



ELSEVIER

Analytica Chimica Acta 388 (1999) 145–159

ANALYTICA  
CHIMICA  
ACTA

# Automatic classification of infrared spectra using a set of improved expert-based features

Plamen N. Penchev<sup>a</sup>, George N. Andreev<sup>a</sup>, Kurt Varmuza<sup>b,\*</sup>

<sup>a</sup>Center of Analytical Chemistry and Applied Spectroscopy, University of Plovdiv, BG-4000, Plovdiv, Bulgaria

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

Received 25 June 1998; received in revised form 18 December 1998; accepted 12 January 1999

## Abstract

Three types of spectral features derived from infrared peak tables were compared for their ability to be used in automatic classification of infrared spectra. Aim of classification was to provide information about presence or absence of 20 chemical substructures in organic compounds. A new method has been applied to improve spectral wavelength intervals as available from expert-knowledge. The resulting set of features proved to be better than features derived from the original intervals and better than features directly derived from peak tables. The methods used for classification were linear discriminant analysis and a back-propagation neural network; the latter gave a better performance of the developed classifiers. © 1999 Elsevier Science B.V. All rights reserved.

*Keywords:* Infrared spectroscopy; Chemometrics; Artificial neural networks; Feature selection; Substructure classification

## 1. Introduction

The methods applied for computer-assisted interpretation of infrared (IR) spectra can be classified into three groups [1]:

1. knowledge-based systems in which spectroscopic and chemical expertise is encoded to assist spectra interpretation,
2. the routine approach of searching in spectral libraries, and
3. pattern recognition techniques which have the ability to recognize structural properties by classifying spectral data.

Nowadays there is a renaissance of the last group of methods through the use of artificial neural networks (ANNs) [2,3]. Computational ANNs are known to have the capability for performing complex mappings between input and output data. They can be applied to different types of problems: classification of objects, modeling of functional relationships, storage and retrieval of information, and representation of large amounts of data [4,5]. This promises a high potential for processing IR data; recent applications cover structure elucidation from IR spectra [6–20], library search in spectral databases [21–23], and peak recognition in IR spectra [24].

The main limitations of the classification models mentioned above are related to the used spectral features. In this article we introduce a set of spectral features which are based on expert knowledge in IR

\*Corresponding author. Tel.: +43-1-58801-16060; fax: +43-1-58801-16091; e-mail: kvarmuza@email.tuwien.ac.at

spectroscopy and have been improved by a new chemometric approach for feature adaptation and selection. Classification of IR spectra by using these features is compared with classification using features which have been directly derived from peak tables. Furthermore results obtained by an ANN classification algorithm are compared with those obtained by linear discriminant analysis (LDA). As a result of the work a program module for IR spectra classification has been developed and implemented in the library search software IRSS [25].

## 2. Spectra and software

*SpecInfo IR Library.* SpecInfo [26] is a multispectral database system, running on workstations. The IR database used contains 13484 full curve spectra together with chemical structures and was available in the JCAMP-DX format. The original spectral range is 400–4000  $\text{cm}^{-1}$  with a sampling interval of 1.93  $\text{cm}^{-1}$  corresponding to 1867 data points; the absorbance values are normalized to the range 0–999. IR spectra, structural data, molecular formulas, and compound names were converted for use in the software TOSIM [27]. The IR spectral data were represented as peak tables containing positions and intensities of spectral bands; the last being the absorbance values normalized to the range 0–100. The threshold value used for peak-picking was 1% of the absorbance of the maximum peak in the spectrum; consequently the minimum peak intensity is equal to 1.0.

TOSIM is a software operating under MS-DOS [27]; it contains tools for the investigation of topological similarities in molecules, such as cluster analysis of chemical structures, and determination of maximum common substructures [28]. The implemented substructure search was used to prepare the learning and test sets for classifier development.

IRIS is a software developed for the application of IR classifiers in practical laboratory situations; it operates under MS Windows and can be started directly from the IR library search system IRSS [25]. The implemented spectral classifiers give evidence for presence or absence of chemical substructures in compounds of unknown chemical structure. IRIS and the software for classifier development (including spectra transformation and ANN classification) were written in Borland

Pascal 7.0. All computations have been performed on 80 586 computers, 200 MHz, running under MS Windows 95.

## 3. Methods

### 3.1. Spectral features

Spectral features are a set of numbers (a vector) that characterize a spectrum. Most of the recent works in IR spectra classification use features based on pre-determined fixed wavelength intervals. We introduce a different approach by choosing the wavelength intervals ( $\nu_1, \nu_2$ ) individually for each classified substructure and by defining two types of features.

Feature INT( $\nu_1, \nu_2$ ) is the intensity of a spectral band as given in Eq. (1), with  $A_{\max}$  being the maximum absorbance in this interval.

$$\text{INT}(\nu_1, \nu_2) = \begin{cases} A_{\max}/100 & \\ 0, & \text{if no peak in } (\nu_1, \nu_2) \end{cases} \quad (1)$$

This feature type is used in most knowledge-based IR spectra interpretation systems (see references in [1] and in works applying ANNs [6,8,10]). Munk et al. [6,8] use the transmittances of spectral bands instead of absorbances.

Feature L12( $\nu_1, \nu_2$ ) is calculated from the logarithmic absorbance ratio as given in Eq. (2), with  $A_{\text{sec}}$  being the absorbance of the second highest peak in the interval.

$$\text{L12}(\nu_1, \nu_2) = \begin{cases} [a - \lg(A_{\max}/A_{\text{sec}})]/a, & a = 2 \\ 0, & \text{if less than two peaks are present in } (\nu_1, \nu_2). \end{cases}$$

This feature considers that some chemical substructures give rise to two or more characteristic bands in a given spectral interval. A similar feature was successfully used for substructure classification from mass spectra [29]. The constant  $a$  in Eq. (2) scales the feature to the range 0–1. Because maximum and minimum absorbances were 100 and 1, respectively, the value of  $a$  has to be equal to 2. The maximum value for L12 is reached when the two largest peaks in the interval are equally sized ( $A_{\max} = A_{\text{sec}}$ ), Fig. 1.



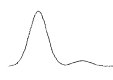
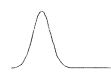

peaks					no peaks 
$L12(\nu_1, \nu_2)$	1.00	0.85	0.50	0.00	0.00

Fig. 1. Examples for feature  $L12(\nu_1, \nu_2)$ , Eq. (2), calculated from different peak absorbances in a wavelength interval.

Selection of appropriate wavelength intervals is a crucial task in feature generation. In this work three approaches have been applied. For expert-based features the intervals have been taken from literature on spectrum-structure correlations [30]. For adjusted expert-based features the interval limits have been optimized for maximum discrimination power of the features; details of this new method are described below. For comparison a third group of fixed-interval features has been generated by dividing the range  $4000\text{--}400\text{ cm}^{-1}$  into predefined 256 intervals [6] and calculating feature  $INT(\nu_1, \nu_2)$  for each of them.

