

# Large and frequently occurring substructures in organic compounds obtained by library search of infrared spectra

K. Varmuza<sup>a,\*</sup>, P.N. Penchev<sup>b</sup>, H. Scsibrany<sup>a</sup>

<sup>a</sup> *Laboratory for Chemometrics, Institute of General Chemistry, Vienna University of Technology, Getreidemarkt 9 / 152, A-1060 Vienna, Austria*

<sup>b</sup> *Center of Analytical Chemistry and Applied Spectroscopy, Faculty of Chemistry, University of Plovdiv, 24 Tsar Assen Street, BG-4000 Plovdiv, Bulgaria*

Received 20 July 1998; received in revised form 13 October 1998; accepted 15 October 1998

## Abstract

Comparing the infrared spectrum of a compound whose chemical structure is unknown with the spectra of a library is a routinely used method to obtain information about the unknown structure. The resulting hitlist contains compounds exhibiting the most similar spectra. If the unknown is not contained in the library, a method based on the maximum common substructure concept can be applied to extract common structural features from the hitlist structures. The result is a set of substructures that are characteristic for the query structure. This approach has been applied to infrared spectra from a series of model compounds and has been compared with information obtained from mass spectra by the same procedure. A complementary chemometric method for evaluating spectral hitlists is principal component analysis of spectral and structural data. © 1999 Elsevier Science B.V. All rights reserved.

*Keywords:* Computer-assisted structure elucidation; Computer chemistry; Chemometrics

## 1. Introduction

Spectral library search is still the most widely used technique for computer-assisted identification of organic compounds [1]. The primary result is a hitlist containing typically 10 to 100 reference spectra (the *hits*) which are most similar to the spectrum of the unknown. If the unknown is present in the library then the correct answer often appears among the first hits and can be identified easily by consider-

ing additional restrictions such as volatility or origin of the investigated sample. However, if the unknown is not contained in the spectral library a more detailed evaluation of the hitlist structures and spectra is necessary. This data interpretation is usually done by the spectroscopist and by relying on the hypothesis (and experience) that similar spectra often indicate similar chemical structures [2]. For infrared spectra (IR), a method based on the concept of maximum common substructures (MCS) has recently been presented as an aid to the spectroscopist [3]. A similar approach has been described for mass spectra [4], <sup>13</sup>C-NMR spectra [5] and considered for IR in combination with the application of fuzzy logic [6].

\* Corresponding author. Tel.: +43-1-58801-16060; Fax: +43-1-581-1915; E-mail: kvarmuza@email.tuwien.ac.at

In this paper we briefly describe the method, present new examples, compare the evaluation of hitlists from IR with those obtained from mass spectra (MS), and finally present preliminary results with a complementary chemometric data evaluation method. A spectral and structural library containing more than 13 000 entries served as the database [7] for this work.

## 2. Method

Details of the MCS-based method for an evaluation of IR hitlists and the influence of several parameters have been described elsewhere [3]; therefore, only a practical-oriented summary is given here.

(1) The used measure for *similarity of two IR spectra* is based on the correlation coefficient of absorbances and is equal to hit quality index  $HQI_4$  used in Ref. [3]. Let  $r_k$  and  $u_k$  be the mean-centered absorbances in wavenumber interval  $k$  of the reference spectrum and the spectrum of the unknown, respectively, with  $\sum r_k = 0$  and  $\sum u_k = 0$  (sum over all intervals in the spectrum). Spectral similarity  $S$  is defined by

$$S = 999 \left[ \frac{\sum r_k u_k}{\text{SQR}(\sum r_k^2 \sum u_k^2)} + 1 \right] / 2$$

and ranges between 0 and 999 (the last value is obtained for identical spectra); the number of equally sized intervals was 801 (interval width  $4 \text{ cm}^{-1}$ , range 500 to  $3700 \text{ cm}^{-1}$ ); SQR denotes the square root. Library search has been performed by software IRSS [8], running under MS Windows. The database consisted of 13 484 IR spectra and the corresponding chemical structures; it is part of the SpecInfo database system [7]. Hitlists consisted of 50 reference spectra and structures; structure duplicates and the query compound have been excluded from the following evaluation.

