Mass Spectral Classifiers for Supporting Systematic Structure Elucidation†

K. Varmuza* and W. Werther

Department of Chemometrics, Technical University of Vienna, A-1060 Vienna, Lehargasse 4/152, Austria

Received September 30, 1995⊗

A set of mass spectral classifiers has been developed to recognize presence or absence of 70 substructures or more general structural properties in a molecule. Classification is based on numerical transformation of low resolution mass spectral data, automatic selection of appropriate features, multivariate discriminant methods, and estimation of the reliability of the classification answer. Examples demonstrate applications in structure elucidation together with automatic isomer generation as well as combination with results obtained by the CHEMICS system.

INTRODUCTION

Development of computer-assisted methods for automatic elucidation of chemical structures of organic compounds is a central theme in computer chemistry and chemometrics since many decades.[1] Research effort in this field stimulated progress in artificial intelligence[2] as well as in applications of multivariate statistical methods in chemistry.[3−5] Essentially, three different approaches have been used in computer-assisted structure elucidation based on spectroscopic data. (1) A number of software tools has been developed for supporting manual and mental work of spectroscopists, a recent example is SpecTool.[6] (2) Library search is now routinely used in most areas of spectroscopy. (3) Systematic and in some sense exhaustive methods for structure elucidation date back to the DENDRAL approach.[1] Pioneering work in this field includes, for instance, the software systems GENOA,[7] CHEMICS,[8] STREC,[9] and SESAMI.[10] Examples for more recently developed systems are SpecInfo,[11,12] CSEARCH,[13] and X-PERT.[14]

Systematic computer-based structure elucidation usually consists of three steps: (1) plan, (2) generate, and (3) test/select. In the first step restrictions about the chemical structure are derived from spectroscopic data, probably the most interesting and challenging part for chemistry. The second step is generation of all isomers agreeing with the structural restrictions. This step requires knowledge of the brutto formula in most systems. A recently developed stand alone and efficiently working isomer generator is MOLGEN.[15] Because the number of generated candidate structures is often high, a third step is necessary to reduce the candidate list. For this purpose usually spectra simulation and comparison with the measured spectrum is applied but also cluster analysis of the candidate structures has been proposed.[16,17]

Automatic recognition of substructures or other more general structural properties from spectral data is still a subject of research. Four groups of different techniques have been applied to this complex problem: (1) Chemical structures obtained in a hitlist as resulting from a library search have been used to predict the presence or the absence of substructures in unknowns.[11,18−20] (2) Correlation tables containing selected spectral data (for instance chemical shift intervals) together with corresponding substructures are widely applied. (3) Knowledge-based methods try to implement more complex spectroscopic knowledge about spectra-structure relationships. (4) Finally, spectral classifiers based on multivariate statistics have been developed for automatic recognition of structural properties.

Because of the relative strict relationships between spectral data and atom-centered substructures, ^{13}C-NMR data are frequently used for computer-assisted structure elucidation. H-NMR and IR data are less often considered. As a consequence of the chemical nature of a mass spectrum relationships between mass spectral data and chemical structures are complex. Therefore, most systems do not use MS data, although a lot of effort has gone into this problem especially during the pioneering phase of chemometrics.[3,21−23]

This paper describes a method for developing classifiers capable of recognizing automatically the presence or the absence of substructures from low resolution mass spectral data. Potential applications of mass spectral classifiers can be expected in trace analysis of organic compounds. Suitable NMR data cannot be measured if an unknown compound is available for identification only at the nanogram level and only for a few seconds, which are typical conditions of trace analyses by GC/MS. Classifier development was based on a specific data transformation of mass spectra and the application of methods from multivariate statistics. Aim of the work was to obtain classification results that can be used together with isomer generator programs. Capabilities and limits of the approach are demonstrated on examples in which only mass spectral data were available. Combination with other spectroscopic data was also tested.

MASS SPECTRA CLASSIFICATION

1. General. A mass spectral classifier is defined here as an algorithm. Input are mass spectral data of a chemical compound. Output is a (numerical) response providing information about the presence or the absence of a particular substructure (or a more general structural property) in the molecular structure of the investigated compound. A simple type of classifier output is a binary *yes/no* answer. Because fragmentation processes of ions occurring in a mass spectrometer are very complex, no general theory is available today that can be used for building classifiers. It is therefore not surprising that knowledge-based classifiers using rules

* Corresponding author. Email: kvarmuza@email.tuwien.ac.at. Fax: +43-1-581-1915, Tel: +43-1-58801-4988.

† Dedicated to the 70th birthday of Prof. S. I. Sasaki.

⊗ Abstract published in *Advance ACS Abstracts,* March 1, 1996.

0095-2338/96/1636-0323$12.00/0

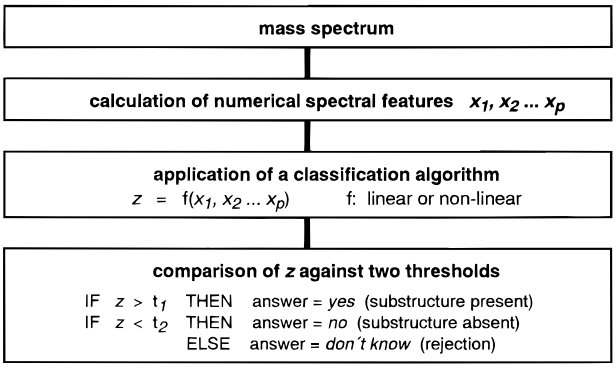


Figure 1. Classification of a mass spectrum.

such as "if peak at mass m has an intensity greater than I_m then substructure S_k is present" are not very effective.[24] An alternative approach is based on statistics and automated learning methods; summaries of works in this field have been published.[3,23,25]

The general scheme of mass spectral classifiers as used in this work is demonstrated in Figure 1. Previous work[26,27] strongly indicated that adequate transformation of the original spectral data is essential for a good performance of classifiers. Linear as well as nonlinear mathematical classification algorithms are then applied to obtain a numerical output z, from which the classifier answer is derived. The parameters of a classification algorithm are calculated in a training phase by using a random sample of spectra originating from compounds with known chemical structures. Application of a classifier to a mass spectrum requires three steps: (a) generation of spectral features, (b) computation of the numerical output, and (c) determination of the classifier answer.

2. Binary Classifiers with a Continuous Response and a Rejection Region. A mass spectrum can be characterized by a set of p numerical features x_j ($j = 1$ to p). A classification function $z = f(x_1,...,x_p)$ has to be developed (trained) with the aims to yield output $z = 1$ for compounds containing the substructure to be classified (class 1), and $z = -1$ for all other compounds (class 2). The training can be based for instance on regression methods or neural networks. Performance of a classifier is estimated by the classification of spectra that have not been used in the training (prediction set).

A simple binary classifier only has the discrete responses *yes* and *no* corresponding to class 1 and class 2, respectively. This response is obtained by defining a threshold t for the classifier output z. Classification results with $z > t$ are interpreted as the answer *yes*, while results $z < t$ yield the answer *no*. Unfortunately, the probability density distributions of z-values for the two classes overlap considerably for most substructure classification problems in mass spectrometry. A simple *yes/no* classifier is therefore not applicable.[26] Even nonlinear neural networks are often not able to separate the classes sufficiently.

Binary classifiers with a continuous response and a rejection region offer possibilities to partly overcome this problem. In those classifiers the value of response z is used to estimate a reliability of classification. Training results indicate that the assumption for this approach is usually fulfilled, namely that the probability of a correct answer depends on the value of z.

