Journal of Molecular Structure 408/409 (1997) 91–96

Journal of
MOLECULAR
STRUCTURE

Systematic structure elucidation of organic compounds by mass spectra classification

K. Varmuza[a,*], P. Penchev[b], F. Stancl[a], W. Werther[a]

[a]Technical University of Vienna, Department of Chemometrics, Getreidemarkt 9/152, A-1060 Vienna, Austria
[b]University of Plovdiv, Center of Analytical Chemistry and Applied Spectroscopy, BG-4000 Plovdiv, Bulgaria

Received 26 August 1996; accepted 6 September 1996

Abstract

Presence or absence of substructures is predicted from low resolution mass spectra by multivariate classification methods. Classification results are evaluated and then transformed to suitable structural restrictions for isomer generation. An example with a compound $C_{10}H_{12}O_3$ demonstrates potential applications of this approach for systematic structure elucidation of organic compounds. © 1997 Elsevier Science B.V.

Keywords: Chemical structures; Chemometrics; Computer chemistry; Mass spectrometry

1. Introduction

Most approaches for systematic structure elucidation of organic compounds date back to the DENDRAL project [1,2] and typically consist of three steps. In the first step (''plan'') restrictions about the chemical structure of an unknown are derived from experimental data, usually from spectra. Today, NMR data are widely used for this purpose [3–5]; infrared data are only rarely considered [6,7]; mass spectral (MS) data are not yet used in commercial software. Recently however, classifiers have been developed [8–11] that recognize chemical substructures from a low resolution mass spectrum, a method that has also been applied in the present paper. Substructures which are considered to be contained in the unknown molecular structure are collected in the so called goodlist, while forbidden substructures constitute the badlist. In the second step (''generate'') all isomers agreeing with the structural restrictions are generated exhaustively. This step requires the brutto formula of the unknown. Because the number of generated candidate structures is often high a third step (''test/select'') is necessary to reduce the list of candidates. For this purpose spectra simulation [12] is widely applied.

''Systematic'' means that each step of the structure elucidation process is clearly defined and can be documented; furthermore an exhaustive solution is obtained for the given structural restrictions. Systematic however does not mean free of errors. The crucial step still is prediction of structural restrictions from spectral data which today has to be performed mainly by heuristic or statistical methods.

This paper deals with mass spectra classification and focuses on methods for interfacing the first and the second step of the procedure described above.

* Corresponding author. Fax: +431-5811915; e-mail: kvarmuza@email.tuwien.ac.at

PII S0022-2860(96)09654-8

Structural information often cannot be formulated in simple substructures. On the other hand goodlist and badlist usually must contain strictly defined substructures because isomer generator programs only have rather limited capabilities for accepting generic substructures or general structural properties.

2. Mass spectra classification

The applied method for obtaining structural restrictions from a low resolution mass spectrum has already been described in detail [8]; only a summary is given here. Classification of mass spectra is based on numerical transformation of the spectral data, and multivariate discriminant methods. MS-classifiers for about 70 substructures or other structural properties have been implemented in the software MSCLASS [11], running under MS-DOS; typical computing time for the application of 100 classifiers to a single mass spectrum is less than 1s (Pentium, 166 MH$_2$. For each substructure up to four different classifiers are applied in parallel; they have been developed from different random samples of spectra by applying either a linear or a non-linear classification method. Result of classification is a list of "yes/no" answers indicating the presence or absence of substructures. If the estimated precision of an answer is below a user-defined threshold then output of the answer is suppressed. Extensive tests of the classifiers indicated that "no" answers are almost always correct, while "yes" answers are sometimes wrong. Table 1 shows results obtained for the mass spectrum [13] of benzene–acetic-acid, 3-hydroxy-, ethylester. A set of 160 classifiers has been applied; six answers are "yes", 53 are "no"; all these are correct; 101 further answers have a precision below 90% and are therefore not considered. This result list contains a lot of structural information which may be useful for inexperienced persons in interpreting the mass spectrum.