(2) For *library search of MS* the database system MassLib has been used<sup>1</sup> with a library [9] contain-

ing more than 130 000 entries (including compound duplicates). The spectral similarity implemented in this system [10] is a heuristic function of five simple similarity criteria (for instance the number of common peaks or a measure for pattern correlation); however, full details have not been published. MassLib has been running on a Vax workstation; hitlists consisted of 50 reference spectra and structures; structure duplicates and the query compound have been excluded from the following evaluation.

(3) The *maximum common substructure (MCS)* of two chemical structures is defined here as a connected substructure of maximum size which is common to both chemical structures. The MCS characterizes common structural properties. The size of the MCS is measured by the number of non-hydrogen atoms; 'common' means that atoms and bond types (single, double, triple, aromatic) have to be equal.

If more than two structures are given the MCS of all  $n$  structures is usually not a good solution to characterize common structural properties because it may be limited by even a single exotic outlier structure [11]. Therefore, a procedure as follows has been applied to generate a set of informative substructures.

(a) Determine the MCS for each of the  $n(n-1)/2$  pairs of hitlist structures.

(b) Determine for each MCS  $i$  in how many ( $n_i$ ) hitlist structures it is contained.

(c) Rank the MCSs by their frequencies  $n_i$ .

(d) Use the MCSs with highest frequencies as a set of substructures which may be characteristic for the unknown.

(e) Optionally, a compressed set of characteristic substructures can be generated by deleting substructures that are contained in others.

MCS determination and substructure searches have been performed by software ToSiM,<sup>2</sup> running under MS Windows.

(4) *Principal component analyses (PCA) of IR spectra and of chemical structures* has been investigated as an additional tool to obtain insight into the data contained in a hitlist; a similar approach has been applied for data interpretation in MS [12], IR

<sup>1</sup> MassLib: Mass Spectra Database and Information System; developed by D. Henneberg, B. Weimann, E. Ziegler (Max-Planck-Institut für Kohlenforschung, Mülheim/Ruhr, Germany). Available from: MSP Friedli, Bindenhausstrasse 46, CH-3098 Koeniz, Switzerland.

<sup>2</sup> ToSiM: Software for Investigation of Topological Similarities in Molecules. Available from author K. Varmuza.











































id	query structure	IR MS	n	characteristic substructures with highest frequencies					av. frequ.
				1	2	3	4	5	
I	 <chem>C5H8O2</chem>	IR	50	46 	44 	42 	42 	40 	86 %
			MS	36	26 	18 	16 <i>wrong</i>	13 	11 
II	 <chem>C6H10O2</chem>	IR	50	47 	41 	40 	36 	35 	80 %
			MS	37	27 	16 	14 <i>wrong</i>	14 	14 
III	 <chem>C8H14O2</chem>	IR	50	43 	43 	42 	41 	40 	84 %
			MS	37	27 	23 	19 	16 	16 
IV	 <chem>C9H16O2</chem>	IR	50	50 	46 	44 	43 	42 	90 %
			MS	36	27 	23 	19 	16 	16 

Fig. 1. Five most frequent characteristic substructures obtained from the IR- and MS-hitlists. For each substructure, the number of occurrences in the hitlist is given; *n*, number of structures in hitlist; 'av. frequ.', averaged frequency of the five substructures in the hitlist (%); 'wrong', substructure not contained in query structure.

[13], and for prediction of IR spectra [14]. Software SCAN,<sup>3</sup> running under MS Windows, has been used for PCA and other chemometric methods [15]. Features (variables) that characterize IR spectra and chemical structures have been generated as follows:

IR spectra have been characterized by a set of features similar to the transformation used for spectral similarity search and similar to approaches described by other authors [16]. The range 500 to 3700  $\text{cm}^{-1}$  has been divided into 200 intervals of width 16  $\text{cm}^{-1}$  and the maximum absorbance per interval was used as the feature value. From these 200 fea-

tures, a set of 20 exhibiting maximum variances [15] has been selected separately for each hitlist.

