applied to the complex problems of substructure recognition or recognition of more general structural properties from spectral data.

(1) *Knowledge-based methods* try to implement spectroscopic knowledge about spectra-structure relationships into computer programs. Because of the lack of generally applicable rules this approach was not successful in MS. However, spectroscopic knowledge has been extensively applied in other methods to guide the construction of mathematical models.
(2) Appropriate *interpretive library search* techniques can be used to obtain structural information if the unknown is not contained in the library.
(3) *Correlation tables* containing characteristic spectral data (key ions) together with corresponding substructures had only limited success because a specific structural property does not always give the same spectral signals.
(4) *Spectral classifiers* are algorithms based on multivariate classification methods or neural networks; they are constructed for an automatic recognition of structural properties from spectral data.

This paper focuses on the application of multivariate data analysis - the typical chemometric approach - to investigate relationships between low resolution electron impact data and chemical structures as well as on the development and use of MS classifiers together with automatic isomer generation [6]. Chemometric methods are successful to some extent in the automatic recognition of substructures or other structural properties from low resolution electron impact mass spectra [6-15]. In some cases a systematic structure elucidation is possible from the molecular formula of an unknown together with restrictions about the presence or absence of substructures (automatically obtained from spectra). From such data an exhaustive set of possible chemical structures can be constructed by an appropriate isomer generator software.

The methods of multivariate data analysis applied to spectra interpretation are all based on the characterization of spectra by a set of variables (spectral features). A spectrum then can be considered as a point in a multidimensional space with the coordinates defined by these spectral features. Several mathematical procedures are available to "look" into the high dimensional space (exploratory data analysis, cluster analysis) or to find decision rules (classifiers) capable to separate for instance a substance class from all other compounds.

The application of multivariate data analysis to spectra interpretation typically consists of the following steps:

(1) Generation of a set of spectra and the corresponding chemical structures (hitlist from a spectral similarity search or result from a database search).
(2) Transformation of peak list data into a set of spectral features and eventually the selection of the most relevant features. Chemical structures have to be encoded by a set of molecular descriptors.
(3) Application of methods form multivariate data analysis, such as for instance principal component analysis (PCA) for exploratory data analysis or multivariate classification for the development of spectral classifiers.

A mathematical description of the used chemometric standard methods [16-24] is beyond the scope of this paper.

2. Transformation of MS data into spectral features

An appropriate mathematical transformation of the original peak list data into a set of suitable features is essential for a successful application of multivariate data analysis [3]. A spectral feature x_j is a number that can be automatically computed from a mass spectrum. Usually nonlinear transformations are applied that sometimes consider spectroscopic knowledge. Aim of data transformation is to obtain a set of variables that are better suited for a structure related spectra interpretation than the original peak list data alone. A summary of mass spectral features is given in Table 1. The typical number of features used is between 10 and several hundreds.

3. Exploratory data analysis

The most prominent method for an exploratory analysis of multivariate data is principal component analysis (PCA). The resulting PCA scatter plots for spectra and for features often present spectra-structure relationships or indicate substance classes that are reflected by the used spectral features and the used chemometric method. The example in Figure 1 shows the clustering of mass spectra from aliphatic and alicyclic ketones according to the number of double bond equivalents (DBE). The data set used consisted of 200 randomly selected spectra from the NIST Mass Spectral Database [25] applying the restrictions: 50 compounds each for DBE = 1, 2, 3, 4, 5; molecular mass 70 to 250; molecular formula $C_nH_mO_1$. The mass spectra have been transformed into the 14 features from the modulo-14-summation type (Table 1, group 5).

Table 1. Summary of mass spectral features. All these features (x_j) are calculated from peak intensities I_m (% base peak intensity) and are in the range 0 to 100 [12,14,26,27].