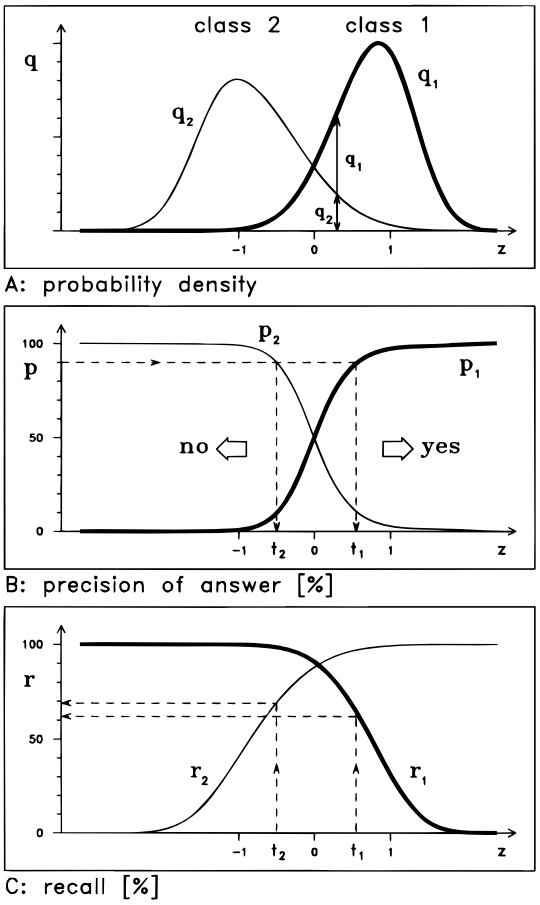


Figure 2. Binary classifier with continuous response z and rejection region t_2 to t_1.

Figure 2A shows probability density curves q_1 and q_2 for the two mutual exclusive classes as obtained from a prediction set by applying a fictive classifier. We assume equal *a priori* probabilities for both classes and consequently the areas between curves q_1 and q_2 and the abscissa are also equal. Precision p of a classifier answer (in %) is defined as the following.

$$\text{precision of answer } yes: \quad p_1 = 100q_1/(q_1 + q_2) \qquad (1)$$

$$\text{precision of answer } no: \quad p_2 = 100q_2/(q_1 + q_2) \qquad (2)$$

Note that the precision of an answer is equivalent to the *a posteriori* probability; it depends on the assumed *a priori* probabilities for both classes.

Figure 2B displays the precision of answers as a function of the classification result z. The curves estimate the probabilities for obtaining correct answers. If a minimum precision (for instance 90%) for of each of the answers is defined, then a rejection region t_1 to t_2 has to be applied.

$$\text{IF } z > t_1 \quad \text{THEN answer} = yes$$

$$\text{IF } z < t_2 \quad \text{THEN answer} = no$$

$$\text{IF } t_2 \le z \le t_1 \quad \text{THEN answer} = don't\ know \text{ (rejection)}$$

A rejection interval prevents the classifier from giving *yes/no* answers with a low reliability. If the probability density curves overlap considerably—or if a high minimum precision is required—then only a small part of the spectra may result

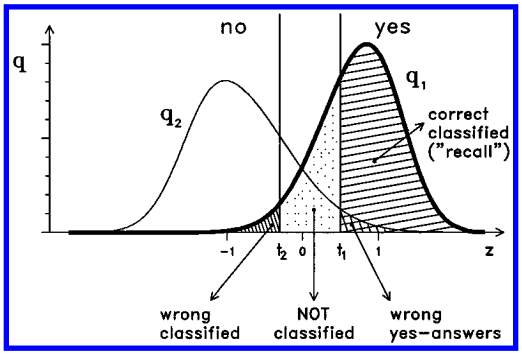


Figure 3. Subsets of class 1 as obtained by the classification results. q_1, q_2: probability densities for class 1 and 2; z: numerical result of classification algorithm; t_1, t_2: classification thresholds for a given minimum precision of answers *yes* (for class 1) and *no* (for class 2).

in a *yes* or *no* answer, while for most spectra the answer may be refused.

Efficiency of a classifier is evaluated by the recalls r_1 and r_2 for correctly classified spectra from class 1 and 2, respectively.

$$r_1 = 100 \int_{t_1}^{\infty} q_1(z)\, dz \tag{3}$$

$$r_2 = 100 \int_{-\infty}^{t_2} q_2(z)\, dz \tag{4}$$

Figure 2C shows the recall as a function of z. A classifier has been considered to be potentially useful if at least one of the classes reaches a minimum recall (for instance 30%) at a minimum precision (for instance 90%). Adjusting the classification thresholds by defining a desired minimum precision of the answers allows the user to adapt the classifier for a particular purpose.

Figure 3 summarizes partitioning of spectra belonging to class 1 into subsets by a classifier. The average precision of *yes* answers (probability of a correct answer) p_{1m} is given by

$$p_{1m} = 100 r_1/(r_1 + e_2) \tag{5}$$

The average precision of *no* answers p_{2m} is given by

$$p_{2m} = 100 r_2/(r_2 + e_1) \tag{6}$$

e_1 and e_2 are the percentages of wrong classified spectra from class 1 and 2, respectively.

$$e_1 = 100 \int_{-\infty}^{t_2} q_1(z)\, dz \tag{7}$$

$$e_2 = 100 \int_{t_1}^{\infty} q_2(z)\, dz \tag{8}$$

Evidently, average precisions of *yes* or *no* answers are higher than the precisions at the borders t_1 and t_2 of the rejection interval.

3. Spectral Features. Mathematical transformation of mass spectra before applying formal interpretation methods dates back to early applications of multivariate methods for spectra interpretation,[21] library search,[28] and pattern recognition.[3] Recently, the importance of an appropriate mathematical representation of mass spectra has been demonstrated again.[26,29−31] A spectral feature x_j (a spectral invariant) is a number that can be automatically computed from a spectrum. Algorithms for the calculation of spectral features often contain spectroscopic knowledge or assumptions; usually they make a nonlinear transformation of the original spectral data. The purpose of spectra transformation is to obtain a set of spectral features that are closely related to molecular structures. Definitions of mass spectral features have already been published,[26,27,30] therefore only an overview is presented here.

Let I_m be the intensity of a peak at mass m, normalized to the highest peak in the spectrum (base peak with intensity 100%). A feature x_j is a linear or nonlinear function of selected (or sometimes all) peak intensities. None of the features used in this work requires knowledge of the molecular weight.

3.1. Feature Group: Logarithmic Intensity Ratios

$$x_j = \ln I_m/I_{m+\Delta m} \tag{9}$$

The mass difference Δm has been varied between 1 and 14, masses m between 28 and 220. Intensities below 1% have been set to 1 in order to avoid arithmetic problems.

3.2. Feature Group: Intensity Sums

$$x_j = \sum I_m \quad (m = m_1, m_2, \ldots m_d) \tag{10}$$

The most prominent member of this group are the 14 features obtained by an intensity summation in mass intervals of 14 (modulo-14-features):

$$x_1 = I_1 + I_{15} + I_{29} + I_{43} + I_{57} + \ldots \tag{11}$$

$$x_{14} = I_{14} + I_{28} + I_{42} + I_{56} + I_{70} + \ldots \tag{12}$$

Summation of intensities at user-selected masses or mass ranges has also been applied to characterize mass spectra.[20,28]

3.3. Feature Group: Autocorrelation

$$x_j = \sum I_m I_{m+\Delta m} \Big/ \sum I_m I_m \tag{13}$$

These features reflect mass differences between peaks. Mass difference Δm has been varied between 1 and 50, considering either the full mass range or only the upper or only the lower half.

3.4. Feature Group: Peak Distribution and Spectra Type. A series of features characterize the distribution of peaks across the mass range; they mainly describe stability of the molecular ion. Further features have been heuristically defined for special classes of compounds, for instance feature

$$x_j = I_{43} I_{57} I_{71} \times 10^{-4} \tag{14}$$

which is characteristic for alkyl groups containing at least five carbon atoms.

In this work each mass spectrum has been transformed to a set of 500−4000 spectral features. The methods applied to select smaller sets containing relevant features is described below.

4. Feature Selection. Relative large sets of spectral features have been used to represent mass spectral information, because no *a priori* knowledge about the best features for a particular classification problem was available. For theoretical and practical reasons a reduction of the number

of features was necessary. Two methods for generating a reduced set containing p features (typical $p = 10$) have been applied.

4.1. Fisher Ratio Selection. The Fisher ratio F_j describes[3,32] the discriminating power of a single feature x_j for separating class 1 and 2. It is a function of the class means ${}_1m_j$ and ${}_2m_j$ and the class variances ${}_1v_j$ and ${}_2v_j$.

$$F_j = ({}_1m_j - {}_2m_j)^2/({}_1v_j + {}_2v_j) \qquad (15)$$

The Fisher ratio is closely related to the t-value as used in statistical tests. For a given data set F_j is proportional to the squared t-value if both classes have equal size and/or if the class variances are equal. For LDA classifiers the ten features with highest Fisher ratio have been selected. This method is fast but suffers from not considering correlations between the features.

