ELSEVIER

Computers and Chemistry 25 (2001) 231–237

Computers & Chemistry

www.elsevier.com/locate/compchem

Characteristic substructures in sets of organic compounds with similar infrared spectra

Plamen N. Penchev [a], Kurt Varmuza [b,*]

[a] *Department of Analytical Chemistry, University of Plovdiv, BG-4000 Plovdiv, Bulgaria*

[b] *Laboratory for Chemometrics, Institute of Food Chemistry and Food Technology, Vienna University of Technology, Getreidemarkt 9/160, A-1060 Vienna, Austria*

Received 3 April 2000; accepted 18 August 2000

Abstract

A method based on the determination of maximum common substructures is applied for the generation of substructures which are characteristic for a given set of molecular structures. The molecular structures are from hitlists obtained by spectral library searches; the hitlists contain those reference compounds, which have infrared spectra most similar to that from the query compound. The influences of various parameters of this method are investigated with the aim to improve the relevance of the obtained substructures for the structure of the query compound. © 2001 Elsevier Science Ltd. All rights reserved.

Keywords: Spectral library search; Structure elucidation; Maximum common substructure; Software

1. Introduction

The most widely used technique for computer-assisted identification of organic compounds is spectra library search with mass spectra (Varmuza, 2000) or infrared spectra (Luinge, 1990). The resulting hitlist typically contains some tens of reference spectra most similar to the spectrum of the unknown. If the unknown is contained in the spectral library a correct identification is often possible; presence of this situation is usually indicated by a high value of the spectra similarity measure for the first hit. If it has to be assumed that the unknown is not a member of the used library the hypothesis is usually applied that similar spectra indicate similar chemical structures (Baumann and Clerc, 1997; Clerc, 1987). Based on this assumption the spectroscopist evaluates structures and spectra in the hitlist with the aim to construct candidates for the unknown molecular structure of the query compound or at least to get hints, which substructures may be present and which may be absent.

Recently a method based on the concept of maximum common substructures (MCS) has been presented as an aid to evaluate hitlists with infrared spectra (Varmuza et al., 1998, 1999). This automatic method generates a set of substructures from the hitlist structures, and it has been shown that these substructures are often characteristic for the molecular structure of the unknown. Extraction of substructure information from the hitlist structures has the advantage to be independent from predefined substance classes and is a complementary method to other computer-assisted approaches such as application of correlation tables (Affolter et al., 1997; Debska et al., 1997), multivariate linear classifiers (Luinge et al., 1995) or neural network classifiers (Ricard et al., 1993; Novic and Zupan, 1995; Klawun and Wilkins, 1996; Munk and Madison, 1996).

* Corresponding author. Tel.: +43-1-5880116060; fax: +43-1-5880116091.

E-mail address: kvarmuza@email.tuwien.ac.at. (K. Varmuza).

0097-8485/01/$ - see front matter © 2001 Elsevier Science Ltd. All rights reserved.
PII: S0097-8485(00)00096-6

Other strategies based on the application of MCSs have been described for mass spectra (Cone et al., 1977; Lebedew et al., 1981; Scsibrany and Varmuza, 1992, 1993; Lebedew and Cabrol-Bass, 1998), ^{13}C-NMR spectra (Chen and Robien, 1994) and for IR spectra in combination with the application of fuzzy logic (Ehrentreich, 1997a,b, 1999).

In this paper a systematic exploration of the applicability of the MCS concept for structure elucidation of organic compounds based on IR spectra is presented. A quantitative measure for the reliability of the obtained substructure set is proposed, and the influence of parameters on the contents of these sets is studied.

2. Spectral libraries and software

The IR database used consists of 13 484 full-curve spectra for the spectral range 500–3700 cm^{-1} with a sampling interval of 4 cm^{-1}, corresponding to 801 data points. Origin of the spectral and structural data is the SpecInfo IR database (SpecInfo, 1996). The format of the original and the converted data has been described previously (Penchev et al., 1999; Varmuza et al., 1998).

The software IRSS (Penchev et al., 1996, 1998) was used for spectra similarity searches in the IR spectra library. Seven different algorithms for the comparison of IR spectra are implemented (Varmuza et al., 1998) comprising three methods for matching peak list data, and four methods for comparing full spectral curves. Software IRSS is available from author P.N.P.