Group	Feature definition	Ref.
1	INTENSITIES AT SINGLE MASSES are useful for informative key fragments. $x_j \;=\; I_m$	-
2	INTENSITIES AT SINGLE MASSES NORMALIZED TO LOCAL ION CURRENT. The local ion current is the sum of peak intensities in a mass interval ± Δm around mass m. $x_j \;=\; 100\, I_m / \Sigma I_k$ with $k \;=\; m - \Delta m \;....\; m + \Delta m$	[28]
3	AVERAGED INTENSITIES OF MASS INTERVALS. This feature group reflects the distribution of peaks in the lower and higher mass ranges. $x_j \;=\; \Sigma I_k / (m_2 - m_1 + 1)$ with $k \;=\; m_1 \;...\; m_2$	-
4	LOGARITHMIC INTENSITY RATIOS. This feature group reflects the better reproducibility of intensity ratios compared with absolute intensities. The equations given avoid arithmetic problems with zero intensities. $x_j \;=\; 100\, (L_m + \ln 100) / (2 \ln 100)$ with $L_m \;=\; \ln I_k / I_{k+\Delta m}$ and $I_z \;=\; \max\, (I_z, 1)$	[14]
5	MODULO-14 SUMMATION. One of the first numerical transformations successfully used for mass spectra is the summation of intensities at masses differing by a multiple of 14. A set of 14 possible features is defined as follows. $x_j \;=\; 100\, s_j / s_{max}$ *with* $s_j \;=\; \Sigma I_{l\,+\,14k}$ $(j, l \;=\; 1 \;...\; 14;\; k \;=\; 0, 1, 2, ...)$ and $s_{max} \;=\; \max\, (s_1, \;...\; s_{14})$	[29]
6	AUTOCORRELATION FEATURES reflect characteristic mass differences between peaks as well as periodicities in a spectrum. $x_j \;=\; 100\, \Sigma I_m I_{m\,+\,\Delta m} / S_0$ with $S_0 \;=\; \Sigma I_m I_m$	[15]
7	SPECTRA TYPE FEATURES characterize the distribution of peaks across the mass range. $x_{j,\, dust} \;=\; 100\, \Sigma I_m / I_{all}$ with $m = 25 \;...\; 78$ and I_{all} = sum of all I_m $x_{j,\, base} \;=\; 100.\, 100 / I_{all}$ $x_{j,\, even} \;=\; 100\, \Sigma I_{2k} / I_{all}$ with for instance $k \;=\; 13 \;...\; 400$	[14]
8	CHARACTERISTIC PEAK SERIES FEATURES. The joint presence of a series of N characteristic masses $m(k)$, $k = 1 \ldots N$ can be described by the following two features. $x_{j,\, product} \;=\; (\, \Pi A_{m(k)}\,)^{1/N}$ with $k = 1 \;...\; N$ and $A_{m(k)} \;=\; \max\, (I_{m(k)}, 1)$ $x_{j,\, mean} \;=\; (N^*/N)\, (\, \Sigma I_{m(k)}\,) / N$ with $k = 1 \;...\; N$ and N^*, number of present peaks	[30]

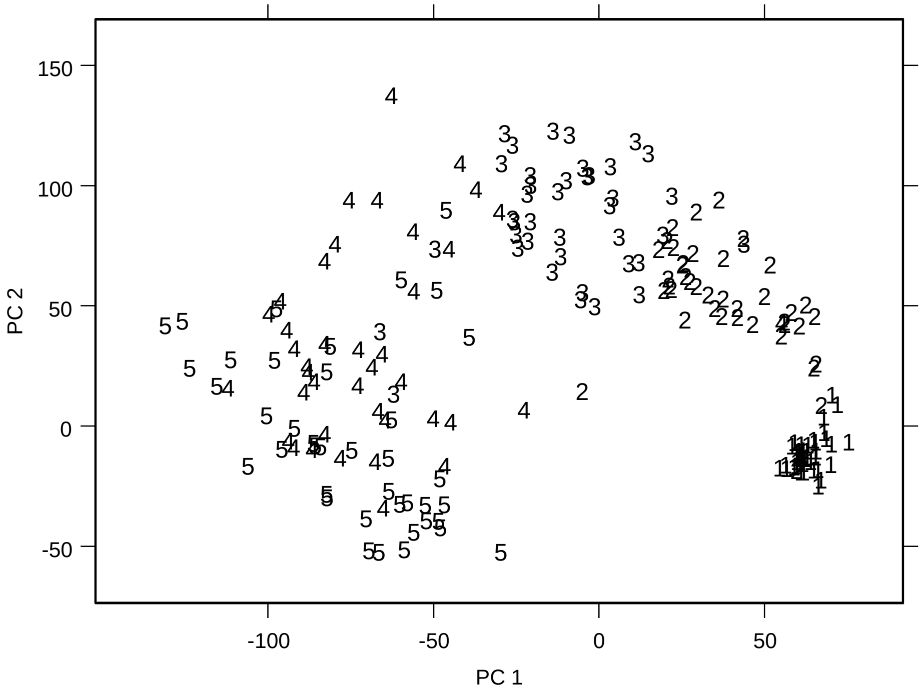

Figure 1. PCA score plot for 200 mass spectra from aliphatic and alicyclic ketones using 14 modulo-14 features (Table 1). The numbers denote the number of double bond equivalents. The axes are the first and second principal component (preserving 35.8 and 21.8 % of the total variance, respectively).