4.2. Stepwise Selection. Contrary to the univariate selection by Fisher ratios this method[33] includes the training process and also partly considers interactions between features. Sets of features are built systematically and tested for their discrimination ability measured by the transinformation.[3,26] The applied algorithm can be summarized as the following: (1) Start with a total of p_o features. (2) Select the feature with the best discrimination ability and set the number of selected features p to 1. (3) Build all $p_o - p$ sets of features containing the p selected features plus one of the remaining features. Increment p by 1. (4) Calculate classifiers for all $p_o - p$ feature sets. Select the set which yields highest discrimination ability. (5) Continue with step 3 until no increase of the performance is measured or a given number of features is obtained. This method has been used for RBF classifiers starting with a preselected set of 300 features exhibiting highest Fisher ratios.

5. Classification Algorithms. A detailed comparison[26] of four complementary classification methods[34] showed that neural networks (NN) produced slightly better classifiers than linear discriminant analysis (LDA) and k-nearest neighbor classification (KNN). The class modeling method SIMCA performed worst. Considering the high computational effort and large storage requirements of KNN classifiers the two methods LDA and NN were selected for this work.

5.1. Linear Discriminant Analysis (LDA). To avoid problems with collinear features a preceding principal component analysis (PCA) was applied; only those PCA scores were used as variables in LDA which had at least 0.1% of the total variance. An LDA-classifier for computing a continuous classifier response z contains p parameters.

5.2. Neural Networks Based on Radial Basis Functions (RBF). This special type of neural networks consists of only one hidden layer and a linear transfer functions for the neurons in the output layer.[35,36] The neurons of the hidden layer are considered as kernels. Each kernel corresponds to a prototype of a cluster and is used as the center of a Gaussian-like potential function (called a radial basis function). An optimum weighted superposition of all potential functions approximates the binary class membership variable. The great advantage of this approach is that no time-consuming iterative learning process is required as necessary for other training methods of neural networks. An RBF classifier requires $2p$ parameters for autoscaling of the features and $(p + 1)h + 1$ parameters for the network containing h hidden neurons.

MASS SPECTRAL LIBRARIES AND SOFTWARE

MassLib. The mass spectroscopic data base system MassLib,[37] version 7.2, is running on a Vaxstation 4000/60 under VMS. MassLib contains more than 140 000 mass spectra in several libraries, the largest is the Wiley/NBS Mass Spectral Database,[38] 4th edition. Selection of compounds was supported by the search capabilities of MassLib. A subset of 31 000 mass spectra fulfilling some plausibility criteria for spectral quality has been used in this work.

NIST Mass Spectral Data Base. Version 4.0 is running under MS-DOS and contains mass spectra from more than 62 000 compounds.[39]

EDAS. This software for exploratory data analysis of spectra[29,40] is running under VMS on a Vaxstation 4000/60. It contains graphics-oriented tools for investigating spectra-structure-relationships and developing spectral classifiers. The implemented methods are based on multivariate statistics. EDAS allows direct access to spectral data in MassLib.

INSPECT is running under MS-DOS and contains a collection of methods for interpreting numerical data, including multivariate statistics and neural networks.[43]

ToSiM is running under MS-DOS; it contains tools for investigating topological similarities of molecules, such as cluster analysis of chemical structures, determination of large and maximum common substructures, and determination of equivalent atoms and bonds in a molecule.[17,41,42] ToSiM is capable of handling data bases containing far more than 100 000 chemical structures (and optionally also mass spectra). The implemented fast substructure search has been used for building random samples of spectra for classifier development and also for evaluating candidate lists as produced by isomer generators. Spectral and structural data of 46 000 compounds from MassLib, and 60 000 compounds from the NIST data base were converted to be directly accessible by ToSiM.

CHEMICS is a powerful computer-assisted structure elucidation system for organic compounds;[8] version 9 was implemented on a Vaxstation 4000/60 under VMS. The isomer generator in CHEMICS builds all possible chemical structures for a given brutto formula. Spectral data are used to select substructures (called components in CHEMICS) that cannot be excluded by the spectral data. In this work only ^{13}C-NMR data have been used, but CHEMICS also accepts infrared- and ^{1}H-NMR data as well as user defined substructures for a goodlist and a badlist. Recent developments of CHEMICS include the use of two-dimensional NMR data.[44] Lists of molecular candidate structures generated by CHEMICS have been transferred to software ToSiM for further evaluation.

MOLGEN. Version 3.0 of this software[15,45] was used under MS-Windows. MOLGEN computes complete sets of connectivity isomers for given brutto formulas. The construction of isomers is redundancy free, complete, and fast; it can be restricted by a goodlist and a badlist. Furthermore lower and upper limits can be defined for ring size as well as a maximum bond multiplicity. The generated outputs, containing connection tables of chemical structures, have been transferred to software ToSiM for further evaluation.

RESULTS

1. Development of Classifiers. The general scheme of classifier development is shown in Figure 4. Selection of

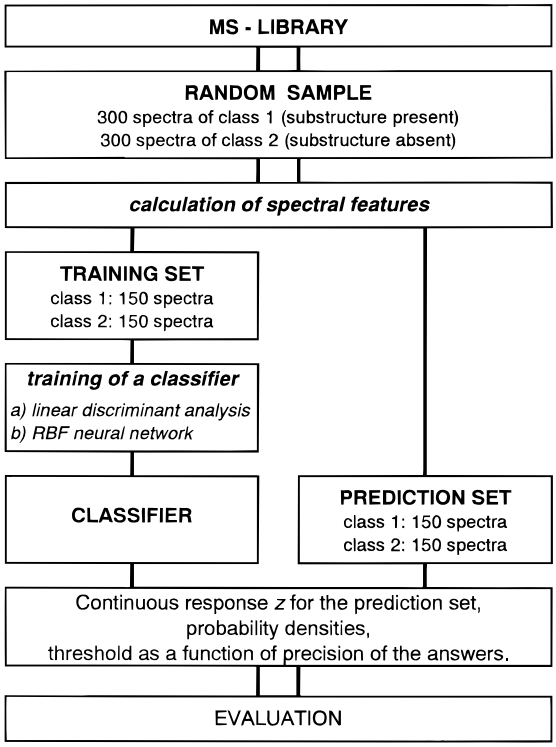


Figure 4. Development and test of a mass spectral classifier.

random samples from mass spectral libraries was performed by substructure searches (using software ToSiM) or by searches using molecular descriptors (as implemented in MassLib). Each training set consisted of 150 compounds from class 1 (substructure present in the molecule) and 150 compounds from class 2 (substructure absent). Prediction sets had the same size but contained only compounds not present in the corresponding training set. For some substructures the libraries contain less than 300 entries and therefore somewhat smaller data sets had to be used.

LDA classifiers were calculated by software EDAS from data sets containing ten features (selected by maximum Fisher ratios). RBF classifiers were developed by software INSPECT; the applied stepwise feature selection resulted in sets of two to ten features (mean 4.7) per classifier. Typical computing time (including feature generation and selection) was for an LDA classifier 60 s (on a Vaxstation 4000/60) and for an RBF classifier 70 h (on a PC-386, 40 MHz). The most frequently selected features were logarithmic intensity ratios, followed by autocorrelation features, and intensity sum features.

Spectra of the corresponding prediction set were used to estimate probability densities of the continuous classifier response *z*. Classification thresholds for precisions between 50 and 99% have been derived as shown in Figure 2. All parameters of a classifier that are necessary for application to spectra are stored in an ASCII file which is separated from software and can be considered as the knowledge base.