### 3.2. Classifier development

The general scheme of classifier development is shown in Fig. 2. The described procedure has been performed separately for each of the investigated 20 substructures as follows.

#### 3.2.1. Learning and test set

First step is the generation of a learning and a test set for the substructure under study. Substructure searches in the database followed by a random selection of compounds results in two files, one from compounds not containing the substructure (class 1), the other from compounds containing the substructure (class 2). Typical size of the files is 500 spectra each; for some substructures, however, only a smaller number of compounds was available in the database. Half or approximately half of each class is used in the learning and in the test set, respectively (Table 1). Isotopically labeled compounds and compounds containing metal atoms have been excluded.

#### 3.2.2. Adjustment of intervals

Next step is feature generation which requires the definition of wavelength intervals ( $\nu_1, \nu_2$ ). We assume that published and widely accepted substructure-spe-

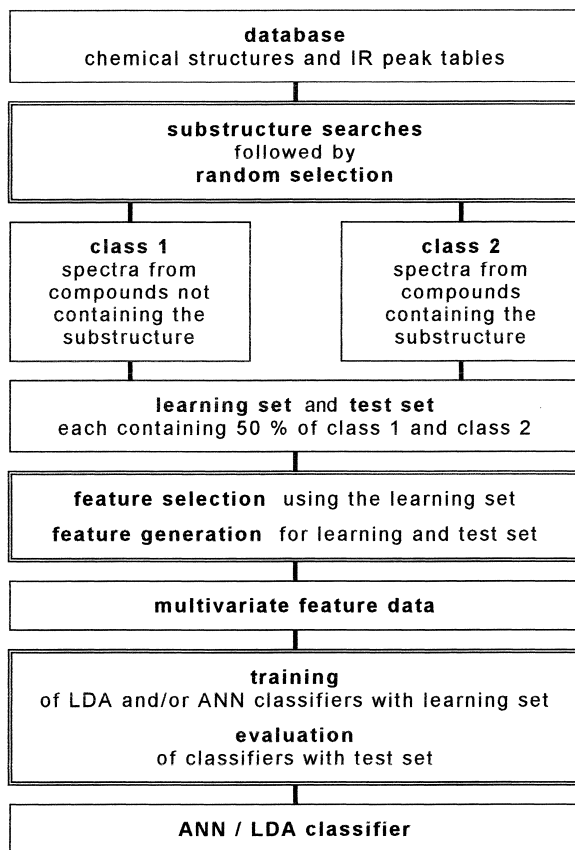


Fig. 2. Scheme of classifier development. Single-line boxes denote data, others denote data processing.

cific intervals are appropriate as initial wavelength regions for feature generation.

For expert-based features published intervals [30] have been directly used for feature generation. However, it has been shown [31] that IR correlation tables alone cannot be reliably used for automatic substructure recognition because the signal often lies outside the predicted range. Therefore for adjusted expert-based features the original intervals were changed by

Table 1  
Classified chemical substructures

No	Substructure	<i>n</i>	<i>L</i> <sub>1</sub>	<i>L</i> <sub>2</sub>	<i>T</i> <sub>1</sub>	<i>T</i> <sub>2</sub>
1	Methyl	9072	250	250	250	250
2	Methylene	11285	250	250	250	250
3	Benzene ring, any subst.	8499	250	250	250	250
4	<i>Ortho</i> substituted benzene	1496	250	250	250	250
5	<i>Meta</i> substituted benzene	657	250	250	250	250
6	<i>Para</i> substituted benzene	2704	250	250	250	250
7	Mono substituted benzene	2087	250	250	250	250
8	Isopropyl	760	250	250	250	250
9	<i>Tertiary</i> -butyl	952	250	250	250	250
10	Methoxy	2061	250	250	250	250
11	<i>cis/trans</i> C=C double bond	1345	250	250	250	250
12	Vinyl	471	235	235	235	235
13	Carboxylic acid	868	250	250	250	250
14	Aldehyde	596	240	240	240	240
15	Primary alcohol	539	215	215	215	215
16	Secondary alcohol	583	235	235	235	235
17	<i>Tertiary</i> alcohol	222	110	110	107	107
18	Phenol	561	250	250	250	250
19	Primary amine/amide	1284	250	250	250	250
20	Secondary amine/amide	1976	250	250	250	250

*n*: Number of occurrences in the SpecInfo database; *L*<sub>1</sub>: number of compounds not containing the substructure in the learning set; *L*<sub>2</sub>: number of compounds containing the substructure in the learning set; *T*<sub>1</sub>, *T*<sub>2</sub>: the corresponding numbers for the test set.

an automatic procedure with the aim to improve classification results. The two parameters to be optimized were the lower end ( $\nu_2$ ) and the width ( $\Delta\nu$ ) of the considered interval. The allowed ranges of these parameters were restricted by  $\nu_1 \leq \nu_{\text{MAX}}$ ,  $\nu_2 \geq \nu_{\text{MIN}}$ , and  $\Delta\nu_{\text{MIN}} \leq \Delta\nu \leq \Delta\nu_{\text{MAX}}$  (see the example given below).

The used optimization criterion *F* is based on the Fisher ratio [32]; it is a measure of the discriminating power of a feature (Eq. (3)).

$$F = g(A_2 - A_1)^2 / (S_2^2 + S_1^2), \quad (3)$$

where *A*<sub>1</sub> and *A*<sub>2</sub> are the arithmetic means of the feature for classes 1 and 2, respectively; *S*<sub>1</sub> and *S*<sub>2</sub> are the corresponding standard deviations; *g* is equal to  $\text{sign}(A_2 - A_1)$ . In a search for a highly discriminating and also spectroscopically relevant interval ( $\nu_1$ ,  $\nu_2$ ) it is essential to consider the sign of the Fisher ratio as is demonstrated by the following examples. Assume the optimum interval for a feature of type INT( $\nu_1$ ,  $\nu_2$ ), Eq. (1), is searched. Evidently an interval would be best if it often contains a peak in spectra from class 2 (substructure present) but does not contain a peak in spectra from class 1; this favorable situation is

reflected by a large positive value of *F*. On the other hand if spectra from class 2 would only rarely contain a peak in an interval – in comparison to spectra of class 1 – a large negative value is obtained for *F*. The same reasoning can be applied to features of type L12, Eq. (2), regarding the presence of two peaks in the interval.

Tests have shown that the response surface *F*( $\nu_1$ ,  $\nu_2$ ) usually is not smooth. Therefore an exhaustive search has been applied to find that pair of values for  $\nu_1$  and  $\nu_2$  which has the maximum value of *F* for a given training set. The following example describes a typical feature generation and optimization.