Table 1
Classification results for mass spectrum [13] of benzene–acetic-acid, 3-hydroxy-, ethylester, $C_{10}H_{12}O_3$

Answers	Precision	Substructure or class of compounds
Y	99	aromatic: CH_2–C_6H_4–O–(*o,m,p*)
YYYY	99	phenol
Y	96	ethyl ester
N	95	alkyl C_4 H_9
NN	98	alkyl C_5 H_{11}
NNN	96	alkyl C_6 H_{13}
NN	98	alkyl C_7 H_{15}
NN	94	alkyl C_8 H_{17}
NN	98	alkyl C_9 H_{19}
NN	98	alkyl C_{10} H_{21}
NN	99	alkyl C_{11} H_{23}
N	92	tertiary butyl
N	95	aryl-N
NN	98	aryl-Cl
N	90	C_6H_4–Br (*o,m,p* substituted)
N	90	naphthaline
N	99	phenol with 2 OH
NNNN	99	chlorophenol
N	92	$(CH_3)_2 > C{=}C$ isopropylidene
N	99	boron (any number)
N	99	bromine (any number)
NNN	95	silicon (any number)
NN	98	acetoxy CH_3–COO
NN	97	acetyl CH_3–CO
N	90	alcohol tertiary
N	92	amine tertiary
N	91	*n*-C_4H_9–O
N	97	C_2H_5–C=O
N	96	CF_3–C=O
N	99	CH_3-O-CH_2
N	99	-$N(CH_3)_2$
N	98	acetic acid ester
NN	92	methyl ester
NN	97	$(CH_2)_6$–C=O
NNNN	99	$(CH_3)_3$ Si

Y: answer "yes"; N: answer "no" (the number of characters indicates how many classifiers gave an answer for a particular substructure); estimated precision of answer (%), has been averaged if more than one answer is available).

Development and evaluation of a classifier for recognizing the substructure "benzene ring disubstituted by a CH_2-group and an oxygen in ortho-, meta- or para- position" is described briefly to demonstrate the method. Compounds containing one of the three substructures are considered as class 1; all others as class 2. The classifier has been calculated by linear discriminant analysis using a training set containing 150 spectra [13,14] from each class. A prediction set of the same size containing spectra not used in the training, served to estimate the precision of classification answers as a function of the discriminant variable [8]. The performance of a classifier is measured by the recall, which is defined as the percentage of spectra correctly classified at a given minimum precision (typical 90%).

Table 2
Classification results for benzyl-oxy substructures $-CH_2-C_6H_4-O-$ (ortho, meta or para). Class 1: compounds containing at least one of these three substructures; class 2: other compounds. General restrictions for compound selection [13]: maximum molecular weight 300, only compounds not containing P, Si or metal atoms

	Minimum precision (%)	Recall (%)				
		Class 1				Class 2
		ortho	meta	para	all	all
No. of tested spectra		153	80	229	462	2978
Correct	90	38.6	47.5	47.6	44.6	37.6
Wrong	90	0.7	0.0	1.3	0.9	1.8
Not classified	90	60.7	52.5	51.1	54.5	60.6
Correct	95	33.3	37.5	40.2	37.5	29.7
Wrong	95	0.0	0.0	0.4	0.2	1.4
Not classified	95	66.7	62.5	59.4	62.3	68.9

Table 2 contains evaluation data for a test set containing mass spectra [13] of 462 compounds from class 1 and 2978 compounds from class 2. If a minimum precision of 90% is required 44.6% of class 1 yield the correct answer, 0.9% the wrong answer, and 54.5% are not classified because the precision is below 90%. Spectra of class 2 are correctly classified in 37.6%, erroneously in 1.8%, and not classified in 60.6%. Increasing the precision to 95% reduces the number of wrong classifications but also reduces the recall. The high percentage of not classified spectra has to be tolerated in order to achieve a sufficient high minimum precision. Recalls for meta and para substituted compounds are slightly higher than for ortho compounds which can be explained by the more complex fragmentation pathways caused by the ortho effect [15].

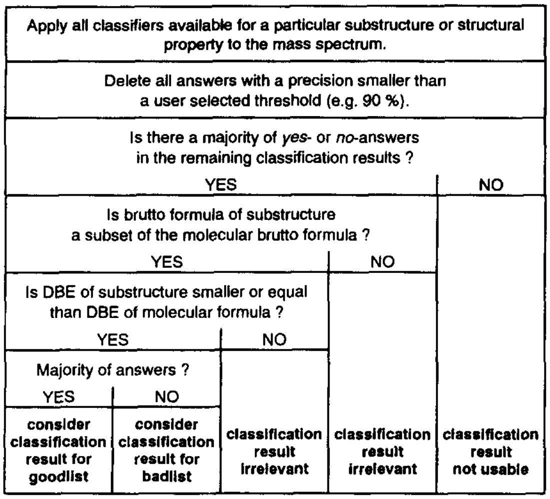
3. Isomer generation

The isomer generator software used was MOLGEN [16–18], version 3.0, running under MS-Windows. MOLGEN computes complete and redundancy free sets of connectivity isomers for given brutto formulas. The applicable structural restrictions are: goodlist (overlapping or not overlapping substructures), badlist, lower and upper limits for bond multiplicity and ring size. The only generalization allowed for substructures are dummy atoms with a fixed valence.