Chemical structures were characterized by a set of 165 binary molecular descriptors (describing presence/absence of substructures or other structural properties) [17]; a set of 20 descriptors exhibiting maximum variances has been selected (separately for each hitlist). Software ToSiM<sup>2</sup> generates a pre-defined set of such descriptors (used in this work) but is also capable to generate descriptors for user-defined substructures. The result of comparing *n* molecular structures with *m* substructures is the substructure isomorphism matrix of size *n.m*, containing '0's and '1's; the later value if a substructure is part of a molecular structure.

(5) *Summary.* Investigation of the infrared or mass spectrum from an unknown mainly consists of

<sup>3</sup> Scan: Software for Chemometric Analysis. Available from: Minitab, 3081 Enterprise Drive, State College, PA 16801-3008, USA.

Table 1  
Substructure isomorphism matrix from an evaluation of the IR spectra hitlist of compound **II**

Hit <i>i</i>	Characteristic substructure															<i>c</i>	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15		
<i>X</i>	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1		
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	15
2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	15
3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	15
4	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	14
5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	15
6	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	15
7	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	15
8	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	15
9	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	13
10	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	15
11	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	15
12	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	14
13	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	13
14	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	10
15	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	15
16	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	12
17	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	12
18	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	15
19	0	0	0	0	0	1	0	0	1	0	0	1	0	1	0	0	4
20	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	15
21	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	15
22	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
23	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	10
24	1	1	1	1	1	1	1	1	1	1	0	1	0	1	1	1	14
25	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	10
26	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	4
27	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	15
28	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	15
29	1	1	1	1	1	0	0	1	0	0	1	0	0	0	0	0	7
30	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	10
31	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	4
32	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	2
33	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	2
34	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
35	1	1	0	0	1	0	0	1	0	0	0	1	0	0	0	0	5
36	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	4
37	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	4
38	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	15
39	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	15
40	1	0	1	0	1	1	1	0	0	0	0	1	0	0	0	0	6
41	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	14
42	1	1	1	1	1	1	0	0	0	1	0	1	0	0	0	0	9
43	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
44	1	1	1	1	1	0	0	1	0	0	1	0	0	0	0	0	7
45	0	0	0	0	0	1	0	0	1	0	0	1	0	1	0	0	4
46	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
47	1	0	1	0	1	1	1	0	0	0	0	0	0	0	0	0	5
48	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2

Table 1 (continued)

Hit <i>i</i>	Characteristic substructure															<i>c</i>	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15		
49	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
50	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	2
<i>f</i>	47	41	40	36	35	34	32	32	31	29	28	26	25	24	22		

Each row corresponds to a hitlist structure (*X* denotes the query structure); each of the 15 most frequent characteristic substructures corresponds to a column.

A '1' in the matrix denotes that the substructure is present in the hitlist structure

*i*: position of hitlist structure (*i* = 1 for the structure with the most similar spectrum).

*c*: number of characteristic substructures contained in a hitlist structure.

*f*: number of hitlist structures containing a characteristic substructure.

three steps: (i) spectral similarity search; (ii) determination of characteristic substructures from the hitlist; (iii) evaluation and application of the results. In some cases, the obtained substructures can be directly applied in a systematic structure elucidation—based on exhaustive isomer generation [18–21]—which needs the molecular formula together with structural restrictions. Evaluation of a hitlist can be supported by a cluster analysis of spectra and structures contained in the hitlist. Examples will demonstrate potentials and drawbacks of this strategy.

### 3. Results

A set of four alkyl-substituted butyrolactones is used to demonstrate the method and to compare results obtained from IR and MS data, respectively. Fig. 1 contains the obtained five most frequent substructures extracted by the described MCS approach from the IR- and MS-hitlists.

All substructures obtained from IR spectra are correct in the sense that they are contained in the query structure; their frequencies in the 50 hitlist structures range between 35 and 50; on the average the substructures are contained in 80 to 90% of the hitlist structures. Although the butyrolactone ring does not appear in the list of the five most frequent substructures, the results are informative for the query

structures: for all compounds a carbonyl group in the neighborhood of another oxygen atom is predicted; a long alkyl chain is obtained for compound **IV**. Results for another four alkyl-substituted butyrolactones—with only IR spectra being available—showed the same good performance of the method.