The interval 3000–2840  $\text{cm}^{-1}$  is considered to be characteristic for the methyl substructure [30]. For an optimization of this interval it has first been widened to the range  $\nu_{\text{MIN}}=2800 \text{ cm}^{-1}$  and  $\nu_{\text{MAX}}=3050 \text{ cm}^{-1}$ . Minimum and maximum interval width have been set to  $\Delta\nu_{\text{MIN}}=20 \text{ cm}^{-1}$  and  $\Delta\nu_{\text{MAX}}=250 \text{ cm}^{-1}$ ; step size for varying  $\Delta\nu$  has been set to  $1 \text{ cm}^{-1}$ . In the exhaustive search first all intervals of width  $20 \text{ cm}^{-1}$  are used for feature calculation: (2820, 2800), (2821, 2801), . . . , (3050, 3030); then all intervals of width 21, 22, . . . ,  $250 \text{ cm}^{-1}$  are tested. For each interval *F* is determined

Table 2  
Optimization of the wavelength interval ( $\nu_1, \nu_2$ ) for feature INT( $\nu_1, \nu_2$ ) when classifying a methyl group

$\nu_1$	$\nu_2$	$F$	Remarks
3000	2840	0.192	Original expert-based interval
2820	2800	0.0024	First tested interval
2821	2801	0.0012	Second tested interval
⋮			
3050	2800	0.170	Last tested interval
2996	2944	0.506	Best interval
2887	2868	0.236	Second best interval

$F$ : optimization criterion (signed Fisher ratio).

from the learning data and the intervals are ranked by their decreasing values of  $F$ . To avoid highly correlated features only intervals are put into the ordered list which do not overlap more than 10% with any of the higher ranked intervals.

In this example 26 881 intervals have to be tested, requiring 10 min computation time for a learning set with 250 spectra in each class. Table 2 contains selected results of this search; the interval with the maximum  $F$  for feature INT( $\nu_1, \nu_2$ ) was found to be  $\nu_1=2996 \text{ cm}^{-1}$ , and  $\nu_2=2944 \text{ cm}^{-1}$ . The found optimal interval limits typically vary by 1–2  $\text{cm}^{-1}$  when different learning sets are used.

### 3.2.3. Training of classifiers

The next step in classifier development is the training of an ANN or the application of LDA. All features in the learning set were scaled to zero mean and unit variance. The means and variances of the learning set were used to transform the data in the test set in the same way. For the training of an ANN the sequence of spectra in the learning set was randomized. Each chemical substructure was classified with a separate neural network.

(a) The applied ANN was a feed-forward one with one hidden layer and employing a back-propagation-of-error algorithm [3,4]. A sigmoidal squashing function was used to transfer the net input  $\text{Net}_j$  of each neuron according to Eq. (4).

$$f(\text{Net}_j) = 1 / (1 + \exp[-\alpha_j(\text{Net}_j + \theta_j)]). \quad (4)$$

The parameters  $\alpha_j$  were set to unity without loss of generality [2], and the threshold parameters  $\theta_j$  were optimized during the learning process. The number of

input neurons was equal to the number of used spectral features. The network coefficients and offsets have been initialized with random values between  $-1$  and  $+1$ . The number of hidden neurons was varied between 2 and 20 to determine the optimal value for each substructure separately. The only output neuron indicates the class membership applying the target values 0 or 1 for absence or presence of the classified substructure, respectively. Constant values of 0.6 for the learning rate, and 0.4 for the momentum factor were used.

The stop criterion used is based on the mean squared error (MSE), Eq. (5).

$$\text{MSE} = \left[ \sum (T_i - O_i)^2 \right] / N. \quad (5)$$

The sum is taken over all  $N$  objects in the learning set;  $O_i$  is the actual ANN output for spectrum  $i$ ;  $T_i$  is the corresponding target value. Tests have shown that the minimum MSE does not correspond to the best classification performance because of over-training. Therefore, the relative change of MSE has been used as the stop criterion. The training was ended when the change of MSE was less than 0.05% in three consecutive sessions. This method avoids the undesired effect that the more frequent class in the overlap region is classified correctly at the expense of the less frequent class. No difficulties with a too early end of the training (because of a flat region of MSE) has been encountered with the used data. A similar problem of ANN training has been discussed by Wilkins et al. [16].

(b) LDA classifiers have been calculated by the standard procedure as already applied for MS classifiers [29].

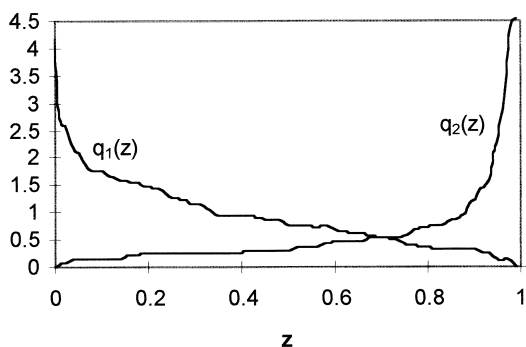


Fig. 3. Probability density distributions  $q_1(z)$  and  $q_2(z)$  of class 1 and class 2, respectively, for the discriminant variable  $z$ . Data: test set, ANN classifier for mono substituted benzene rings.

### 3.2.4. Test of classifiers

In the final step classification thresholds are determined and the performance of the classifier is evaluated. The classifier (either an ANN or a LDA classifier) is applied to the spectra of the test set resulting in a value for the discriminant variable  $z$  for each test spectrum. Fig. 3 shows typical probability density distributions  $q_1(z)$  and  $q_2(z)$  for class 1 and class 2, respectively, obtained by an ANN classifier. For LDA classifiers these distributions are often bell-shaped [29]. Many substructure classifiers, linear as well as non-linear, either for MS or IR, exhibit a considerable overlap of the classes. Therefore a simple yes/no classifier is not applicable. This problem can be partly overcome by estimating the classification performance as a function of the discriminant variable [29]. LDA and ANN classifiers have been treated by the same method.

Let  $N_{k,m}(z_m)$  be the number of objects from the test set belonging to class  $k$  and being classified to class  $m$  when thresholds  $z_m$  have been applied as follows:

IF  $z \leq z_1$  THEN assign spectrum to class 1,  
 IF  $z \geq z_2$  THEN assign spectrum to class 2,  
 ELSE reject classification.