4. Classification of substructures

The two most important computer-assisted strategies for the recognition of structural information from mass spectral data are:

(1) The structures in the hitlist from a spectra similarity search are used to estimate the probability of substructures in the unknown [31].

(2) Random samples of mass spectra are first characterized by a set of spectral features and then multivariate classification methods or neural networks are applied to develop spectral classifiers. Only this approach will be briefly treated in this paper [6,12,27].

A spectral classifier is a mathematical function (or algorithm) with spectral features as input and one new variable - the discriminant variable - as output. Let $x_1, x_2, \dots x_p$ be the spectral features, and z a discriminant variable; a linear classifier is defined by

$$z = b_1 x_1 + b_2 x_2 + \dots + b_p x_p$$

For a discrimination of two mutual exclusive classes of spectra the coefficients b_j can be determined with the aim to obtain values for z near +1 for one class and values near -1 for the other. The coefficients b_j define a vector in the feature space which corresponds to the axis of the discriminant variable. Classification means to project a point (corresponding to a spectrum) onto this axis and thereby determining the value of z. A simple approach for class assignment uses the sign of z (if $z > 0$ then class 1, else class 2). The number of wrong classifications can be reduced if rejection thresholds are used, however, at the cost of not classified spectra.

The coefficients b_j are estimated from a data set containing spectra of both classes (training set); the resulting classifier has to be tested with spectra that have not been used for the training (prediction set). A widely used method for the calculation of linear classifiers is linear discriminant analysis (LDA) - preferably combined with a preceding PCA. The typical strategy for the development of spectral classifiers is shown in Figure 2.

Today's performance of mass spectra classification by multivariate methods can be summarized as follows [3]:

(1) Only a rather small number of substructures can be recognized with a low error rate.
(2) Predictions of the absence of substructures are usually more accurate than predictions of their presence.
(3) Erroneous classifications cannot be avoided completely; therefore interaction of a human expert and the parallel use of other spectra interpretation methods are advisable.
(4) For small molecules a systematic and almost complete structure elucidation is sometimes possible by mass spectra classification and by application of the obtained structural restrictions in automatic isomer generation. An example is shown in the next chapter.
(5) Classification by neural networks [7-9,14] or feature selection by genetic algorithms [32,33] improved the performance in some cases but did not enable a break-through.

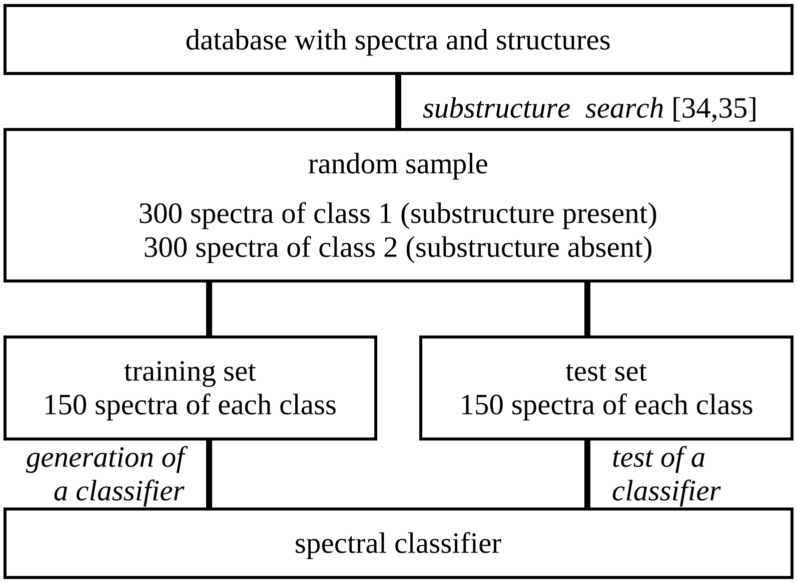
database with spectra and structures
substructure search [34,35]
random sample
300 spectra of class 1 (substructure present)
300 spectra of class 2 (substructure absent)
training set
150 spectra of each class
test set
150 spectra of each class
generation of a classifier
test of a classifier
spectral classifier

Figure 2. Development and test of a spectral classifiers for the recognition of a substructure [6].