The software ToSiM (Scsibrany and Varmuza 1994) was used for the evaluation of hitlist structures. It contains tools for the investigation of topological similarities in molecules, such as cluster analysis of chemical structures, as well as determination of large and maximum common substructures in a given set of structures. The software SubMat (Varmuza and Scsibrany, 2000) was applied for automatic determination which of given k substructures are contained in given n molecular structures. Software ToSiM and SubMat are available from author K.V.

All computations have been performed on Pentium personal computers, 300 MHz, running under MS-Windows 95 or NT.

3. Methods

3.1. Characteristic substructures

The maximum common substructure (MCS) of two chemical structures is defined here as the largest connected substructure that is present in the two given structures. The MCS can be considered as a measure and a description of the similarity of two structures. The MCS of a set of n structures may be very small or may even not exist if one or more exotic structure is contained in the set, a situation common with spectral hitlists. Characteristic structural properties of a set with n structures can be described by a set of appropriate substructures; in the applied approach each of them is the MCS of a pair of the given molecular structures (Scsibrany and Varmuza 1992; Varmuza et al., 1998, 1999). The applied method to generate a set with characteristic substructures has been described previously (Varmuza et al., 1998) and is only briefly summarized here. For each of the $n(n-1)/2$ possible pairs of molecular structures the MCS is determined; then for each MCS$_i$ the number of occurrences, n_i (frequency), in the n structures is counted. Finally the MCSs are ordered by their decreasing ranking weight R_i as defined in Eq. (1). The ranking considers both the frequency and the size of the substructures (which is given by the number of non-hydrogen atoms).

$$R_i = (1-f)\, n_i/n + f\, A_i/A_{\max} \qquad (1)$$

A_i is the number of non-hydrogen atoms in MCS$_i$; $A_{\max}$ is the maximum number of non-hydrogen atoms in the n investigated molecular structures; f is a user-adjustable factor ranging between 0 and 1. If f is zero only the frequency counts for the ranking; if f is 1 only the size is considered; the influence of this parameter is reported in Section 4. The obtained set of characteristic, large, and frequently occurring substructures characterizes common structural properties of the molecular hit list structures; the result is only less affected by outlier structures.

The isomorphism of substructures and the determination of MCSs are controlled by the parameters listed in Table 1. The IR spectrum of a compound depends on both the masses of the atoms and on the strengths of the bonds between them. Taking into account these facts the parameters were chosen as listed. The influence of two parameters marked by 'varied' is investigated in Section 4.

3.2. Effectiveness of a set with characteristic substructures

A library search algorithm is effective in the view of this work if the hitlist structures resemble the structure of the unknown compound to a great extent. A measure of effectiveness, E, is defined to characterize semi-quantitatively how well the found substructures fit to the query structure.

$$E = \sum p_i\, n_i\, A_i/(k\, n\, A) \qquad i = 1, \ldots, k \qquad (2)$$

n is the number of used hitlist structures; k is the number of characteristic substructures derived from the hitlist structures by the described MCS approach; n_i

Table 1
Parameters for the MCS algorithm

No.	Parameter	Values	Used value
1	Atom type to be checked	Yes/No	Yes
2	Heteroatoms considered as equivalent	Yes/No	No
3	Topology of atoms to be checked (atom is part of a ring, of an aromatic ring, of a chain)	Yes/No	No
4	Equal number of H-atoms on non-hydrogen atoms to be checked	Yes/No	Varied
5	Bond type to be checked	Yes/No	Yes
6	Multiple bonds considered as equivalent	Yes/No	No
7	Minimum number of non-hydrogen atoms in a MCS	2–10	2
8	Factor *f* for ranking the MCSs (Eq. (1))	0–1	Varied

and A_i are frequency and size of substructure i, respectively; A is the size of the query structure (measured by the number of non-hydrogen atoms). The penalty coefficient p_i is set to 1 if substructure i is present in the query structure, and to -1 otherwise. The sum is calculated from all k substructures. Division by A makes E less dependent of the size of the query structure. The maximum value of E is 1, however, the minimum value is not defined. Scaling of E for instance to the range 0–1 would be possible by using a penalty coefficient of zero (instead of -1) in the cases the substructure is not present in the query structure. A disadvantage of this penalty value is that the size of an erroneous substructure would have no influence on E; for this reason a not-defined minimum value for E seems to be the better choice. The size and frequency of the found characteristic substructures mainly influence the used effectiveness measure. The structural diversity of the substructures was not considered in this work.