Assuming equal a priori probabilities for both classes precisions  $P_1(z_1)$  and  $P_2(z_2)$  of classification answers can be estimated by Eqs. (6a) and (6b).

$$P_1(z_1) = 100 N_{1,1}(z_1) / [N_{1,1}(z_1) + N_{2,1}(z_1)], \quad (6a)$$

$$P_2(z_2) = 100 N_{2,2}(z_2) / [N_{2,2}(z_2) + N_{1,2}(z_2)]. \quad (6b)$$

The precisions depend on the used thresholds  $z_1$  and  $z_2$ ; for practical applications classification thresholds

are applied that yield a precision of 90% or 95%. Increasing the interval between  $z_1$  and  $z_2$  usually increases the precisions; however, at the cost of more non-classified spectra. Therefore, also the recall values of a classifier have to be considered during evaluation. Recalls  $R_1(z_1)$  and  $R_2(z_2)$  are defined as the percentage of correctly classified spectra of class 1 and class 2, respectively, at a given precision and thus depend on the thresholds  $z_1$  and  $z_2$  (Eqs. (7a) and (7b)).

$$R_1(z_1) = 100 N_{1,1}(z_1) / N_1, \quad (7a)$$

$$R_2(z_2) = 100 N_{2,2}(z_2) / N_2, \quad (7b)$$

where  $N_1$  and  $N_2$  are the number of test objects in class 1 and class 2, respectively. For an evaluation of the practical applicability of a classifier the recalls at a precision of 90% have been used in this work. For the determination of the optimum number of hidden neurons the mean of  $R_1(z_1)$  and  $R_2(z_2)$  has been considered. Application of the test set for the determination of the number of hidden neurons and the determination of the classification thresholds leads to some overestimation of the classifier performances; however, the comparison of the different feature sets may be only less affected.

## 4. Results and discussion

### 4.1. Training

Twenty chemical substructures which are infrared active were selected for this study (Table 1) and classifiers have been developed using three sets of features.

(a) For expert-based features Eqs. (1) and (2) were applied to the characteristic spectral intervals given in [30]. For the classification of substructures containing a benzene ring feature L12 has been calculated for the interval 1625–1475  $\text{cm}^{-1}$  instead of (1625,1575) and (1525,1475), [30]. The number of features in this set depends on the number of characteristic intervals and was between 4 and 13 (Table 4).

(b) The set with adjusted expert-based features has been generated as follows: The characteristic intervals given in [30] were first widened at both ends by 50  $\text{cm}^{-1}$ , then the intervals were optimized separately for each feature type as described before, and features according to Eqs. (1) and (2) were calculated. For the development of classifiers only features with  $F > 0.01$  were considered. Table 3 contains the adjusted inter-

Table 3  
Adjusted expert-based features for 20 substructures

	<i>F</i>	Type	Interval
<i>Methyl</i>			
	0.506	INT	2996–2944
	0.236	INT	2887–2868
	0.147	INT	1379–1357
	0.545	L12	3015–2927
	0.322	L12	2945–2868
	0.292	L12	1404–1357
<i>Methylene</i>			
	1.017	INT	2984–2927
	0.573	INT	2940–2839
	0.112	INT	2861–2790
	0.103	INT	1369–1364
	2.257	L12	3007–2855
	0.212	L12	1468–1367
<i>Benzene</i>			
	0.102	INT	3070–3055
	0.331	INT	1617–1588
	0.758	INT	1550–1471
	0.260	INT	870–820
	0.617	INT	838–739
	0.353	INT	716–670
	0.426	L12	3136–3023
	1.672	L12	1616–1476
	0.725	L12	766–670
	0.410	L12	883–783
<i>Ortho subst. benzene</i>			
	0.170	INT	1602–1575
	0.111	INT	1490–1483
	0.226	INT	1473–1433
	0.976	INT	764–717
	0.297	L12	1634–1530
	0.385	L12	1508–1433
	0.337	L12	797–706
	0.136	L12	710–650
<i>Meta subst. benzene</i>			
	0.179	INT	1595–1571
	0.243	INT	1504–1472
	0.099	INT	1473–1428
	0.068	INT	848–837
	0.173	INT	808–769
	0.051	INT	751–730
	0.321	INT	704–673
	0.361	L12	1606–1486
	0.258	L12	1504–1421
	0.187	L12	905–763
	0.300	L12	802–680
<i>Para subst. benzene</i>			
	0.205	INT	1613–1601
	0.578	INT	1519–1490

Table 3 (Continued)

	<i>F</i>	Type	Interval
	0.788	INT	853–821
	0.100	INT	826–765
	0.076	L12	3112–2990
	0.092	L12	1616–1564
	0.224	L12	1544–1459
	0.310	L12	852–802
<i>Mono subst. benzene</i>			
	0.432	INT	3070–3055
	0.283	INT	3050–3022
	0.093	INT	1605–1594
	0.450	INT	1500–1488
	0.247	INT	1459–1443
	0.829	INT	770–699
	2.514	INT	705–683
	1.241	L12	3073–3024
	0.269	L12	1609–1576
	0.527	L12	1507–1445
	0.644	L12	770–687
<i>Isopropyl</i>			
	0.633	INT	2989–2954
	0.411	INT	2885–2864
	0.133	INT	1396–1383
	0.154	INT	1379–1366
	0.055	INT	1222–1140
	0.441	L12	2985–2864
	0.584	L12	1399–1358
	0.065	L12	1053–1023
<i>Tertiary butyl</i>			
	1.414	INT	2988–2935
	0.304	INT	2950–2869
	0.842	INT	2876–2859
	0.215	INT	1483–1476
	0.183	INT	1396–1392
	1.095	INT	1371–1361
	0.405	L12	2981–2858
	0.248	L12	1484–1457
	1.331	L12	1405–1360
	0.068	L12	1250–1196
<i>Methoxy</i>			
	0.125	INT	3019–2990
	0.138	INT	2938–2916
	0.239	INT	2893–2850
	0.211	INT	2841–2831
	0.062	INT	1466–1448
	0.210	INT	1445–1432
	0.300	INT	1288–1036
	0.389	L12	3019–2940
	0.189	L12	2938–2850
	0.170	L12	1466–1432
	0.195	L12	1219–1110

Table 3 (Continued)

	<i>F</i>	Type	Interval	
<i>-CH=CH-</i>	0.041	INT	3027–3021	
	0.023	INT	3051–3048	
	0.023	INT	3043–3042	
	0.089	INT	1697–1645	
	0.270	INT	985–962	
	0.072	INT	746–734	
	0.054	L12	3158–3017	
	0.085	L12	1700–1604	
	0.072	L12	997–959	
	0.055	L12	815–754	
<i>Vinyl</i>	0.071	INT	3089–3071	
	0.057	INT	3023–3008	
	0.121	INT	1655–1631	
	0.574	INT	1004–981	
	0.293	INT	952–906	
	0.037	L12	3089–3008	
	0.220	L12	1661–1600	
	0.219	L12	1036–989	
	0.098	L12	1000–879	
	<i>Carboxylic acid</i>	0.108	INT	3405–2992
0.310		INT	3100–2990	
0.890		INT	2667–2513	
0.913		INT	1733–1676	
0.188		INT	1320–1248	
0.183		INT	1436–1401	
0.062		INT	953–887	
1.001		L12	2741–2503	
<i>Aldehyde</i>		0.376	INT	2733–2699
		0.934	INT	1732–1661
	0.626	L12	2892–2689	
	0.099	L12	1696–1620	
<i>Primary alcohol</i>	0.824	INT	3578–3278	
	0.587	INT	1079–1003	
<i>Secondary alcohol</i>	0.828	INT	3629–3210	
	0.426	INT	1096–1028	
	0.222	INT	1032–981	
	0.198	L12	1092–1014	
	0.103	L12	1219–1109	
<i>Tertiary alcohol</i>	0.875	INT	3650–3201	
	0.079	INT	1416–1406	
	0.211	INT	1385–1374	