2. Classified Substructures. Selection of substructures, or more general structural properties, which may be reflected by mass spectral data is a difficult problem. Because no sufficient spectroscopic knowledge is usually available for predicting suitable substructures, a trial and error approach had to be used, including already existing proposals from mass spectroscopic data base systems. In MassLib chemical structures are coded by a set of 180 binary molecular descriptors;[19,26] most of them can be described by a substructure; 162 of them have been investigated in this work. Another source for substructures was STIRS.[20,46] From the suggested more than 600 structural properties in STIRS 70 have been selected that can be described by a Boolean combination of substructures. Another 30 substructures have been defined because of their high relevance in organic chemistry. By applying the two training methods LDA and RBF and using different random samples of spectra, more than 600 classifiers have been developed up to now.

Table 1. Substructures and Compound Classes for Which Mass Spectral Classifiers Are Available Together with Software MSclass[a]

classifier group	substructure or class of compounds		
alkyl	hydrocarbon	$(CH_3)_3C$	C_nH_{2n+1} $n = 4-11$
aromatic compounds	bz	bz-O	bz-N
	bz-C	bz-O-CH_2	bz-S
	bz-CH	bz-O-CH_3	bz-F
	bz-CH_2CH_2	bz-CH_2-O	bz-Cl
		bz-CO	bz-Br
		bz-CO-CH_2	
		bz-CO-O	
	phenyl	alkyl-substituted phenols	
	naphthyl	alkyl-substituted chlorophenols	
	biphenyl	condensed aromatic rings	
functional groups	CH_3CH_2O	CH_2CH_2NH	
	$CH_3(CH_2)_3O$	$(CH_3)_2N$	
	tertiary alcohol	tertiary amine	
	CH_3OCH_2	CH_2-S	
	CH_3CO	$(CH_3)_3Si$	
	CH_3CH_2CO	$(CH_3)_3SiO$	
	CH_3COO	CF_3	
	CH_3COOCH methyl ester ethyl ester	CF_3O	
elements	N_x, Cl_x, Br_x, P_x, S_x, B_x, Si_x		

[a] bz, benzene ring; CO, carbonyl; free valences, H or any substitution.

The following criteria were applied for selecting classifiers for potential applications in structure elucidation. (A) Classifiers were discarded if visual inspection of the probability density curves showed great deviations from the ideal shape as given in Figure 2. (B) Classifiers were discarded if recall (eqs 3 and 4) was below 30% for both classes at a minimum precision of 90%. About 160 classifiers survived this selection and were compiled to a knowledge base for the new software MSclass (described below). Table 1 presents an overview of substructures and compound classes for which classifiers are currently available. Only a small sector of organic chemistry is covered by these classifiers. Furthermore, it has to be considered that a sharp definition of compound classes is often not feasible. Consider for instance a classifier for substructures with formula C_7H_{15}. In this case aim of the training was to recognize compounds containing alkyl groups C_7H_{15}. However, as tests indicated *yes* answers are also obtained for a few compounds with a slightly smaller or larger alkyl group.

Table 2 contains a selection of 16 classifiers, characterized by the recalls for both classes (eqs 3 and 4) and the average precisions of *yes* and *no* answers (eqs 5 and 6); classification thresholds were adjusted to obtain a minimum precision of 90%. In Table 3 the averaged recall values for 89 LDA

Table 2. Examples for Mass Spectral Classifiers

substructure or compound class (class 1)	method	for a minimum precision[a] of 90% recall[b] % class 1	recall[b] % class 2	average precision[c] % *yes*	average precision[c] % *no*
C7 H15	LDA	41	53	97	98
tert-butyl	RBF	47	55	93	93
phenyl	LDA	45	41	97	97
aryl - CH	LDA	10	55	97	98
aryl - N	LDA	17	38	94	96
aryl - Cl	RBF	66	73	96	98
biphenyl	LDA	33	50	95	97
acetyl	LDA	61	73	96	98
methyl ester	LDA	61	79	98	96
ethyl ester	RBF	67	61	96	95
dimethylamino	RBF	72	68	93	98
trimethylsilyl	RBF	93	98	99	99
nitrogen (any no. of atoms)	RBF	36	7	97	96
chlorine (any no. of atoms)	RBF	51	0	96	0
bromine (any no. of atoms)	RBF	81	73	95	97
phosphorous (any no. of atoms)	LDA	34	19	98	97

[a] Minimum probability for a correct *yes/no* answer (eqs 1 and 2). [b] Percentage of correctly classified spectra (eqs 3 and 4). [c] Average probability for a correct answer *yes/no* (eqs 5 and 6).

Table 3. Averaged Recalls of Mass Spectral Classifiers

method	no. of classifiers	recall (%) of class 1 at minimum precision 90%	95%	99%	recall (%) of class 2 at minimum precision 90%	95%	99%
LDA	89	39	27	13	48	37	23
RBF	78	43	25	11	48	33	15
LDA + RBF	167	41	26	12	48	35	19

classifiers, 78 RBF classifiers, and all 167 classifiers are listed. In general, class 2 has higher recall than class 1; that means it is more successful (or easier) to predict the absence of a substructure than its presence. If a minimum precision of 90% is required, less than half of the spectra (41% of class 1, 48% of class 2) results in a *yes* or *no* answer. No significant difference appears between the performances of LDA and RBF classifiers, although the stepwise feature selection used for the nonlinear classifiers required an enormous computing effort. In general, recall decreases almost linearly when increasing the minimum precision from 90% to 99%.

As illustrated in Table 2 the investigated substructures exhibit quite different classification behavior. In many cases the recall is in the region of 40−70% for both classes; a few classifiers reach more than 90% recall. Some classifiers are able to recognize only one of the classes with a sufficient precision of the answer. An example for this behavior is the chlorine classifier; 51% of chlorine-containing compounds resulted in the correct answer *yes* (the remaining 49% did not give an answer); however it was impossible to reach 90% precision for *no* answers and therefore recall for not-chlorine-containing compounds is zero.

3. Classifier Example. An LDA classifier to recognize the phenyl substructure (C_6H_5-) is described here in more detail. Figure 5 shows probability densities, precision of *yes/no* answers, and recall for both classes as obtained by testing the classifier with a prediction set containing 150 spectra from each class. Assuming a minimum precision of 90%

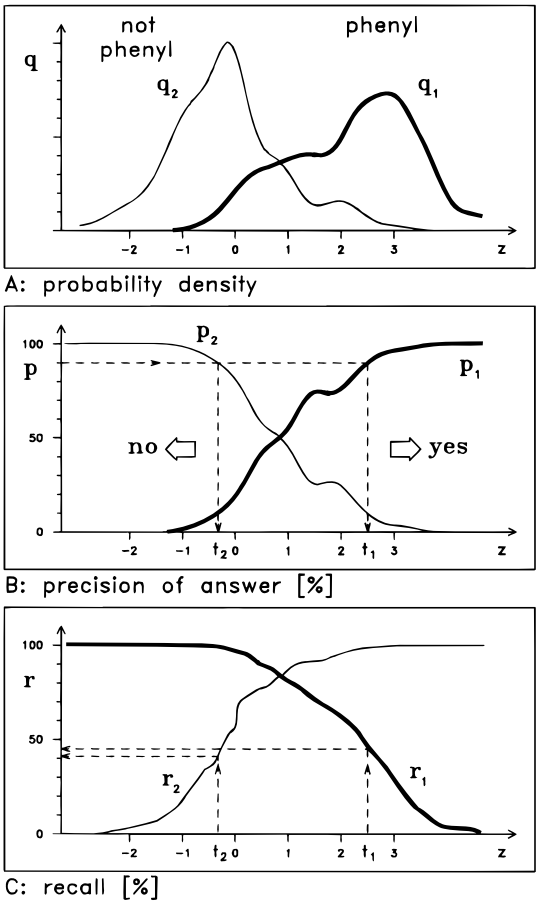


Figure 5. LDA classifier for the recognition of phenyl-substructures.