For a given unknown the contents of the set with characteristic substructures in general depends on: (1) the quality of the spectral library; (2) the structural diversity of the reference compounds with respect to the query structure; (3) the size of the library; (4) the spectral similarity measure applied; (5) the number of hitlist structures used, and; (6) the parameters used for the determination of MCSs. The influence of different spectral similarity measures, of the number of hitlist structures used, and of some MCS parameters are reported Section 4.

4. Results and discussion

4.1. Library searches

Ten compounds randomly selected from the library were chosen as 'unknown' query compounds (Table 2). The number of non-hydrogen atoms in these compounds is between 5 and 38. The corresponding IR spectra were searched in the library using the spectra similarity criteria described below. The first hit (exact match) was removed from the hitlist, and the remaining compounds were used for the determination of MCSs.

Table 2
Ten compounds used as unknowns; size of molecules is measured by the number of non-hydrogen atoms

No	Compound name <CAS registry number>	Molecular formula	Size
1	Butylamine <109-73-9>	$C_4H_{11}N$	5
2	1-Pentanol, 5-bromo <34626-51-2>	$C_5H_{11}OBr$	7
3	Tetrahydropyran-4-methanol <14774-37-9>	$C_6H_{12}O_2$	8
4	1,3-Dioxolane-4-methanol, 2-vinyl- <4313-32-0>	$C_6H_{10}O_3$	9
5	Benzene, 1-methoxy-3-(1-propenyl)- <20112-91-8>	$C_{10}H_{12}O$	11
6	1-Amino-naphthalene <134-32-7>	$C_{10}H_9N$	11
7	1-Hexanone, 1-(3-pyridinyl)- <81418-03-3>	$C_{11}H_{15}NO$	13
8	4,7-Methano-1H-indene, 6-(diethoxymethyl)-3A,4,5,6,7,7A-hexane <67633-93-6>	$C_{15}H_{24}O_2$	17
9	1H-1,2,4-Triazole, 1-[2-[3-(2-fluorophenyl)-2-methylpropoxy]-3,3-dimethyl-1-butenyl]- <101975-44-4>	$C_{18}H_{24}N_3OF$	23
10	Proline, 1-benzoyl-4-(2,5-dichlorobenzoyl)-3-(1,1-dimethylethyl)-5-phenyl, ethyl ester <103430-68-8>	$C_{31}H_{31}NO_4Cl_2$	38

Table 3
Effectiveness E (Eq. (2)) of sets containing characteristic substructures for the ten unknowns listed in Table 2[a]

Compound No.	Effectiveness E					
	CC	SP	SD	AD	PM	m(1–4)
1	−0.547	−0.603	−0.549	−0.506	−0.483	−0.548
2	−0.127	−0.132	−0.134	−0.115	−0.223	−0.130
3	0.119	0.119	0.039	0.031	−0.097	0.079
4	0.369	0.359	0.365	0.308	−0.243	0.362
5	−0.060	0.008	−0.003	−0.045	−0.033	−0.024
6	−0.036	−0.048	−0.061	−0.047	−0.261	−0.048
7	0.121	0.086	0.061	0.096	0.067	0.091
8	0.237	0.232	0.232	0.241	0.010	0.235
9	0.060	−0.019	0.083	0.088	−0.028	0.072
10	0.102	0.099	0.079	0.097	0.080	0.098
Low quartile	−0.060	−0.048	−0.061	−0.047	−0.243	
Median	0.081	0.047	0.050	0.060	−0.065	
High quartile	0.121	0.119	0.083	0.097	0.010	

[a] Five different measures for spectral similarity (CC, SP, SD, AD, PM) have been applied. The first four similarity measures give similar results (CC is best); measure PM is significantly worse. m(1–4) is the median of the first four measures. Compound 4 and 1 yielded best and poorest results, respectively.