Table 3 (Continued)

	<i>F</i>	Type	Interval
	0.084	INT	1169–1124
	0.060	INT	1087–1084
	0.056	INT	1052–1045
	0.054	L12	1142–1113
	0.117	L12	1061–1019
	0.068	L12	1022–980
<i>Phenol</i>	0.347	INT	3603–3205
	0.492	INT	1318–1200
	0.468	INT	1264–1130
	0.086	L12	1448–1207
	0.072	L12	1273–1072
<i>Primary amine/amide</i>	0.768	INT	3502–3326
	0.322	INT	1666–1611
	1.088	L12	3509–3272
<i>Secondary amine/amide</i>	0.242	INT	3332–3263
	0.238	INT	1674–1584
	0.387	INT	1587–1531
	0.087	INT	1327–1291
	0.069	L12	1253–1219
	0.116	L12	788–753
	0.105	L12	726–697

*F*: signed Fisher ratio (Eq. (3)); the feature type is defined in Eqs. (1) and (2); wavelength intervals are given in  $\text{cm}^{-1}$ .

vals for all 20 substructures together with the Fisher ratios of the generated features. The number of features in this set depends on the number of characteristic intervals and was between 2 and 11 (Table 4).

(c) For the generation of fixed-interval features the spectral range 400–4000  $\text{cm}^{-1}$  was divided into 256 intervals with the widths continuously increasing with growing wave number [6]. It has been reported [6,8] that classifiers developed with features from such intervals are better than those based on features from equally sized intervals. Features have been calculated by applying Eq. (1) to each interval; the 20 features with highest absolute values of the Fisher ratio were selected.

For a comparison of the different feature types LDA classifiers were developed for all three feature sets. ANN classifiers were only developed for adjusted expert-based features because of the high computational effort necessary.



Table 4  
 Characteristics of classifiers for the recognition of 20 substructures using different types of features and classification methods

Substructure	Classification type	$nF/nH$	$R_1$	$R_2$	$A_{50}$
Methyl	LDA/interval	20	25.6 <sup>a</sup>	31.2	83.1
	LDA/expert	6	45.2 <sup>a</sup>	7.6	66.3
	LDA/adjusted	6	56.4 <sup>a</sup>	34.0	85.7
	ANN/adjusted	6/16	68.4 <sup>a</sup>	41.2	86.6
Methylene	LDA/interval	20	3.6	44.0	86.6
	LDA/expert	6	43.2	52.8	89.9
	LDA/adjusted	6	63.6	39.6	88.2
	ANN/adjusted	6/5	64.8	49.6	89.4
Benzene	LDA/interval	20	52.8	46.8	89.4
	LDA/expert	7	64.8	14.0	81.7
	LDA/adjusted	10	73.6	42.8	88.5
	ANN/adjusted	10/6	70.8	50.0	89.9
<i>Ortho</i> substituted benzene	LDA/interval	20	41.2	16.4	80.7
	LDA/expert	7	35.6	4.4	80.1
	LDA/adjusted	8	51.6	8.8	80.8
	ANN/adjusted	8/6	53.6	37.2	83.3
<i>Meta</i> substituted benzene	LDA/interval	20	46.0	55.2 <sup>a</sup>	80.5
	LDA/expert	13	27.6	5.2 <sup>a</sup>	71.8
	LDA/adjusted	11	36.0	3.2 <sup>a</sup>	72.3
	ANN/adjusted	11/4	55.6	64.0	79.6
<i>Para</i> substituted benzene	LDA/interval	20	47.2	49.6	89.4
	LDA/expert	7	42.4	28.0	82.8
	LDA/adjusted	8	61.2	55.6	91.2
	ANN/adjusted	8/6	65.6	40.8	86.8
Mono substituted benzene	LDA/interval	20	79.2	71.6	94.6
	LDA/expert	11	69.6	77.6	96.9
	LDA/adjusted	11	86.0	86.8	99.2
	ANN/adjusted	11/6	92.8	92.4	97.8
Isopropyl	LDA/interval	20	18.0	57.2 <sup>a</sup>	80.7
	LDA/expert	8	51.2	59.6 <sup>a</sup>	85.7
	LDA/adjusted	8	62.8	77.2 <sup>a</sup>	86.2
	ANN/adjusted	8/4	60.4	80.4 <sup>a</sup>	88.7
<i>Tertiary</i> butyl	LDA/interval	20	42.4	40.8	87.5
	LDA/expert	8	56.0	19.2	80.7
	LDA/adjusted	10	76.8	66.0	93.8
	ANN/adjusted	10/6	78.8	75.2	97.2
Methoxy	LDA/interval	20	4.0	37.6	85.5
	LDA/expert	8	13.6	<2.8	78.1
	LDA/adjusted	11	28.8	40.8	87.7
	ANN/adjusted	11/4	28.8	57.6	93.3
<i>Cis-</i> or <i>trans-</i> substituted carbon-carbon double bond	LDA/interval	20	8.0 <sup>b</sup>	4.4	66.5
	LDA/expert	8	- <sup>c</sup>	5.2	67.5
	LDA/adjusted	10	39.6 <sup>b</sup>	9.2	67.2
	ANN/adjusted	10/4	7.6 <sup>b</sup>	20.8	62.5

Table 4 (Continued)