5. Systematic structure elucidation with isomer generation

The most important *systematic* approach for structure elucidation of organic compounds is still based on the DENDRAL project [36] (Figure 3). Central tool is an isomer generator software capable to generate the exhaustive set of isomers from a given molecular formula. The generator also considers structural restrictions which are usually obtained from spectral data. Substructures which have to be present in the unknown molecular structure are collected in the so-called *goodlist* while forbidden substructures are put into the *badlist* [13]. Mass spectrometry can contribute to this approach in two aspects: The molecular formula can be determined from high resolution data, and structural information can be derived from low resolution data.

A successful example for this approach of structure elucidation is presented in Figure 4. The compound *benzene acetic acid, 2-hydroxy, ethyl ester* has been considered as unknown [3]. Given is the molecular formula and the mass spectrum. Application of software MSclass [6,12,13] resulted in substructures for the goodlist and the badlist; only substructures relevant to the molecular formula are shown.

The isomer generator software used was MOLGEN [37-39]. This program computes the complete set of connectivity isomers for a given molecular formula; structural restrictions can be defined by a goodlist (separated into substructures that may not overlap and those that may overlap), and a badlist. Furthermore, lower and upper limits can be defined for bond multiplicity and ring size, as well as the number of H-atoms at C-, N-, and O-atoms. The construction of isomers is free from duplicates and fast.

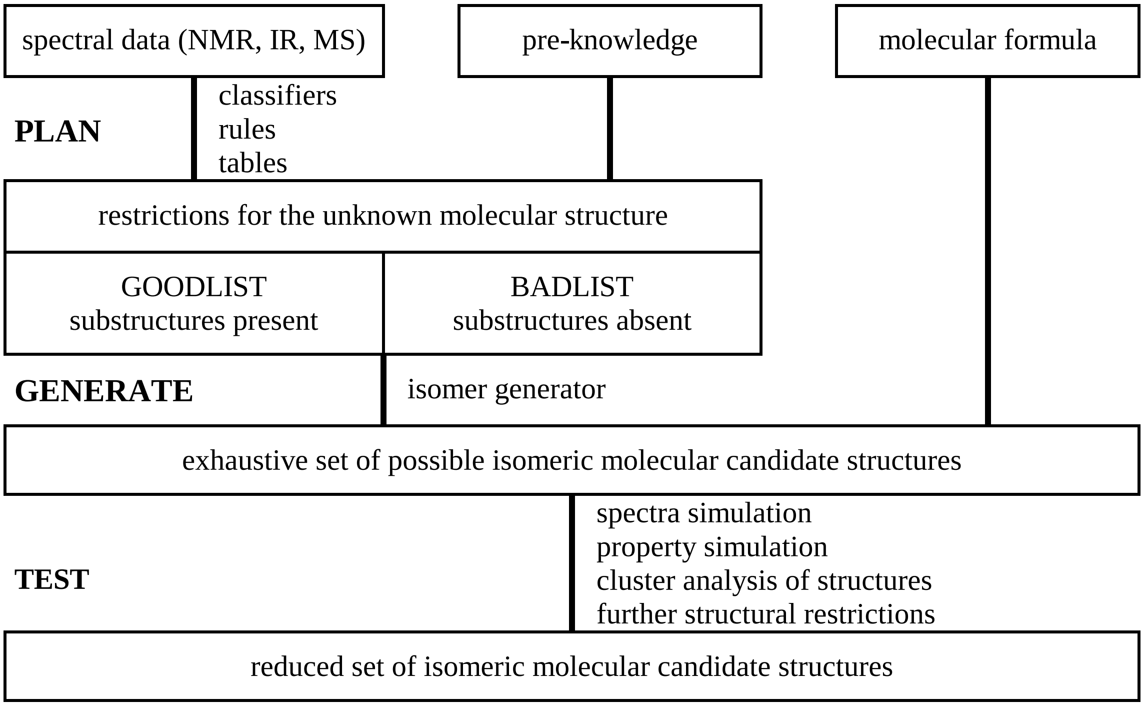
spectral data (NMR, IR, MS)
pre-knowledge
molecular formula
PLAN
classifiers
rules
tables
restrictions for the unknown molecular structure
GOODLIST
substructures present
BADLIST
substructures absent
GENERATE
isomer generator
exhaustive set of possible isomeric molecular candidate structures
TEST
spectra simulation
property simulation
cluster analysis of structures
further structural restrictions
reduced set of isomeric molecular candidate structures

Figure 3. Systematic structure elucidation of organic compounds based on the DENDRAL approach [6].