For MS data the size of the hitlist varied between 36 and 37 because of the different number of duplicates that had to be removed. The frequencies of the substructures range between 11 and 27; on the average the substructures are contained in 46 to 56% of the hitlist structures. The substructures obtained from MS data are less informative in this example than those obtained from IR for several reasons: two substructures are wrong; most substructures do not contain oxygen atoms; the hitlists contain a variety of different structures resulting in lower frequencies of the found substructures.

For a better insight, more details of the evaluation of IR data from compound **II** (4,4-dimethyl-butylolactone) are presented and a complementary multivariate method (PCA) is applied. Table 1 shows the substructure isomorphism matrix for the most frequent 15 characteristic substructures (columns) and the 50 hitlist structures (rows); for comparison also the query structure (*X*) is included. In the first ten hits almost all characteristic substructures are contained. In general, a low similarity of the spectra (corresponding to a high position number in the hitlist) causes a smaller number of matching substructures. However, exceptions of this trend occur:

hit 19 (a dichloro-bornanone) contains only four of the substructures while hit 38 (a tetrahydro-phthalic-anhydride) contains all 15.

Cluster analyses of the spectra and the structures have been performed by PCA to search for the different classes of compounds which are present in the hitlist. Fig. 2 shows scatter plots for the IR spectra (2a) and the corresponding structures (2b). The spectral data show a compact cluster (A) with 28 compounds; from the principal component loadings and the features can be concluded that the spectra of this cluster have their C=O band mainly in the interval 1764–1780  $\text{cm}^{-1}$ . Another cluster (B) contains nine compounds; the spectra all have their C=O band in interval 1780–1796  $\text{cm}^{-1}$ . The small cluster C contains four compounds, with the C=O band in interval 1748–1764  $\text{cm}^{-1}$ . Nine other compounds are spread across the plot. The spectrum of the unknown is located near the center of cluster A.

An inspection of the PCA plot for the corresponding chemical structures indicates existing spectra-structure relationships. The structural data show a cluster D with 26 compounds; 21 of them correspond to compounds in cluster A of the spectral data; 25 of them contain a butyrolactone ring; none of them is aromatic; in this case the hypothesis of 'similar spectra are from similar structures' is valid. Cluster E contains seven compounds; all have a benzene ring; however, the spectra of these compounds are not clustered. The remaining structures can be separated into two parts. Group F contains seven aro-

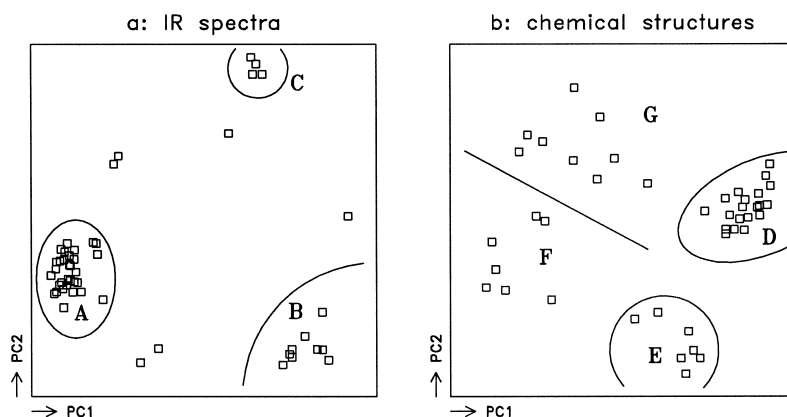


Fig. 2. PCA plots for IR spectra and corresponding chemical structures from the hitlist obtained for compound **II**. PC1, PC2, first and second principal component, respectively; (a) data from 50 IR spectra, retained variances in PC1 and PC2 are 28.2 and 16.2% of total variance, respectively; (b) data from 50 chemical structures, retained variances in PC1 and PC2 are 33.2 and 15.5%, respectively.

matic structures; four are acidic chlorides. The 10 structures in group G are all non-aromatic; six are acidic chlorides. The spectra corresponding to groups F and G do not cluster significantly. A data analysis by PLS [13,15] using spectral and structural data together did not yield more information about spectra–structure relationships than the two separate PCAs.