Substructure	Classification type	$nF/nH$	$R_1$	$R_2$	$A_{50}$
Vinyl	LDA/interval	20	28.1	14.9	83.6
	LDA/expert	10	21.3	14.9	80.7
	LDA/adjusted	9	34.5	26.8	76.7
	ANN/adjusted	9/2	55.8	43.0	87.3
Carboxylic acid	LDA/interval	20	18.4	71.6	93.7
	LDA/expert	6	29.2	10.8	74.8
	LDA/adjusted	8	64.0	77.2	97.7
	ANN/adjusted	8/2	68.0	85.2	>98.5
Aldehyde	LDA/interval	20	52.5 <sup>a</sup>	68.8 <sup>a</sup>	92.3
	LDA/expert	6	62.1 <sup>a</sup>	55.5 <sup>a</sup>	84.1
	LDA/adjusted	4	72.5 <sup>a</sup>	76.3 <sup>a</sup>	91.1
	ANN/adjusted	4/4	79.2 <sup>a</sup>	78.8 <sup>a</sup>	94.5
Primary alcohol	LDA/interval	20	62.8	54.4	93.2
	LDA/expert	4	12.1	37.7	77.7
	LDA/adjusted	2	61.9	56.3	92.3
	ANN/adjusted	2/2	82.3	84.2	97.6
Secondary alcohol	LDA/interval	20	48.5	40.9	86.8
	LDA/expert	4	4.3	32.8	86.2
	LDA/adjusted	5	22.1	48.1	89.1
	ANN/adjusted	5/20	73.2	60.0	92.3
Tertiary alcohol	LDA/interval	20	27.1	24.3	82.3
	LDA/expert	4	73.8	25.2	73.8
	LDA/adjusted	9	71.1	24.3	85.6
	ANN/adjusted	9/2	73.8	31.7	85.6
Phenol	LDA/interval	20	14.1	10.2	71.7
	LDA/expert	4	29.3	25.3	74.4
	LDA/adjusted	5	37.8	24.0	73.9
	ANN/adjusted	5/2	44.9	24.0	76.3
Primary amine/amide	LDA/interval	20	23.6	34.0	86.4
	LDA/expert	6	30.8	38.0	87.0
	LDA/adjusted	3	36.8	55.6	94.0
	ANN/adjusted	3/6	38.4	68.8	95.4
Secondary amine/amide	LDA/interval	20	4.6	20.8	78.1
	LDA/expert	6	<2.4	<1.6	69.0
	LDA/adjusted	7	18.0	<4.4	78.8
	ANN/adjusted	7/20	2.0	30.8	81.2

LDA, linear discriminant analysis; ANN, artificial neural network; interval, fixed-interval features; expert, expert-based features; adjusted, adjusted expert-based features;  $nF$ , number of features;  $nH$ , number of hidden neurons;  $R_1$ ,  $R_2$ , recall at 90% precision for class 1 and 2, respectively;  $A_{50}$ , precision for class 2 at a recall of 50%.

<sup>a</sup> Recall at precision 80%.

<sup>b</sup> Recall at precision 70%.

<sup>c</sup> Could not be determined at precision 70% or above.

#### 4.2. Feature reliability

The majority of the fixed-interval features have negative Fisher ratios  $F$ . For example, in the classification of methyl groups only 68 out of the 256 fixed-interval features have positive values for  $F$ . A negative Fisher ratio means that class 1 compounds (substructure absent) have on the average larger peaks in the interval than class 2 compounds (substructure present). Use of such a feature obviously contradicts basic principles of IR spectroscopy.

One reason for large negative Fisher ratios is an unbalanced distribution of an accompanying substructure in the two classes. The learning set for a methyl classifier is used as an example to demonstrate this effect. In class 1 (consisting of 250 randomly selected compounds without a methyl group) 189 structures contain a benzene ring; in class 2 (consisting of 250 randomly selected compounds with a methyl group) only 146 contain a benzene ring. The numbers of *ortho*-, *meta*-, *para*-, and mono-substituted benzene rings in class 1 are 40, 14, 50, and 55; in class 2 they are 26, 8, 43 and 30. It is therefore not surprising that an automatic feature selection will also result in features that are responsible for the benzene substructure but not for the methyl group. Actually the 14 fixed-interval features with the largest negative Fisher ratios (from the 20 features used for classification) are from the intervals 3155–3136 (14), 3117–3099 (11), 3099–3080 (13), 3080–3062 (7), 3062–3043 (18), 1547–1534 (20), 1508–1495 (6), 890–880 (12), 841–831 (15), 756–747 (8), 693–685 (4), 685–676 (17), 659–650 (19), and 537–529 (16); the number in brackets is the rank when all 256 features are ordered by their decreasing absolute value of the Fisher ratio. Only the interval 1547–1534  $\text{cm}^{-1}$  (rank 20) is not characteristic for a benzene ring. A methyl classifier based on this simple method of feature selection therefore would contain many features irrelevant to the methyl group.

To reduce such spectra-structure miscorrelations the sign of the Fisher ratio has been considered in this work for feature selection. Other strategies restrict the selected features to those with positive loadings (in LDA classifiers) or positive coefficients (in ANN classifiers) [6].

The problem of miscorrelations is even more severe if compounds from class 2 contain an additional

substructure more often than compounds from class 1. In this case neither the use of the signed Fisher ratio nor the restriction to positive LDA loadings can avoid the selection of irrelevant features.

The strategy to reduce the chance of miscorrelations, which has been applied in this work, utilizes spectroscopic knowledge to define wavelength intervals that are characteristic for the classified substructure, followed by an automatic optimization procedure. Results demonstrate that a proper adjustment of the expert-based interval limits improves the performance of the classifier significantly. Because of the immense variety of chemical structures structurally well balanced learning and test sets are very difficult to obtain and therefore miscorrelations never can be excluded completely. Considering additional spectroscopic expertise helps in some cases: for instance substructures with a benzene ring usually give rise to only one peak per characteristic interval; therefore features of type L12 (Eq. (2)) are not relevant and can be excluded. Applying these methods resulted in adjusted expert-based features with Fisher ratios being all positive and being significantly higher than those for other feature types. The LDA loadings calculated from these features were mostly positive.

#### 4.3. Classification efficiency

The obtained binary classifiers have been evaluated by the recalls  $R_1$  and  $R_2$  for class 1 and class 2, respectively, at a precision of 90%. Results for all 20 substructures are summarized in Table 4. When it was impossible to determine the recall at 90% precision – because this value was not reached – the recalls are given for a precision of 80% or even for only 70%. To allow a comparison with other works also the measure  $A_{50}$  for the predictive ability [6,8] is given; this criterion is equivalent to the precision for class 2 at a recall of 50%.

Most of the investigated substructures have acceptable high values for the recall and also the criterion  $A_{50}$ ; this demonstrates the capabilities of the applied methods for substructure recognition from IR spectra. Exceptions are the classifiers for *cis*- or *trans*-substituted carbon–carbon double bonds and for secondary amines or amides, which exhibit only a poor classification ability. Corresponding to experiences from IR spectroscopy the classification of compounds from

class 1 (substructure absent) was found to be in general more accurate than those from class 2; exceptions are the substructures methyl, methoxy, mono-substituted benzene ring, carboxylic acids and primary amines.

Klawun and Wilkins [18] presented a comprehensive comparison of results from IR substructure classification. The 12 substance classes common in [18] and in our work show similar trends in the prediction rates. For instance the success of classification increases from tertiary to secondary and to primary alcohols, and from secondary to primary amines (or amides); carbon–carbon double bonds and aldehydes seem to be difficult to recognize, while benzene rings, carboxylic acids, and primary alcohols are easier to identify.

For an objective comparison of the three different feature sets (and the two applied classification methods) the paired *t*-test has been applied. Each set of classifiers has been compared with the three others by using the differences of  $R_1$ ,  $R_2$  and  $A_{50}$  for all 20 substructures. The tabulated *t*-value for 19 degrees of freedom and 90% statistical significance is 1.73 for a two-sided test. Table 5 summarizes the results of this comparison. A *t*-value greater than 1.73 indicates that the method given in the row is significant better than the method given in the column; if *t* is smaller than  $-1.73$  the method in the column is significant better than that in the row.