Considering these structural restrictions, six isomers are possible, including the correct solution. Computation time on a modern personal computer is only a few seconds when the structural restrictions of the goodlist and badlist are considered.

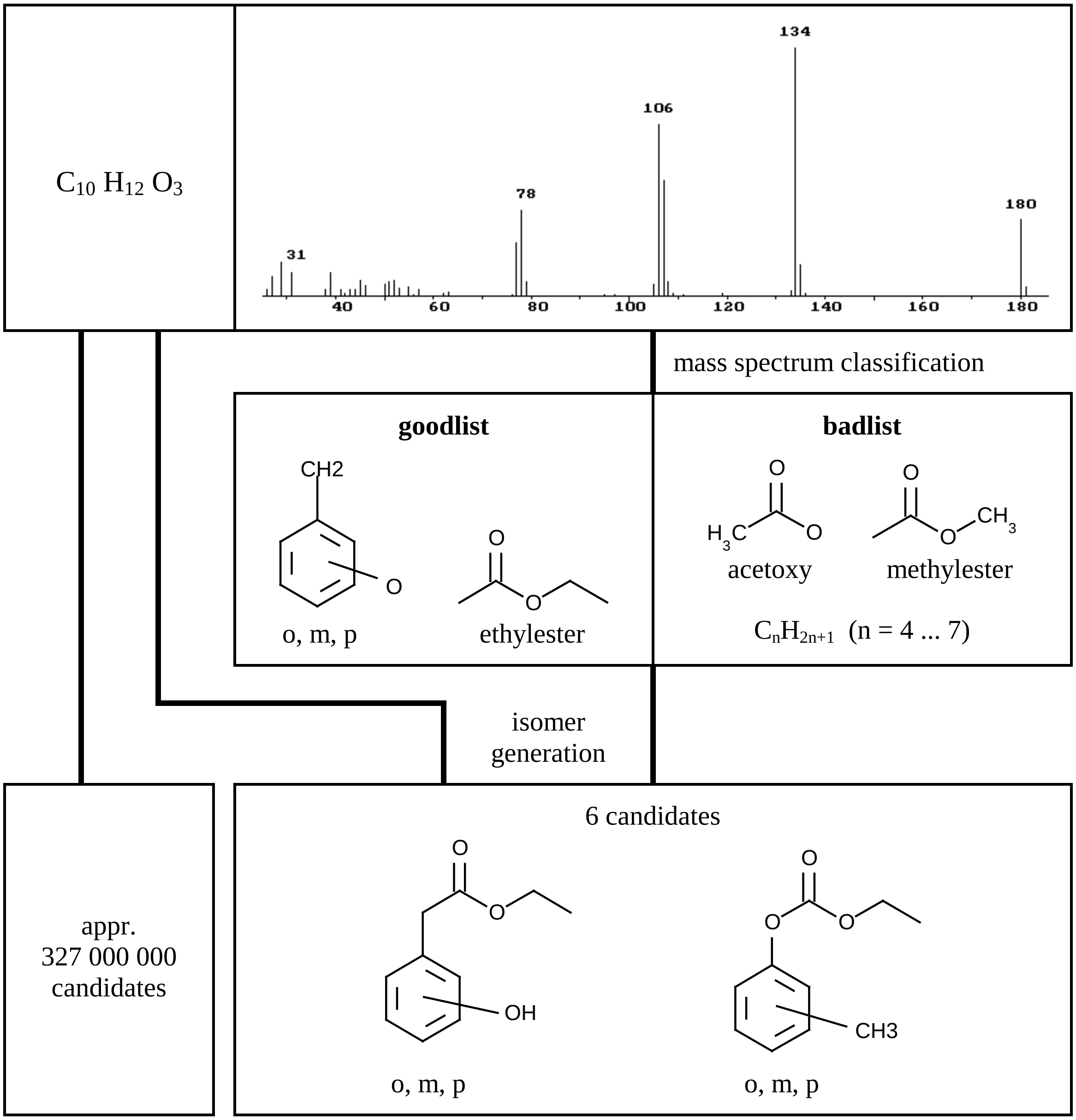

Figure 4. Systematic structure elucidation using the molecular formula of the unknown, structural restrictions from automatic mass spectra classification and exhaustive isomer generation. Considered as unknown: benzene acetic acid, 2-hydroxy, ethyl ester [3].