4.2. Comparison of spectral similarity measures

Four different similarity measures for full-curve spectral matches have been tested (Varmuza et al., 1998): correlation coefficient (CC), scalar product (SP), sum of squared differences (SD), and sum of absolute differences (AD); additionally a simple peak-matching algorithm (PM) for a forward peak search (Clerc, 1987) was used. For each unknown the IR spectrum was searched in the library applying all five-similarity measures separately; size of the hitlist in this investigation was 50 reference compounds. For each resulting substructure set the effectiveness E (Eq. (2)) was determined; the results are summarized in Table 3. The median of the effectiveness is highest for hitlists obtained with the correlation coefficient measure, and lowest for applying the peak-matching algorithm. This result is confirmed by a Wilcoxon matched-pairs test (Massart et al., 1997). Each of the four full-curve spectral similarity measures gives significantly better effectiveness than the peak-match algorithm (at a maximum statistical risk α of 0.1). These results are in agreement with previous reports (Ehrentreich 1997b; Penchev 1998; Varmuza et al., 1998) showing that the peak-match criterion usually gives good results only for spectral identity searches with the unknown contained in the library. The performances of the four similarity measures for full-curve data decreases in the order $CC > AD > SD > SP$ but only CC and SP exhibit a statistically significant difference. The effectiveness values obtained with the five spectra similarity measures exhibit high correlations but show great differences for the ten unknowns. The medians, m(1–4), calculated from the similarity measures CC, AD, SD and SP show that compounds 4 and 8 yielded best results while compound 1 is an outlier with worst effectiveness. For compound 4 (best results) and compound 1 (poorest results) the ten best ranked characteristic substructures (using the correlation coefficient measure for spectral similarity) are displayed in Fig. 1. For compound 4 all ten substructures are part of the query structure; for compound 1 only three substructures are contained in the query structure, the other seven are very similar to it.

Use of the intersection or the union of several hitlists—which have been obtained by applying different similarity measures — does not improve the results. In general, intersecting hitlists produces larger substructures than single hitlists. These carry more structural information than smaller ones, but on the other hand erroneous substructures are also larger and thus decrease the effectiveness; consequently the results for intersecting hitlists are typically the average of the effectiveness values from the single hitlists.

4.3. Optimum number of hitlist structures

The use of a too short hitlist for the generation of characteristic substructures may cause erroneous results if the query compound is not contained in the library. A too large hitlist may result in rather small and less informative substructures. In an extreme case the whole spectral library may be considered as the hitlist; of course this approach is useless because the found sub-