Table 4. Test Results for the Phenyl Classifier as Defined by Eq 16 Using Large and Small Data Sets

	answers *yes* #[a]	answers *yes* %	answers *no* #[a]	answers *no* %	rejections #[a]	rejections %	sum #[a]	sum %
				Phenyl				
large set	2807	*46.32*	82	*1.35*	3171	*52.33*	6060	*100*
small set	67	*44.67*	1	*0.67*	82	*54.67*	150	*100*
				Not-Phenyl				
large set	209	*3.59*	2283	*39.20*	3332	*57.21*	5824	*100*
small set	2	*1.33*	62	*41.33*	86	*57.33*	150	*100*

[a] #, number of spectra or answers; minimum precision was 90%.

for *yes* and for *no* answers the recall is 45% for phenyl compounds and 41% for compounds not containing a phenyl substructure (numerical values are given in Table 2). Actually, for 67 spectra (44.7%) of the tested 150 phenyl compounds the correct *yes* answer was obtained; for 82 spectra (54.7%) the answer was refused, and one spectrum (0.7%) was classified erroneously. From the 150 not-phenyl compounds 61 (40.7%) were correctly classified by the answer *no*, 87 (58.0%) were not classified, and 2 (1.3%) were erroneously classified. Average precision (eqs 5 and 6) of the *yes* answers as well as that of the *no* answers are both 97% (assuming equal *a priori* probabilities of both classes).

All ten features for this classifier are logarithmic intensity ratios. The continuous classifier response *z* is given by

$$z = 0.273 \ln(I_{51}/I_{53}) + 0.299 \ln(I_{51}/I_{52}) + 0.163 \ln(I_{77}/I_{79}) + 0.328 \ln(I_{49}/I_{51}) + 0.053 \ln(I_{78}/I_{80}) + 0.391 \ln(I_{52}/I_{53}) - 0.252 \ln(I_{75}/I_{77}) - 0.475 \ln(I_{50}/I_{51}) - 0.035 \ln(I_{76}/I_{77}) + 0.297 \ln(I_{91}/I_{93}) \quad (16)$$

I_m is the peak intensity at mass m, normalized to a base peak intensity of 100. The final classification answer for a minimum precision of 90% is obtained by comparing z with two thresholds:

IF $z > 2.25$ THEN answer = *yes* (phenyl present)

IF $z < -0.31$ THEN answer = *no* (phenyl absent)

ELSE rejection of classification

Thresholds for 95% minimum precision of answers would be 2.50 and −0.54. Because eq 16 includes scaling of features, the resulting rejection interval does not correspond to the original target values +1 and −1 for z. From the mass spectroscopist's point of view most of the automatically selected features are evident; for instance mass 77 corresponds to $C_6H_5^+$; mass 51 is the typical fragment ion $C_4H_3^+$ of aromatic compounds. Note, that presence of a peak at mass 77 alone is not significant for monosubstituted benzene rings because probability density curves of peak intensities at this mass overlap considerably. The features of the classifier are intensity ratios and thereby reflect the different yields of competing fragmentation processes which are more characteristic than absolute peak intensities.

The phenyl classifier as defined by eq 16 has also been tested with larger data sets than used for training and deriving the classification thresholds. For class 1 a set containing 6060 mass spectra was selected from the NIST mass spectral data base;[39] actually all compounds in the molecular weight range 78−300 containing a phenyl group but no metal atom were selected by software ToSiM. For class 2 a corresponding set containing randomly selected 5824 not-phenyl compounds was prepared. Results for the large spectra sets are very similar to those obtained from the prediction sets (Table 4). It can be concluded that the generally used small prediction sets gave good estimations of the classifier performances.

4. Software MSclass. For practical applications of mass spectral classifiers the software MSclass,[47,48] running under MS-DOS, has been developed. It contains subroutines for calculation of spectral features; the necessary classifier parameters are read from an ASCII file (currently containing 167 classifiers for about 70 substructures or structural properties). Operation of MSclass is graphics- and mouse-oriented; the user does not need any knowledge about the mathematics of the classifiers. Import of some proprietary mass spectral data formats is supported. After selection of mass spectra to be investigated the classification process runs automatically. Result is a list (or file) for each mass spectrum containing *yes/no* answers. The minimum precision of answers can be adjusted by the user. Typical computing time for 167 classifiers is 2 s per mass spectrum (PC-386, 40 MHz). Furthermore, user-selected single classifiers can be applied as a chemometric detector to spectra series as obtained in GC/MS analyses; result is a chromatogram that selectively indicates a particular class of compounds.[27,47]

EXAMPLES

The examples presented here are simple problems from structure elucidation in which classification of mass spectra was able to produce useful hints. A typical result for applying the currently available 167 classifiers to a mass spectrum is a list containing only a few *yes* answers but many *no* answers. About 40−70% of the classifiers do not give an answer for a certain mass spectrum at the 90% level for minimum precision. Numerous tests showed that *no* answers are almost always correct; *yes* answers may unfortunately be wrong. Because for most substructures more than one classifier is available, erroneous classifications can be detected if inconsistent results occur. If the brutto formula is considered to be known, then, in general, only a small number of classification answers remain relevant.

1. $C_6H_{12}O_2$ (MS). One of the 1313 isomers is *n*-pentanoic acid methyl ester; Figure 6 shows the mass spectrum[37] of this compound. Table 5 summarizes the classification results as obtained from 167 classifiers; all are correct in this example. Only three *yes* answers were given, all predicting the substructure methyl ester; 70 answers were *no*, and 94 classifiers did not result in an answer because precision was below 95%. The list of answers may support inexperienced persons in interpreting the mass spectrum. From the results in Table 5 one may conclude the unknown is a methyl ester, is not aromatic, and does not contain a CH_3OCH_2 substructure; furthermore it is probably not a hydrocarbon and does not contain nitrogen.

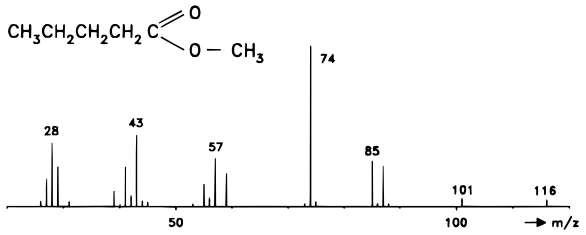


Figure 6. Mass spectrum of *n*-pentanoic acid methyl ester,[37] example 1.

If the brutto formula of the compound is assumed to be known a more systematic use of the classification results is possible.[47] Only two classification answers remain relevant: methyl ester (*yes*) and CH_3OCH_2- (*no*); the badlist substructure is even redundant, because if the unknown is a methyl ester it cannot contain this substructure. Considering these restrictions only four isomers are possible, namely the four isomers of pentanoic acid methyl ester. Discrimination between the four candidates is not possible with the available data but may be achieved for instance by using retention time data. Computation time on a PC-386, 40 MHz was 3 s for classification by software MSclass and 4 s for generation of all 1313 isomers by software MOLGEN.

2. C_7H_7OCl (MS). One of the 62 625 isomers is 2-chloro,5-methylphenol; Figure 7 shows the mass spectrum[37] of this compound. Application of 167 mass spectral classifiers resulted in 7 *yes* answers, 43 *no* answers, and 117 rejections (minimum precision was 95%). Considering the brutto formula only two goodlist substructures survive,

Table 5. Classification Result for Mass Spectrum of *n*-Pentanoic Acid Methyl Ester (Example 1)[a]

answers	prec	substructure or class of compounds
YYY	97	methyl ester
N	99	alkyl $C_{10}H_{21}$
N	99	alkyl $C_{11}H_{23}$
NN	99	hydrocarbon
NN	98	aromatic compound
NN	99	aryl-CH═O
NN	98	aryl- −C−O or −C═O or −N═N
N	97	aryl- $-CH_2$ or $-CH_3$
NN	96	aryl- −N═ or −NHN
NN	96	aryl- C (ring bond)
N	97	aryl- C═O
N	98	aryl-CH
N	98	aryl-CH_2CH_2
N	99	aryl-Cl
NN	99	aryl-$COCH_2$
N	99	aryl-F
N	96	aryl-N
N	96	aryl-N (incl. NO_x)
N	99	aryl-N (ring-bond)
NN	98	aryl-OCH_2
N	99	aryl-OCH_3
NN	99	aryl-S (S in ring)
N	99	C_6H_4-Br (o,m,p)
N	98	$C_6H_5CH_2O$
NN	99	C_6H_5-
NN	98	$CH_2C_6H_4O-$ (*o*,*m*,*p*)
NN	97	condensed rings
N	99	ester $C_6H_4COOCH_2-$ (and subst. at *o*, *m*, or *p*)
NN	99	more than 1 aromatic ring (any type)
NN	99	phenol (1 OH), alkyl-subst.
NN	97	phenol (1−3 OH), alkyl-subst.
NN	99	phenol-Cl (1 OH, 1 Cl), alkyl-subst.
NN	99	phenol-Cl (1−3 OH, 1−3 Cl), alkyl-subst.
N	98	boron (any number)
N	98	phosphorous (any number)
N	99	silicon (any number)
NN	99	silicon: ≥ 2 atoms
N	99	CF_3
NN	96	CF_3CO
N	99	CH_3OCH_2
NN	99	$N(CH_3)_2$
NNNN	99	$Si(CH_3)_3$
NN	99	$OSi(CH_3)_3$
NN	96	ring: CHCCH bridge, typic. terpene