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. *Automated spectra interpretation and library search systems*.

[2] F. W. McLafferty, S. Y. Loh, D. B. Stauffer, in: H. L. C. Meuzelaar (Eds.), Computer-enhanced analytical spectroscopy, Plenum Press, New York, 1990, p. 163-181. *Computer identification of mass spectra*.

[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. *Chemical structure information from mass spectrometry*.

[4] F. W. McLafferty, R. H. Hertel, Org. Mass Spectrom. **8** (1994) 690-702. *Probability based matching of mass spectra*.

[5] S. E. Stein, D. R. Scott, J. Am. Soc. Mass Spectrom. **5** (1994) 856-866. *Optimization and testing of mass spectral library search algorithms for compound identification*.

[6] K. Varmuza, W. Werther, in: E. J. Karjalainen, A. E. Hesso, J. E. Jalonen, U. P. Karjalainen (Eds.), Advances in mass spectrometry, Elsevier, Amsterdam, 1998, p. 611-626. *Systematic structure elucidation of organic compounds based on mass spectra classification and isomer generation*.

[7] D. Cabrol-Bass, C. Cachet, C. Cleva, A. Eghbaldar, T. P. Forrest, Can. J. Chem. **73** (1995) 1412-1426. *Application pratique des reseaux neuro mimetiques aux donnees spectroscopiques (infrarouge et masse) en vue de l´elucidation structurale*.

[8] B. Curry, D. E. Rumelhart, Tetrahedron Comput. Methodol. **3** (1990) 213-237. *MSnet: A neural network which classifies mass spectra*.

[9] C. Klawun, C. L. Wilkins, J. Chem. Inf. Comput. Sci. **36** (1996) 249-257. *Joint neural network interpretation of infrared and mass spectra*.

[10] K. Varmuza, W. Werther, D. Henneberg, B. Weimann, Rapid Comm. Mass Spectrom. **4** (1990) 159-162. *Computer-aided interpretation of mass spectra by a combination of library search with principal component analysis*.

[11] K. Varmuza, Int. J. Mass Spectrom. Ion Proc. **118/119** (1992) 811-823. *Chemometrics in mass spectrometry*.

[12] K. Varmuza, W. Werther, J. Chem. Inf. Comput. Sci. **36** (1996) 323-333. *Mass spectral classifiers for supporting systematic structure elucidation*.

[13] K. Varmuza, W. Werther, F. Stancl, A. Kerber, R. Laue, in: J. Gasteiger (Eds.), Software development in chemistry, vol. **10**, Gesellschaft Deutscher Chemiker, Frankfurt am Main, 1996, p. 303-314. *Computer-assisted structure elucidation of organic compounds, based on mass spectra classification and exhaustive isomer generation*.

[14] W. Werther, H. Lohninger, F. Stancl, K. Varmuza, Chemometrics Intell. Lab. Syst. **22** (1994) 63-76. *Classification of mass spectra. A comparison of yes/no classification methods for the recognition of simple structural properties*.

[15] S. Wold, O. H. J. Christie, Anal. Chim. Acta **165** (1984) 51-59. *Extraction of mass spectral information by a combination of autocorrelation and principal components models*.

[16] M. J. Adams, *Chemometrics in analytical spectroscopy*. The Royal Society of Chemistry, Cambridge, 1995.

[17] K. R. Beebe, R. J. Pell, M. B. Seasholtz, *Chemometrics: A practical guide*. John Wiley & Sons, New York, 1998.

[18] P. Geladi, B. R. Kowalski, Anal. Chim. Acta **185** (1986) 1-17. *Partial least squares regression: a tutorial*.

[19] H. Martens, T. Naes, *Multivariate calibration*. Wiley, Chichester, 1989.