Table 5  
Signed *t*-values (statistical *t*-test) from comparisons of LDA and ANN classifiers using different types of features

	LDA/interval	LDA/expert	LDA/adjusted
<i>Comparison of recall <math>R_1</math> (class 1, substructure absent)</i>			
LDA/expert	0.62		
LDA/adjusted	4.44	6.29	
ANN/adjusted	6.91	5.27	1.53
<i>Comparison of recall <math>R_2</math> (class 2, substructure present)</i>			
LDA/expert	-3.10		
LDA/adjusted	0.85	4.04	
ANN/adjusted	6.55	6.54	3.85
<i>Comparison of criterion <math>A_{50}</math></i>			
LDA/expert	-3.35		
LDA/adjusted	1.58	4.02	
ANN/adjusted	4.48	5.16	2.78

For spectra from class 2 (substructure present) conclusions can be drawn from Table 5 as follows: The criteria  $R_2$  and  $A_{50}$  give nearly the same results in the comparison of the feature sets; this is reasonable because both criteria estimate the classifier performance for class 2 spectra. For LDA classifiers adjusted expert-based features gave similar results as fixed-interval features but both are significantly better than the original expert-based features. At a first glance it is surprising that expert-based features are less powerful than fixed-interval features. An explanation for this result is that for fixed-interval features the most discriminating intervals (without regard to spectroscopic relevance) have been selected. On the other hand the original expert-based features use wide intervals to include spectral bands of the substructure in various structural environments; thus, into these intervals also fall bands of compounds which do not belong to class 2. The adjustment of the original expert-based intervals by the optimization method described enhanced the classification performance significantly.

For spectra from class 1 (substructure absent) conclusions can be drawn as follows: for LDA classifiers the adjusted expert-based features gave much better results than the original expert-based features which are almost equivalent to the fixed-interval features. For the classification of class 1 spectra even the wide original expert-based intervals are successful: if a spectrum has no peaks in the wide interval then the presence of the substructure is very improbable. However, also for class 1 the adjustment of the original expert-based intervals enhanced the classification performance considerably.

#### 4.4. Comparison of ANN with LDA

The paired *t*-test has also been used to compare the performance of ANN and LDA classifiers. Because of the high computational effort necessary to train ANNs only the set of adjusted expert-based features has been used for ANN classifiers. The *t*-values in Table 5 show that classification of class 2 is in general significantly better with ANN than with LDA classifiers. For class 1 the differences are not significant although for 16 substructures the ANN classifiers exhibit higher recall values  $R_1$  than the LDA classifiers (Table 4). The only substructures for which ANN could not improve the

(very poor) performance of LDA classifiers are *cis*- or *trans*-substituted carbon–carbon double bonds and secondary amines or amides.

The better performance of ANN classifiers is graphically demonstrated by an example in which only two features are used. For the recognition of primary alcohols the intervals 3578–3278  $\text{cm}^{-1}$  and 1079–1003  $\text{cm}^{-1}$  have been found to be optimal for the adjusted expert-based features of type INT, (Eq. (1)). In a LDA classifier the discriminant variable  $z_{\text{LDA}}$  is a linear function of INT(3578, 3278) and INT(1079, 1003), and was found as given by Eq. (8).

$$z_{\text{LDA}} = 0.8287 I_1 + 0.5598 I_2, \quad (8)$$

where  $I_1$  and  $I_2$  are the features autoscaled by the means and standard deviations from the learning set data, Eqs. (9a) and (9b).

$$I_1 = (\text{INT}(3578, 3278) - 0.3214)/0.306, \quad (9a)$$

$$I_2 = (\text{INT}(1079, 1003) - 0.4463)/0.286. \quad (9b)$$

For the ANN classifier the discriminant variable  $z_{\text{ANN}}$  is calculated from the same features by using the non-linear algorithm implemented in the network, Eq. (10)

$$z_{\text{ANN}} = f [4.672f(16.50 I_1 - 0.9246 I_2 + 9.864) - 3.498f(3.455 I_1 - 8.991 I_2 - 6.947) - 2.416] \quad (10)$$

with  $f(\cdot)$  being the squashing function from Eq. (4) with  $\alpha_j=1$  and  $\theta_j=0$ .

Fig. 4 contains the 90% precision borders for assigning spectra to class 1 or class 2; the axes of the co-ordinate system correspond to the absorbance values of the highest peaks in the two intervals. Each point corresponds to one or more spectra from the test set (triangles denote primary alcohols, circles denote other compounds). The straight lines describe the classification thresholds for the LDA classifier, while the curved lines are for the ANN classifier. If a spectrum is located in the area between the two classes then the precision of the answer is below 90% and the classification is rejected. For example a spectrum that gives the value 0.4 for both features would be assigned by the ANN classifier to class 2 (primary alcohol) but would not result in an answer when the LDA classifier is applied. The non-linear ANN classifier is able to separate the two classes better – although not excellent – than the LDA classifier and consequently has higher

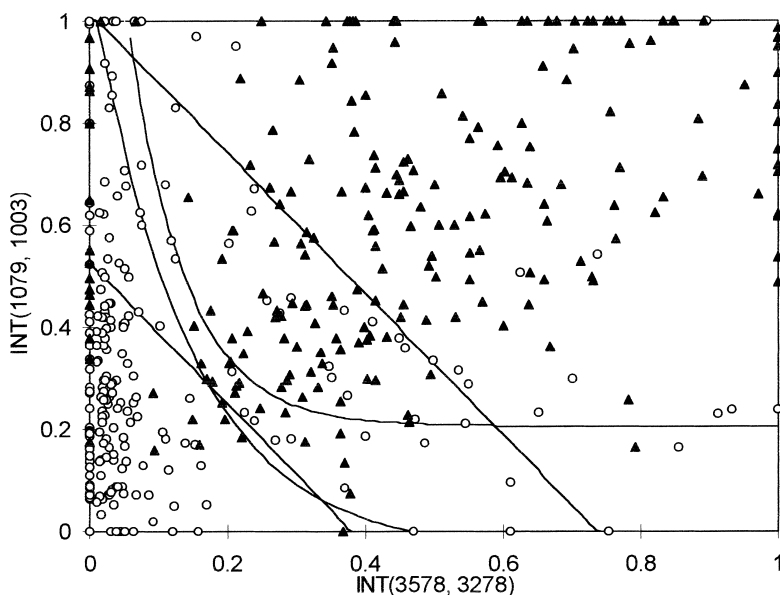


Fig. 4. Recognition of primary alcohols (test set) by using features INT(3578, 3278) and INT(1079, 1003) for adjusted expert-based intervals. The feature space has been separated by the training into areas for class 1, class 2 and rejection of classification (90% precision). Straight lines, LDA; curved lines, ANN; triangles, primary alcohols (class 2); circles, other compounds (class 1).

recall values (82% instead of 62% for  $R_1$ , and 84% instead of 56% for  $R_2$ ).