[a] Y, answer *yes*; N, answer *no* (the number of characters indicates how many classifiers gave an answer for the particular substructure); prec, precision of answer (%, averaged if more than one classifier answer). Minimum precision, 95%; number of *yes* answers, 3; number of *no* answers, 70; number of rejections of answer, 94; sum of used classifiers, 167.

namely phenol and chlorinated benzene ring (note that goodlist substructures may overlap). Assuming these structural restrictions only ten molecular structures are possible, the ten isomers of methylchlorophenol. Computation time on a PC-386, 40 MHz was 3 s for classification by software MSclass and 110 s for generation of all 62 625 isomers but only 0.4 s if the two goodlist substructures are considered, by software MOLGEN.

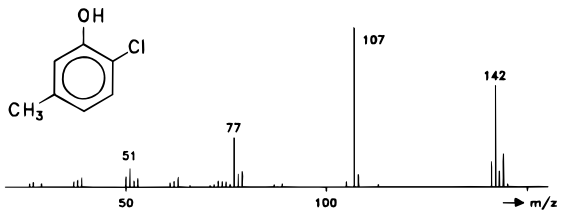
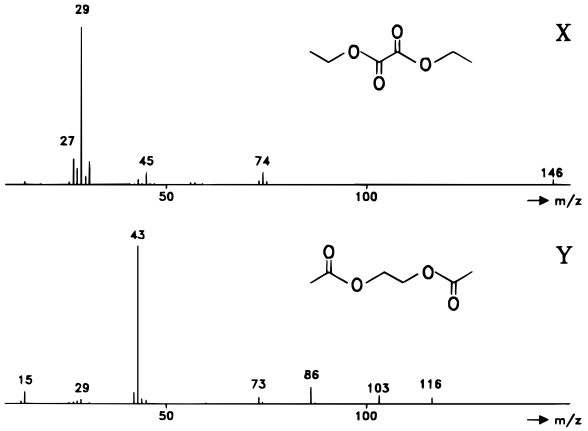


Figure 7. Mass spectrum of 2-chloro,5-methylphenol,[37] example 2.

3. $C_6H_{10}O_4$ (^{13}C-NMR + MS). Two of the 97 394 isomers have been investigated: ethanedioic acid, diethyl ester (**X**) and 1,2-ethanediol, diacetate (**Y**); Figure 8 shows ^{13}C-NMR data[11] and mass spectra[37] of these compounds. Results from a combined use of both spectroscopic data by CHEMICS and MSclass is summarized in Figure 9.



^{13}C-NMR chemical shifts (multiplicity)
X: 13.9 (Q), 62.9 (T), 158.0 (S), **Y**:20.7 (Q), 62.3 (T), 170.6 (S)

Figure 8. Mass spectra[39] and ^{13}C-NMR data[11] for the two compounds ethanedioic acid, diethyl ester (X), and 1,2-ethanediol, diacetate (Y), example 3.

Table 6. Classification Results for Mass Spectra from Example 3 (**X**, Ethanedioic Acid, Diethyl Ester and **Y**, 1,2-Ethanediol, Diacetate[a]

	no. of answers				substructures	
compd	*yes*	*no*	reject.	wrong	present	absent
X	3	97	67	0	C_2H_5COO-	methyl ester
						CH_3CO-
						CH_3COO-
						C_2H_5CO-
						n-C_4H_9O-
						tertiary alcohol
						$(CH_3)_3C-$
						$(CH_3)_2C{=}C<$
						(ring)C═C(ring)
Y	7	76	84	0	CH_3CO-	methyl ester
					CH_3COO-	CH_3OCH_2-
					CH_3COOCH_2-	C_2H_5CO-
						$(CH_3)_3C-$
						$(CH_3)_2C{=}C<$

[a] Classified substructures are only given if in agreement with the brutto formula $C_6H_{10}O_4$.

Structure elucidation system CHEMICS has been applied to obtain all molecular candidate structures agreeing with the ^{13}C-NMR spectra. From the 630 substructures (components) that are considered in CHEMICS a set of 17 survived for compound **X** and a set of 16 for compound **Y**. The isomer generator implemented in CHEMICS produced seven molecular candidates for **X** and five for **Y**. Computation time on a Vaxstation 4000/60 was 30 s for each of the two compounds.

Results of applying 167 mass spectral classifiers by software MSclass are summarized in Table 6. All *yes* answers and all *no* answers given at the 90% minimum precision level were correct. The recognized substructures are partly redundant and partly overlapping. Bad- and goodlist for isomer generation by software MOLGEN have

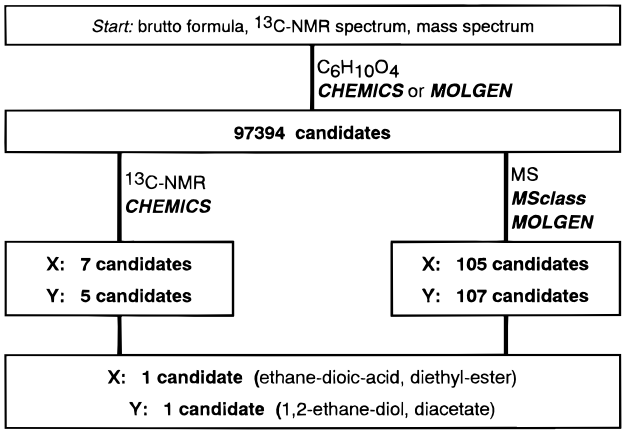


Figure 9. Automatic structure elucidation of the two compounds **X** and **Y**, both with brutto formula $C_6H_{10}O_4$, based on mass spectra and ^{13}C-NMR data, using software CHEMICS, MSclass, and MOLGEN (example 3).

been built by considering the known brutto formula. Structural restrictions obtained from mass spectra are less selective than those obtained from ^{13}C-NMR data; the number of candidate structures when using only mass spectral data is 105 for compound **X** and 107 for compound **Y**.

Combination of the candidate lists obtained from CHEMICS and MSclass by a logical intersection gives as a final result only one (the correct) candidate for each compound.

4. $C_8H_{10}O_3$ (^{13}C-NMR + MS). Mass spectrum[11] and ^{13}C-NMR data[11] of 3-acetoxy-2,4-dimethylfuran are given in Figure 10. The number of isomers is 3 868 967 (counted by software MOLGEN, computation time 70 min). Using the^{13}C-NMR data CHEMICS was able to reduce the candidate list to 161 molecular structures (computation time 210 s on a Vaxstation 4000/60). MSclass produced at 90% minimum precision 3 *yes* answers, 63 *no* answers, and 118 rejections; all answers are correct; goodlist and badlist are given in Table 7. The only relevant information for reducing the candidate list is "presence of an acetoxy group". This substructure is present in 36 of the 161 candidate structures. Furthermore, eight candidates can be eliminated, because their structures have only seven topological different carbon atoms which contradicts to the eight peaks in the ^{13}C-NMR spectrum.

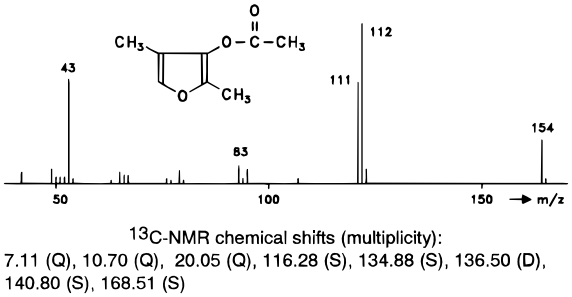


Figure 10. Mass spectrum[11] and ^{13}C-NMR data[11] for 3-acetoxy-2,4-dimethylfuran, formula $C_8H_{10}O_3$ (example 4).