The PCA plots indicate which classes of compounds are present in the hitlist and for which a close relationship between spectra and structures exist. Cluster analysis of the IR spectra recognized automatically several classes of compounds according to their different carbonyl stretching frequencies.

#### 4. Conclusion

The described MCS-based approach for evaluation of IR spectra hitlists has been successfully applied to a set of lactones. The automatically obtained substructures are characteristic for the unknown and can be used in structure elucidation of compounds which are not present in the library. For the investigated compounds MS yielded less informative substructures than IR spectra. PCA has been shown as a useful chemometrics tool for an evaluation of a set of similar IR spectra and their corresponding chemical structures. The applied methods cannot provide automatic structure elucidation but may be a powerful support—even in routine work—if it would be implemented in spectral library search systems.

#### Acknowledgements

We thank R. Neudert of 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 format. We thank E. Ziegler from Max-Planck Institut für Kohlenforschung (Mülheim a.d. Ruhr, Germany) and

F. Friedli (MSP Friedli, Koeniz, Switzerland) for providing the MS database MassLib.

#### References

- [1] H.J. Luinge, *Vib. Spectrosc.* 1 (1990) 3.
- [2] J.T. Clerc, in: H.L.C. Meuzelaar, T.L. Isenhour (Eds.), *Computer-Enhanced Analytical Spectroscopy*, Plenum, New York, 1987, p. 145.
- [3] K. Varmuza, P.N. Penchev, H. Scsibrany, *J. Chem. Inf. Comput. Sci.* 38 (1998) 420.
- [4] H. Scsibrany, K. Varmuza, *Fresenius J. Anal. Chem.* 344 (1992) 220.
- [5] L. Chen, W. Robien, *J. Chem. Inf. Comput. Sci.* 34 (1994) 934.
- [6] F. Ehrentreich, *Fresenius J. Anal. Chem.* 357 (1997) 527.
- [7] SpecInfo: Spectroscopic Information System. Available from: Chemical Concepts, PO Box 100202, D-69442 Weinheim, Germany.
- [8] P.N. Penchev, A.N. Sohau, G.N. Andreev, *Spectroscopy Letters* 29 (1996) 1513.
- [9] *Wiley Mass Spectral Database*, 4th edn., Wiley, New York, USA.
- [10] H. Damen, D. Henneberg, B. Weimann, *Anal. Chim. Acta* 103 (1978) 289.
- [11] H. Scsibrany, K. Varmuza, in: D. Ziessow (Ed.), *Software Development in Chemistry*, Vol. 7, Gesellschaft Deutscher Chemiker, Frankfurt/Main, 1993, p. 77.
- [12] K. Varmuza, W. Werther, D. Henneberg, B. Weimann, *Rapid Communications in Mass Spectrometry* 4 (1990) 159.
- [13] W. Werther, K. Varmuza, *Fresenius J. Anal. Chem.* 344 (1992) 223.
- [14] K. Baumann, J.T. Clerc, *Anal. Chim. Acta* 348 (1997) 327.
- [15] D.L. Massart, B.G.M. Vandeginste, L.M.C. Buydens, S. DeJong, P.J. Lewi, J. Smeyers-Verbeke, *Handbook of Chemometrics and Qualimetrics*, Elsevier, Amsterdam, 1997.
- [16] M. Novic, J. Zupan, *J. Chem. Inf. Comput. Sci.* 35 (1995) 454.
- [17] K. Varmuza, H. Scsibrany, in: R. Moll (Ed.), *Software Development in Chemistry*, Vol. 9, Gesellschaft Deutscher Chemiker, Frankfurt/Main, 1995, p. 81.
- [18] K. Varmuza, W. Werther, *J. Chem. Inf. Comput. Sci.* 36 (1996) 323.
- [19] K. Varmuza, P.N. Penchev, F. Stancl, W. Werther, *J. Mol. Struct.* 408–409 (1997) 91.
- [20] T. Wieland, A. Kerber, R. Laue, *J. Chem. Inf. Comput. Sci.* 36 (1996) 413.
- [21] H. Schriber, E. Pretsch, *J. Chem. Inf. Comput. Sci.* 37 (1997) 879.