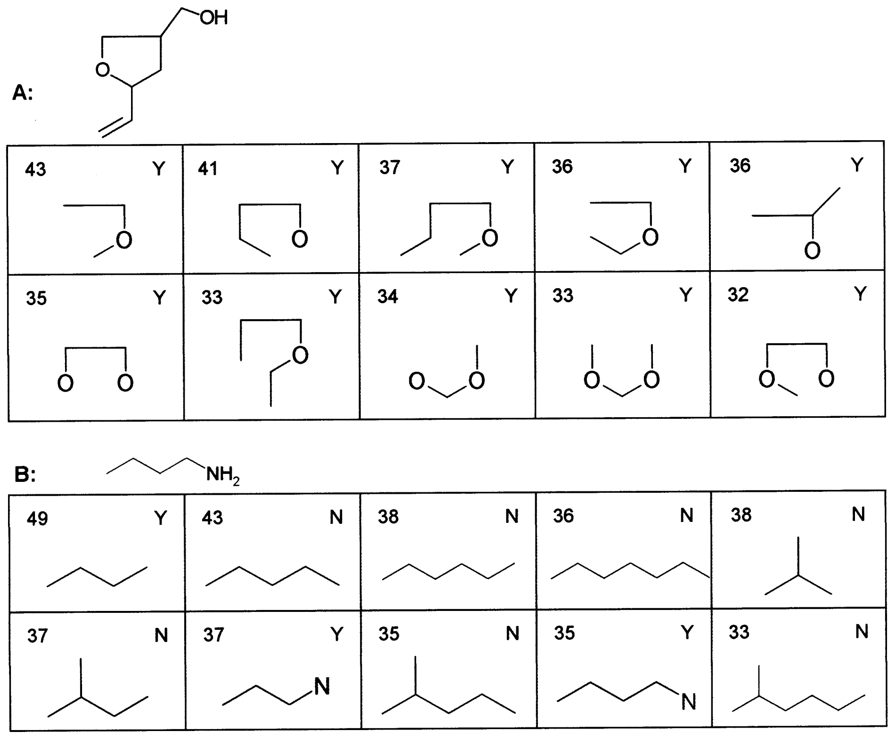

Fig. 1. Examples for characteristic substructures found by the MCS approach. (A) compound 4 (highest effectiveness); (B) compound 1 (lowest effectiveness). Compounds are listed in Table 2. *Y*, substructure is part of the query structure, *N*, substructure is not part of the query structure; the number given is the frequency (number of hitlist structures containing the substructure). Spectral similarity measure was the correlation coefficient, size of hitlist was 50, factor *f* for ranking was 0.3.

structures are characteristic for the library but not for the query compound. Based on these considerations an optimum number of hitlist structures may exist yielding maximum effectiveness. This assumption has been tested using the IR spectra from the ten compounds listed in Table 2; the spectral similarity measure was the correlation coefficient and the size of the hitlist was varied between 20 and 70. The mean, E_m, of the effectiveness values for the ten unknowns is plotted versus the number of used hits, n, in Fig. 2. Maximum performance is obtained with approximately 50 reference compounds in the hitlist. Considering that the optimum size of the hitlist also depends on the size and diversity of the library a range of 40–60 reference spectra can be recommended for the hitlist to be used in the MCS approach.

4.4. Parameter for MCS determination

The MCS of two structures depends on the used isomorphism criteria as listed in Table 1; some of them are relevant to IR spectroscopy. For instance, many characteristic bands in IR spectra are due to the vibrations of X–H bonds in the molecules. Therefore one may expect that the obtained substructures are more informative if the isomorphism criterion 'equal number of H-atoms on non-hydrogen atoms' is set to 'Yes'. This assumption has been tested using the IR spectra from the ten compounds listed in Table 2 as unknowns for library search; the spectral similarity measure was the correlation coefficient. From hitlists containing 50 reference compounds the MCSs have been determined

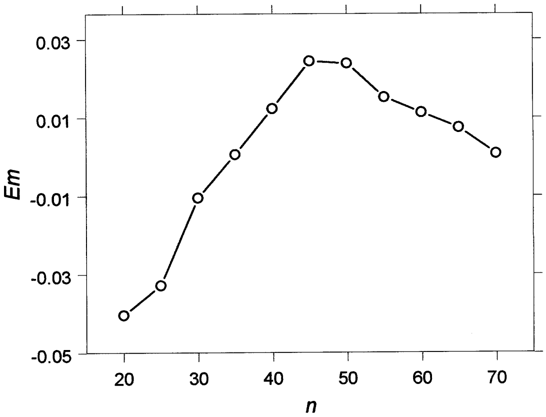

Fig. 2. Averaged effectiveness, E_m, versus number of used hitlist structures, n.

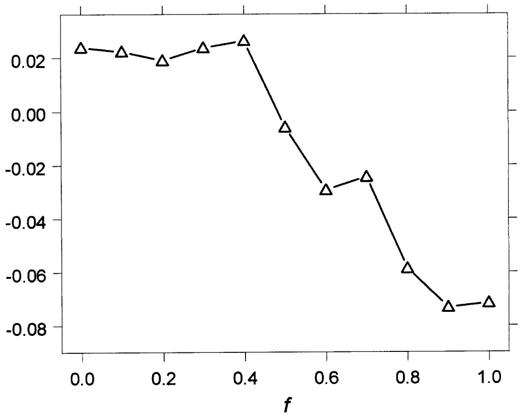

Fig. 3. Averaged effectiveness, E_m versus parameter f for ranking the maximum common substructures (Eq. (1)).

with this parameter set once to 'No' and then to 'Yes'. The Wilcoxon matched-pairs test on the two sets with effectiveness values indicated no significant difference at the $\alpha = 0.1$ level. An explanation for this results is that typically several X–H bonds of the same type are present in a molecule giving rise to vibrations in the same characteristic wavenumber interval with the wavenumber only less influenced by the number of hydrogen atoms at the X atom.