## 5. Conclusions

The introduced adjusted expert-based features for IR spectra are based on reliable concepts from spectroscopy combined with mathematical optimization. Results obtained for 20 substructures show that this feature type leads to significantly better classifiers than obtained from fixed-interval features or features calculated from the original expert-based wavelength intervals. Introducing human expertise to the learning process and thereby also setting some limits had a positive effect on the performance of the classifiers. The signed Fisher ratio was successfully used for feature selection and reduced spectroscopic misclassifications.

ANN classifiers were found to be in general better than LDA classifiers; for recall values of class 2 (substructure present) the improvement is highly significant. This result is of benefit because spectroscopists are primarily interested on the substructures present in an unknown. Also in systematic structure elucidation – which creates all isomers for a given molecular formula that fulfil defined structural restrictions – positive restrictions are most powerful [29,33,34]. The developed IR classifiers were implemented into the new software IRIS for easy use in practical laboratory situations; thus IR spectra classification serves as a complementary tool to IR library search. Tests have shown that IR classifiers provide additional structural information to that obtained from MS classifiers [34]. The independent application of MS and IR classifiers for establishing structural restrictions often reduces appreciably the number of generated isomeric molecular structures in comparison to the use of only one type of classifiers [35].

## Acknowledgements

We thank R. Neudert from Chemical Concepts (Weinheim, Germany) for providing the SpecInfo IR database. We are grateful to the late J.T. Clerc and to E. Pretsch (ETH Zurich, Switzerland) for making this database available in an appropriate for-

mat. We appreciate constructive remarks from two unknown reviewers.

## References

- [1] H.J. Luinge, *Vib. Spectrosc.* 1 (1990) 3.
- [2] J. Zupan, J. Gasteiger, *Anal. Chim. Acta* 248 (1991) 1.
- [3] J. Zupan, J. Gasteiger, *Neural Networks for Chemists: An Introduction*, VCH Publishers, Weinheim, 1993.
- [4] D.E. Rumelhart, J.L. McClelland, *Parallel Distributed Processing: Explorations in the Microstructures of Cognition*, vols. I and II, MIT Press, Cambridge, MA, 1986.
- [5] T. Kohonen, *Self-organizing Maps*, Springer, Berlin, 1995.
- [6] E.W. Robb, M.E. Munk, *Mikrochim. Acta [Wien]* I (1990) 131.
- [7] R.J. Fessenden, L. Gyorgyi, *J. Chem. Soc., Perkin Trans. II* (1991) 1755.
- [8] M.E. Munk, M.S. Madison, E.W. Robb, *Mikrochim. Acta [Wien]* II (1991) 505.
- [9] M. Meyer, T. Weigelt, *Anal. Chim. Acta* 265 (1992) 183.
- [10] U.M. Weigel, R. Herges, *J. Chem. Inf. Comput. Sci.* 32 (1992) 723.
- [11] J.R.M. Smits, P. Schoenmakers, A. Stehmann, F. Sijstermans, G. Kateman, *Chemom. Intell. Lab. Syst.* 18 (1993) 27.
- [12] D. Ricard, C. Cachet, D. Cabrol-Bass, T.P. Forrest, *J. Chem. Inf. Comput. Sci.* 33 (1993) 202.
- [13] Q.C. Van Est, P.J. Schoenmakers, J.R.M. Smits, W.P.M. Nijssen, *Vib. Spectrosc.* 4 (1993) 263.
- [14] J. Gasteiger, X. Li, V. Simon, M. Novic, J. Zupan, *J. Mol. Struct.* 292 (1993) 141.
- [15] T. Visser, H.J. Luinge, J.H. van der Maas, *Anal. Chim. Acta* 296 (1994) 141.
- [16] C. Klawun, C.L. Wilkins, *J. Chem. Inf. Comput. Sci.* 34 (1994) 984.
- [17] M. Novic, J. Zupan, *J. Chem. Inf. Comput. Sci.* 35 (1995) 454.
- [18] C. Klawun, C.L. Wilkins, *J. Chem. Inf. Comput. Sci.* 36 (1996) 69.
- [19] M.E. Munk, M.S. Madison, E.W. Robb, *J. Chem. Inf. Comput. Sci.* 36 (1996) 231.
- [20] C. Klawun, C.L. Wilkins, *J. Chem. Inf. Comput. Sci.* 36 (1996) 249.
- [21] K. Tanabe, T. Tamura, H. Uesaka, *Appl. Spectrosc.* 46 (1992) 807.
- [22] A. Bruchmann, H.J. Gotze, P. Zinn, *Chemom. Intell. Lab. Syst.* 18 (1993) 59.
- [23] C. Klawun, C.L. Wilkins, *Anal. Chem.* 67 (1995) 374.
- [24] B.J. Wythoff, S.P. Levine, S.A. Tomellini, *Anal. Chem.* 62 (1990) 2702.
- [25] P.N. Penchev, N.T. Kochev, G.N. Andreev, *Compt. Rend. Bulg. Sci.* 51(1) (1998). Software IRSS is available from one of the authors (PNP).
- [26] The spectroscopic database SpecInfo is available from Chemical Concepts, PO Box 100202, D-69442 Weinheim, Germany.

- [27] H. Scsibrany, K. Varmuza, in: C. Jochum (Ed.), *Software Development in Chemistry*, vol. 8, Gesellschaft Deutscher Chemiker, Frankfurt am Main, 1994, p. 235, Software TOSIM is available from one of the authors (KV).
- [28] K. Varmuza, P.N. Penchev, H. Scsibrany, *J. Chem. Inf. Comput. Sci.* 38 (1998) 420.
- [29] K. Varmuza, W. Werther, *J. Chem. Inf. Comput. Sci.* 36 (1996) 323.
- [30] E. Pretsch, T.J. Clerc, J. Seibl, W. Simon, *Tables of Spectral Data for Structure Determination of Organic Compounds*, Springer, Berlin, 1989.
- [31] C. Affolter, K. Baumann, J.T. Clerc, H. Schriber, E. Pretsch, *Mikrochim. Acta (Suppl.)* 14 (1997) 143.
- [32] K. Varmuza, *Pattern Recognition in Chemistry*, Springer, Berlin, 1980.
- [33] K. Varmuza, P.N. Penchev, F. Stancl, W. Werther, *J. Mol. Struct.* 408/409 (1997) 91.
- [34] K. Varmuza, W. Werther, *Adv. Mass Spectrometry* 14 (1998) 611.
- [35] K. Varmuza, P.N. Penchev, (1998), unpublished results.