[20] D. L. Massart, B. G. M. Vandeginste, L. C. M. Buydens, S. De Jong, J. Smeyers-Verbeke, *Handbook of chemometrics and qualimetrics: Part A*. Elsevier, Amsterdam, 1997.

[21] B. G. M. Vandeginste, D. L. Massart, L. C. M. Buydens, S. De Jong, J. Smeyers-Verbeke, *Handbook of chemometrics and qualimetrics: Part B*. Elsevier, Amsterdam, 1998.

[22] K. Varmuza, in: P. v. R. Schleyer, N. L. Allinger, T. Clark, J. Gasteiger, P. A. Kollman, I. H. F. Schaefer, P. R. Schreiner (Eds.), The encyclopedia of computational chemistry, Wiley, Chichester, 1998, p. 346-366. *Chemometrics: Multivariate view on chemical problems*.

[23] S. Wold, in: J. Brandt, I. K. Ugi (Eds.), Computer applications in chemical research and education, Hüthig Verlag, Heidelberg, 1989, p. 101-128. *Multivariate data analysis: Converting chemical data tables to plots*.

[24] J. Zupan, J. Gasteiger, *Neural networks in chemistry and drug design*, edn 2. Wiley-VCH, Weinheim, 1999.

[25] NIST, *NIST ´98 Mass spectral database*. National Institute of Standards and Technology, Gaithersburg, MD 20899, 1998.

[26] W. Werther, W. Demuth, F. R. Krueger, J. Kissel, E. R. Schmid, K. Varmuza, submitted (2000). *Evaluation of mass spectra from organic compounds assumed to be present in cometary grains. Exploratory data analysis*.

[27] K. Varmuza, J. Kissel, F. R. Krueger, E. R. Schmid, submitted (2000). *Chemometrics and TOF-SIMS of organic compounds near a comet*.

[28] F. Erni, J. T. Clerc, Helv. Chim. Acta **55** (1972) 489-500. *Strukturaufklärung organischer Verbindungen durch computerunterstützten Vergleich spektraler Daten*.

[29] L. R. Crawford, J. D. Morrison, Anal. Chem. **40** (1968) 1469-1474. *Computer methods in analytical mass spectrometry. Empricial identification of molecular class*.

[30] K. Varmuza, W. Werther, F. R. Krueger, J. Kissel, E. R. Schmid, Int. J. Mass Spectrom. **189** (1999) 79-92. *Organic substances in cometary grains: Comparison of secondary ion mass spectral data and Californium-252 plasma desorption data from reference compounds*.

[31] S. E. Stein, J. Am. Soc. Mass Spectrom. **6** (1995) 644-655. *Chemical substructure identification by mass spectral library searching*.

[32] D. Broadhurst, R. Goodacre, A. Jones, J. J. Rowland, D. B. Kell, Anal. Chim. Acta **348** (1997) 71-86. *Genetic algorithms as a method for variable selection in multiple linear regression and partial least squares regression, with applications to pyrolysis mass spectrometry*.

[33] H. Yoshida, R. Leardi, K. Funatsu, K. Varmuza, submitted (2000). *Feature selection by genetic algorithms for mass spectral classifiers*.

[34] K. Varmuza, H. Scsibrany, J. Chem. Inf. Comput. Sci. **40** (2000) 308-313. *Substructure isomorphism matrix*.

[35] K. Varmuza, H. Scsibrany, Vienna University of Technology (Austria), Laboratory for Chemometrics. Information: WWW.LCM.TUWIEN.AC.AT (2000). *Software SubMat (generation of binary substructure descriptors)*.

[36] N. A. B. Gray, *Computer-assisted structure elucidation*. Wiley, New York, 1986.

[37] T. Grüner, A. Kerber, R. Laue, M. Meringer, K. Varmuza, W. Werther, Match **38** (1998) 173-180. *MASSMOL*.

[38] T. Grüner, A. Kerber, R. Laue, M. Meringer, Match **37** (1998) 205-208. *MOLGEN 4.0*.

[39] A. Kerber, R. Laue, M. Meringer, University of Bayreuth (Germany), Institute for Mathematics II. Information: WWW.MOLGEN.DE (2000). *Isomer generator software MOLGEN*.

Acknowledgments. Collaborations with W. Werther, H. Scsibrany, F. Stancl, W. Demuth, A. Kerber, R. Laue, M. Meringer, K. Funatsu, H. Yoshida, and R. Leardi are gratefully acknowledged.