The other MCS parameter investigated is the factor f (ranging from 0 to 1) in Eq. (1), which influences the ranking, of the obtained MCSs. The larger f is the more the size of the MCSs becomes important instead of frequency. This dependence was investigated using again the IR spectra from the ten compounds listed in Table 2; the spectral similarity measure was the correlation coefficient, size of the hitlist was 50; factor f was varied between 0 and 1 in ten steps. In Fig. 3 the averaged effectiveness, E_m, is displayed versus f; the optimum value for f lies between 0 and 0.4. For larger values of f the size of the selected MCSs is much more important than their frequency; consequently MCSs may be selected that are large but are present in only a few hitlist structures (sometimes even only in structures at the end of the hitlist). Such substructures tend to be not characteristic for the query structure and therefore decrease the performance.

5. Conclusions

The structure-oriented evaluation of hitlists from spectral library searches deserves special attention if it has to be considered that the query compound is not a member of the library. For infrared spectroscopy an approach based on the determination of maximum common substructures from all pairs of hitlist structures can be successfully applied; result is a set of substructures, which are often relevant to the query structure. For an improvement of the method an effectiveness measure has been defined to judge semi-quantitatively the structural relevance of the derived substructures. Using this measure five spectral similarity criteria have been compared; best results were obtained with a criterion based on the correlation coefficient for full-curve spectra. The optimum size of the hitlist was found to be about 50 reference compounds. Also some parameters that control the determination of maximum substructures influence the effectiveness; tests show that the frequency of the substructures in the hitlist structures is more important than the size of the substructures.

The results from this investigation lead to an improved method for a structure-oriented evaluation of hitlists from IR library searches. The set of substructures automatically obtained by the MCS approach is capable to assist in structure elucidation of unknowns, which are not contained in the spectral library.

Acknowledgements

We thank Neudert of Chemical Concepts (Weinheim, Germany) for providing the SpecInfo IR database. We are grateful to the late Clerc and to Pretsch (ETH Zurich, Switzerland) for making this database available in an appropriate format. We thank Scsibrany for developing the ToSiM and SubMat software.

References

Affolter, C., Baumann, K., Clerc, J.T., Schriber, H., Pretsch, E., 1997. Automatic interpretation of infrared spectra. Mikrochim. Acta 14, 143–147 Suppl.

Baumann, K., Clerc, J.T., 1997. Computer-assisted IR spectra prediction — linked similarity searches for structures and spectra. Anal. Chim. Acta 348, 327–343.

Chen, L., Robien, W., 1994. Application of the maximum common substructure algorithm to automatic interpretation of ^{13}C-NMR spectra. J. Chem. Inf. Comput. Sci. 34, 934–941.

Clerc, J.T., 1987. Automated spectra interpretation and library search systems. In: Meuzelaar, H.L.C., Isenhour, T.L. (Eds.), Computer-Enhanced Analytical Spectroscopy. Plenum, New York, pp. 145–162.

Cone, M.M., Venkataraghavan, R., McLafferty, F.W., 1977. Molecular structure comparison program for the identification of maximal common substructures. J. Am. Chem. Soc. 99, 7668–7671.

Debska, B.J., Guzowska-Swider, B., 1997. Knowledge discovery in an infrared database. Comput. Chem. 21, 51–59.

Ehrentreich, F., 1997a. Derivation of substructures from infrared band shapes by fuzzy logic and partial cross correlation functions. Fresenius J. Anal. Chem. 359, 56–60.

Ehrentreich, F., 1997b. Representation of extented infrared spectrum-structure correlations based on fuzzy theory. Fresenius J. Anal. Chem. 357, 527–533.

Ehrentreich, F., 1999. Joined knowledge- and signal processing for infrared spectrum interpretation. Anal. Chim. Acta 393, 193–200.

Klawun, C., Wilkins, C.L., 1996. Optimization of functional group prediction from infrared spectra using neural networks. J. Chem. Inf. Comput. Sci. 36, 69–81.