The remaining set of 28 candidates contains very different chemical structures. Further evaluation is supported by a cluster analysis of chemical structures. It has been shown[17] that for sets containing a maximum of about 1000 structures principal component analysis (PCA) is a powerful tool for getting insight into the structural variety. Application of this method is briefly demonstrated here for the final set of 28

Table 7. Classification Results for Mass Spectrum of 3-Acetoxy-2,4-dimethylfuran (Example 4)[a]

structural restriction	substructure or class of compounds		
goodlist total of 3 *yes* answers	CH_3CO-	CH_3COO-	
badlist total of 63 *no* answers	aromatic substructures with 12 different substituents C_2H_5O- $-(CH_2)_6CO-$ (ring)C=C(ring)	n-C_4H_9O- methyl ester	CH_3OCH_2- ethyl ester

[a] Classified substructures are only given if in agreement with the brutto formula $C_8H_{10}O_3$.

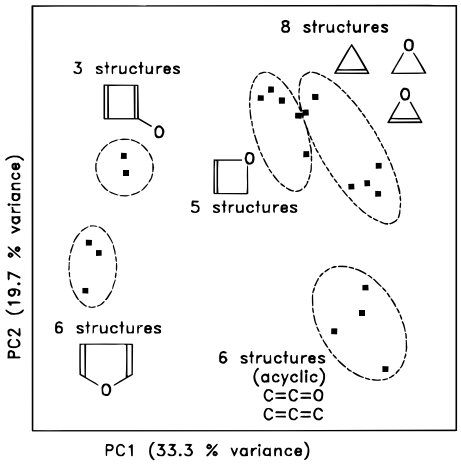


Figure 11. Cluster analysis of 28 remaining candidate structures of example 4. Structures have been characterized by 20 binary descriptors and clustered by principal component analysis using software ToSiM.

structures. Chemical structures have been characterized by a set of 165 binary molecular descriptors;[17,42] 20 descriptors with maximum variance have been used for PCA. Projection of the data onto the first and second principal component separates well the different classes of compounds as shown in Figure 11. For instance all substituted furans are located in a single cluster; other clusters only contain acyclic compounds, or only compounds with a cyclobutadiene substructure, or only compounds with a three-membered ring. Potential applications of this data presentation for a selection of appropriate prototype structures prior to spectra simulation is under investigation.

CONCLUSION

Mass spectra have been characterized by a set of several hundred numerical spectral features. Automatic selection of the most relevant features resulted in multivariate data suitable for the development of classifiers that recognize presence or absence of substructures in unknown molecules from low resolution mass spectral data. However, simple *yes/no* classifiers did not possess sufficient performance because of the too large number of wrong answers. Higher precision of substructure prediction has been achieved by restricting the output of answers to cases in which the estimated probability for giving a correct answer reaches a certain threshold (for instance 90%). The cost of this approach is rejection of 40−70% of the classification answers because of too low precision.

No answers were almost always correct; *yes* answers were sometimes wrong. Incorrect classifications can be detected

by the appearance of contradictory results in the list of answers; a software tool for supporting evaluation of classification results is under development. Examples demonstrate that the obtained list of answers often supports spectra interpretation and structure elucidation. For small monofunctional molecules a systematic and in some cases complete structure elucidation is possible even if only mass spectral data are available. In general, however, additional spectroscopic data or structure information is necessary to end up with a reasonable-sized list of candidate structures.

Combined use of ^{13}C-NMR data interpretation by the structure elucidation system CHEMICS and of mass spectra classification demonstrated that mass spectra may provide complementary structural information to that obtained from ^{13}C-NMR. Candidate lists solely based on NMR data can be reduced effectively by considering the results of MS classifications.

Aim of this work was to investigate methods for a systematic structure elucidation of organic compounds applicable also in trace analysis. Systematic means that the different steps in the interpretation process are strictly defined: First step is a set-up of premises about the unknown structure; second step is the complete and redundancy-free generation of all possible candidate structures that do not contradict with the premises. Obtaining suitable restrictions for an unknown chemical structure is still difficult; this work demonstrates that it can be supported by mass spectral classifiers. Human interpreters and library search systems are in general not capable of strictly separating these two steps and producing complete solutions.

A preliminary set of mass spectral classifiers has been implemented for use by the new software MSclass. The reported method is available for tests by practicing spectroscopists. A fruitful use of this approach, however, can only be expected if the limitations are considered. A statistics-based method can only produce suggestions; accepting them or not is within the responsibility of the chemist. Classification of mass spectra is considered as a complementary technique to library search.

In the present status only a small area of organic chemistry is covered by the available classifiers. Substantial progress can be expected by definition of new spectral features (eventually based on spectroscopic interpretation rules) and by a large scale development of up to some 10 000 classifiers for a great variety of substructures and structural properties.

ACKNOWLEDGMENT

This work was supported by the Austrian ministry *Bundesministerium für Wissenschaft und Forschung*. We express our deep gratitude to S. I. Sasaki and K. Funatsu (*Toyohashi University of Technology*, Japan) for making available their software CHEMICS. We also thank R. Neudert of *Chemical Concepts* (Weinheim, Germany) and D. Henneberg and E. Ziegler (*Max-Planck-Institut für Kohlenforschung*, Mülheim a.d. Ruhr, Germany) for providing mass spectral libraries. We are grateful to A. Kerber and R. Laue (*University of Bayreuth*, Germany) for making available their software MOLGEN. We wish to thank F. Stancl, H. Lohninger, W. Czerni, and T. Kirchtag for fruitful discussions, software development, and calculating mass spectral classifiers.