Lebedew, K.S., Cabrol-Bass, D., 1998. New computer aided methods for revealing structural features of unknown compounds using low resolution mass spectra. J. Chem. Inf. Comput. Sci. 38, 410–419.

Lebedew, K.S., Tormyshev, V.M., Derendyaev, B.G., Koptyug, V.A., 1981. A computer search system for chemical structure elucidation based on low-resolution mass spectra. Anal. Chim. Acta 133, 517–525.

Luinge, H.J., 1990. Automated interpretation of vibrational spectra. Vib. Spectrosc. 1, 3–18.

Luinge, H.J., van der Maas, J.H., Visser, T., 1995. Partial least squares regression as a multivariate tool for the interpretation of infrared spectra. Chemom. Intell. Lab. Syst. 28, 129–138.

Massart, D.L., Vandeginste, B.G.M., Buydens, L.C.M., de Jong, S., Lewi, P.J., Smeyers-Verbeke, J., 1997. Handbook of Chemometrics and Qualimetrics: Part A. Elsevier, Amsterdam.

Munk, M.E, Madison, M.S., 1996. The neural network as a tool for multispectral interpretation. J. Chem. Inf. Comput. Sci. 36, 231–238.

Novic, M., Zupan, J., 1995. Investigation of infrared spectra-structure correlation using Kohonen and counterpropagation neural network. J. Chem. Inf. Comput. Sci. 35, 454–466.

Penchev, P.N., 1998. Application of chemometric methods for identification of organic compounds from their infrared spectra, Ph.D. Thesis, Bulgarian Academy of Sciences, Sofia.

Penchev, P.N., Andreev, G.N., Varmuza, K., 1999. Automatic classification of infrared spectra using a set of improved expert-based features. Anal. Chim. Acta 388, 145–159.

Penchev, P.N., Kochev, N.T., Andreev, G.N., 1998. IRSS: A program system for infrared library search. Comptes Rendus de l'Academie Bulgares des Sciences 50 (1–2), 67–70.

Penchev, P.N., Sohou, A.N., Andreev, G.N., 1996. Description and performance analysis of an infrared library search system. Spectrosc. Lett. 29, 1513–1522.

Ricard, D., Cachet, C., Cabrol-Bass, D., 1993. Neural network approach to structural feature recognition from infrared spectra. J. Chem. Inf. Comput. Sci. 33, 202–210.

Scsibrany, H., Varmuza, K., 1992. Common substructures in groups of compounds exhibiting similar mass spectra. Fresenius J. Anal. Chem. 344, 220–222.

Scsibrany, H., Varmuza, K., 1993. Toplogical similarity of molecules based on maximum common substructures. In: Ziessow, D. (Ed.), Software Development in Chemistry, vol. 7. Gesellschaft Deutscher Chemiker, Frankfurt am Main, pp. 77–87.

Scsibrany, H., Varmuza, K., 1994. ToSiM: PC-software for the investigation of topological similarities in molecules. In: Jochum, C. (Ed.), Software Development in Chemistry, vol. 8. Gesellschaft Deutscher Chemiker, Frankfurt am Main, pp. 235–249.

SpecInfo: Spectroscopic Information System, Chemical Concepts: PO Box 100202, D-69442 Weinhein, Germany 1996.

Varmuza, K., Penchev, P.N., Scsibrany, H., 1998. Maximum common substructures of organic compounds exhibiting similar infrared spectra. J. Chem. Inf. Comput. Sci. 38, 420–427.

Varmuza, K., Penchev, P.N., Scsibrany, H., 1999. Large and frequently occuring substructures in organic compounds obtained by library search of infrared spectra. Vib. Spectrosc. 19, 407–412.

Varmuza, K., Scsibrany, H., 2000. Substructure isomorphism matrix. J. Chem. Inf. Comput. Sci. 40, 308–313.

Varmuza, K., 2000. Chemical structure information from mass spectrometry. In: Lindon, J.C., Tranter, G.E., Holmes, J.L. (Eds.), Encyclopedia of Spectroscopy and Spectrometry. Academic Press, London, pp. 232–243.