REFERENCES AND NOTES

(1) Gray, N. A. B. *Computer-Assisted Structure Elucidation*; John Wiley: New York, 1986.
(2) Hippe, Z. *Artificial Intelligence in Chemistry*; Elsevier: Amsterdam, 1991.
(3) Varmuza, K. *Pattern Recognition in Chemistry*; Springer-Verlag: Berlin, 1980.
(4) Zupan, J. *Algorithms for Chemists*; John Wiley: Chichester, 1989.
(5) Adams, M. J. *Chemometrics in Analytical Spectroscopy*; The Royal Society of Chemistry: Cambridge, 1995.
(6) Cadish, M.; Farkas, M.; Clerc, T. J.; Pretsch, E. SpecTool: A Hypermedia Toolkit for Structure Elucidation. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 286−290.
(7) Carhart, R. E.; Smith, D. H.; Gray, N. A. B.; Nourse, J. G.; Djerassi, C. GENOA: A Computer Program for Structure Elucidation Utilizing Overlapping and Alternative Substructures. *J. Org. Chem.* **1981**, *46*, 1708−1718.
(8) Funatsu, K.; Miyabayashi, N.; Sasaki, S. I. Further Development of Structure Generation in the Automated Structure Elucidation System CHEMICS. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 18−28.
(9) Gribov, L. A.; Elyashberg, M. E.; Koldashov, V. N.; Pletnjov, I. V. A Dialogue Computer Program System for Structure Recognition of Complex Molecules by Spectroscopic Methods. *Anal. Chim. Acta* **1983**, *148*, 159−170.
(10) Christie, B. D.; Munk, M. E. Structure Generation by Reduction: A New Strategy for Computer-Assisted Structure Elucidation. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 87−93.
(11) *SpecInfo: Spectroscopic Information System*, vers. 3.0, 1995. Available from: Chemical Concepts, P.O. Box 100202, D-69442 Weinheim, Germany.
(12) Martinelli, E.; Redeker, D.; Neudert, R.; Körnig, S. The Spectroscopic Interpretation System SpecInfo. Past, Present and Future. *Chim. Oggi* **1994**, *12*, 33−37.
(13) Kalchhauser, H.; Robien, W. CSEARCH: A Computer Program for Identification of Organic Compounds and Fully Automated Assignment of Carbon-13 Nuclear Magnetic Resonance Spectra. *J. Chem. Inf. Comput. Sci.* **1985**, 5*8*, 103−108.
(14) Thiele, H. *X-PERT: A New Expert System for Structure Elucidation*. In *Software Development in Chemistry*; Moll, R., Ed.; Springer-Verlag: Berlin, 1995; Vol. 9, pp 305−317.
(15) Grund, R.; Kerber, A.; Laue, R. MOLGEN, ein Computeralgebra-System für die Konstruktion molekularer Graphen. *MATCH* **1992**, *2*, 87−131.
(16) Lipkus, A. H.; Munk, M. E. Automated Classification of Candidate Structures for Computer-Assisted Structure Elucidation. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 9−18.
(17) Varmuza, K.; Scsibrany, H. Cluster Analysis of Chemical Structures Based on Binary Molecular Descriptors and Principal Component Analysis. In *Software Development in Chemistry*; Moll, R., Ed.; Springer-Verlag: Berlin, 1995; Vol. 9, pp 81−90.
(18) Stein, S. E. Chemical Substructure Identification by Mass Spectral Library Searching. *J. Am. Soc. Mass Spectrom.* **1995**, *6*, 644−655.
(19) Henneberg, D.; Weimann, B.; Zalfen, U. Computer-Aided Interpretation of Mass Spectra Using Data Bases with Spectra and Structures. I. Structure Searches. *Org. Mass Spectr.* **1993**, *28*, 198−206.
(20) Kwok, K. S.; Venkataraghavan, R.; McLafferty, F. W. Computer-Aided Interpretation of Mass Spectra. III. A Self-Training Interpretive and Retrieval System. *J. Am. Chem. Soc.* **1973**, *95*, 4185−4194.
(21) Crawford, L. R.; Morrison, J. D. Computer Methods in Analytical Mass Spectrometry. Empirical Identification of Molecular Class. *Anal. Chem.* **1968**, *40*, 1469−1474.
(22) Jurs, P. C.; Kowalski, B. R.; Isenhour, T. L. Computerized Learning Machines Applied to Chemical Problems. Molecular Formula Determination from Low Resolution Mass Spectrometry. *Anal. Chem.* **1969**, *41*, 21−27.
(23) Chapman, J. R. *Computers in Mass Spectrometry*; Academic Press: London, 1978.
(24) Luinge, H. J. A Knowledge-Based System for Structure Analysis from Infrared and Mass Spectral Data. *Trends Anal. Chem.* **1990**, *9*, 66−69.
(25) Varmuza, K. Chemometrics in Mass Spectrometry. *Int. J. Mass Spectrom. Ion Proc.* **1992**, *118/119*, 811−823.
(26) Werther, W.; Lohninger, H.; Stancl, F.; Varmuza, K. Classification of Mass Spectra. A Comparison of yes/no Classification Methods for the Recognition of Simple Structural Properties. *Chemometrics Intelligent Laboratory Systems* **1994**, *22*, 63−76.
(27) Lohninger, H.; Varmuza, K. Selective Detection of Classes of Chemical Compounds by Gas Chromatography / Mass Spectrometry / Pattern Recognition: Polycyclic Aromatic Hydrocarbons and Alkanes. *Anal. Chem.* **1987**, *59*, 236−244.
(28) Erni, F.; Clerc, J. T. Strukturaufklärung organischer Verbindungen durch computerunterstützten Vergleich spektraler Daten. *Helv. Chim. Acta* **1972**, *55*, 489−500.

(29) Werther, W.; Varmuza, K. EDAS-MS: Exploratory Data Analysis of Mass Spectra. In *Software-Development in Chemistry*; Gasteiger, J., Ed.; Springer-Verlag: Berlin, 1990; Vol. 4, pp 175−185.
(30) Curry, B.; Rumelhart, D. E. MSnet: A Neural Network which Classifies Mass Spectra. *Tetrahedron Comput. Methodol.* **1990**, *3*, 213−237.
(31) Nekrasov, Y. S.; Sukharev, Y. N.; Molgacheva, N. S.; Tepfer, E. E.; Nekrasov, S.Y. Generalized Characteristics of Mass Spectra of Aromatic Compounds and their Correlation with the Constants of Substituents. *Russ. Chem. Bull.* **1993**, *42*, 1986−1990.
(32) Duda, R. O.; Hart, P. E. *Pattern Classification and Scene Analysis*; John Wiley: NewYork, 1973.
(33) Lohninger, H. Feature Selection Using Growing Neural Networks: The Recognition of Quinoline Derivatives from Mass Spectral Data. In *Software Development in Chemistry*; Ziessow, D., Ed.; Gesellschaft Deutscher Chemiker: Frankfurt am Main, 1993; Vol. 7, pp 25−37.
(34) Massart, D. L; Vandeginste, B. G. M.; Deming, S. N.; Michotte, Y.; Kaufmann, L. *Chemometrics: a Textbook*; Elsevier: Amsterdam, 1988.
(35) Jokinen, P. A. Dynamically Capacity Allocating Neural Networks for Continuous Learning Using Sequential Processing of Data. *Chemometrics Intelligent Laboratory Systems* **1991**, *12*, 121−145.
(36) Lohninger, H. Evaluation of Neural Networks Based on Radial Basis Functions and Their Application to the Prediction of Boiling Points From Structural Parameters. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 736−744.
(37) *MassLib: Mass Spectra Database and Information System*; developed by Henneberg, D., Weimann, B., and Ziegler, E. (Max-Planck-Institut für Kohlenforschung, Mülheim/Ruhr, Germany). Available from: Chemical Concepts, P.O. Box 100202, D-69442 Weinheim, Germany.
(38) *Wiley Mass Spectral Database*, 4th ed.; Electronic Publishing Division, John Wiley & Sons, Inc.: New York.
(39) *NIST Mass Spectral Database*, version 4.0; National Institute of Standards and Technology: Gaithersburg, MD 20899, 1992. Available from: HD Science Ltd, 4a Bessell Lane, Nottingham NG9 7BX, UK.
(40) Varmuza, K.; Werther, W.; Henneberg, D.; Weimann, B. Computer-Aided Interpretation of Mass Spectra by a Combination of Library Search with Principal Component Analysis. *Rapid Commun. Mass Spectrom.* **1990**, *4*, 159−162.
(41) Scsibrany, H.; Varmuza, K. ToSiM: PC-Software for the Investigation of Topological Similarities in Molecules. In *Software Development in Chemistry*; Jochum, C., Ed.; Gesellschaft Deutscher Chemiker: Frankfurt am Main, 1994; Vol. 8, pp 235−249.
(42) *ToSiM: Software for Investigation of Topological Similarities of Molecules*. Available from Varmuza, K., Department of Chemometrics, Technical University Vienna, Lehargasse 4/152, A-1060 Vienna, Austria.
(43) Lohninger, H. INSPECT: a Program System to Visualize and Interpret Chemical Data. *Chemometrics Intelligent Laboratory Systems* **1994**, *22*, 147−153.
(44) Funatsu, K.; Susuta, Y.; Sasaki, S. I. Introduction of Two-Dimensional NMR Spectral Information to an Automated Structure Elucidation System CHEMICS. *J. Chem. Inf. Comput. Sci.* **1989**, *93*, 6−11.
(45) *MOLGEN: Isomer Generator Software*. Available from Kerber, A. and Laue, R., University of Bayreuth, Institute for Mathematics II, D-95440 Bayreuth, Germany.
(46) Haraki, K. S.; Venkataraghavan, R.; McLafferty, F. W. Prediction of Substructures from Unknown Mass Spectra by the Self-Training Interpretive and Retrieval System. *Anal. Chem.* **1981**, *53*, 386−392.
(47) Varmuza, K.; Stancl, F.; Lohninger, H.; Werther, W. Automatic Recognition of Substance Classes From Data Obtained by Gas Chromatography/Mass Spectrometry. *Chemometrics Intelligent Laboratory Systems*, in press.
(48) *MSclass: Software for Classification of Mass Spectra.* Authors: Stancl, F.; Lohninger, H.; Varmuza, K.. Available from Varmuza, K., Department of Chemometrics, Technical University Vienna, Lehargasse 4/152, A-1060 Vienna, Austria.

CI9501406