4. Evaluation and transformation of classification results

The classification result as obtained by software MSCLASS is a list of yes- and no-answers indicating presence or absence of substructures or more general structural properties. Structural information about the unknown may be enriched by results from other spectroscopic and analytical data as well as by preknowledge about the compound. This pool of information has to be used cautiously because it may be (a) only an incomplete description of the unknown structure, (b) redundant, (c) contradictory, (d) partly irrelevant, (e) hardly to be judged and or systematically be used by the chemist. The software MOLIN has been developed to serve as an interface between the result file produced by MSCLASS and the input file necessary for MOLGEN. MOLIN runs under MS-DOS, searches for contradictions in the classification results, summarizes the results and finally generates a file that defines as much as possible structural restrictions for direct use by MOLGEN.

If the molecular formula is considered to be known only a part of the classification answers may remain relevant. Assume for instance the molecular formula does not contain nitrogen. Then ''no''-answers for nitrogen-containing substructures are redundant information; ''yes''-answers however, indicate erroneous classifications. Fig. 1 shows the used decision scheme for checking classification results against the known



Apply all classifiers available for a particular substructure or structural property to the mass spectrum.

Delete all answers with a precision smaller than a user selected threshold (e.g. 90 %).

Is there a majority of *yes*- or *no*-answers in the remaining classification results ?

YES | NO

Is brutto formula of substructure a subset of the molecular brutto formula ?

YES | NO

Is DBE of substructure smaller or equal than DBE of molecular formula ?

YES | NO

Majority of answers ?

YES | NO

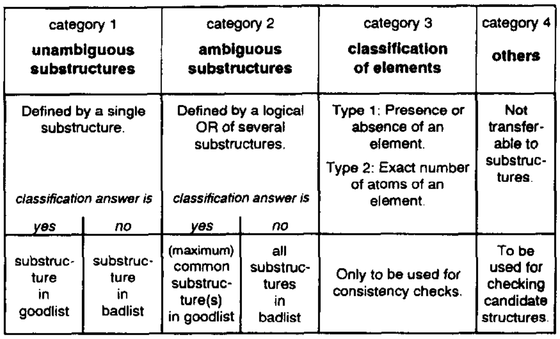
YES	NO			
consider classification result for goodlist	**consider classification result for badlist**	**classification result irrelevant**	**classification result irrelevant**	**classification result not usable**

Fig. 1. Check of the relevance of spectra classification results if the molecular formula is known. DBE, number of double bond equivalents.

molecular formula. If more than one classifier is applied for the same substructure, a simple majority rule determines the final answer; in the case of a tie the answer is suppressed.

Classification results can be divided into four categories (Fig. 2). Most useful for isomer generating are unambiguous substructures (category 1) that completely describe the structural information given by classifiers. Examples from Table 1 are ethyl ester and phenol. Full structural information can be used by the isomer generator in these cases, either in the goodlist or in the badlist.

A classifier of the second category recognizes a structural property that can be described by a logical OR of several unambiguous substructures. The above described classifier for "benzene ring, disubstituted by a CH_2-group and an oxygen in ortho-, meta- or para-position" is an example for this type. If the classification answer is "yes" then one of the three defining substructures must be present in each generated molecular structure (Fig. 3(a)). However, none of these three substructures can be put into the goodlist; only smaller substructures that are contained in all three defining substructures are appropriate. Usually a high amount of information can be obtained if the maximum common substructure (MCS) is used for the goodlist. In the example two equally sized MCSs exist (both containing a CH_2-group). Two other large common substructures (each containing an oxygen) however provide additional structural restrictions. Therefore, all four common substructures are put into the goodlist if the answer is "yes". Notice, that the complete structural information cannot be utilized by this approach. This would be possible if the isomer generator itself accepts a logical OR combination of goodlist substructures. If the answer of the classifier is "no" then all three defining substructures become members of the badlist and the complete information can be utilized. Two other examples for classifiers of category 2 are shown in Fig. 3(b,c). Determination of MCSs has been supported by software TOSIM [14]; the



category 1 **unambiguous substructures**		category 2 **ambiguous substructures**		category 3 **classification of elements**	category 4 **others**
Defined by a single substructure. *classification answer is*		Defined by a logical OR of several substructures. *classification answer is*		Type 1: Presence or absence of an element. Type 2: Exact number of atoms of an element.	Not transferable to substructures.
yes	*no*	*yes*	*no*		
substructure in goodlist	substructure in badlist	(maximum) common substructure(s) in goodlist	all substructures in badlist	Only to be used for consistency checks.	To be used for checking candidate structures.

Fig. 2. Categories of classification results and their use as structural restrictions in isomer generation.

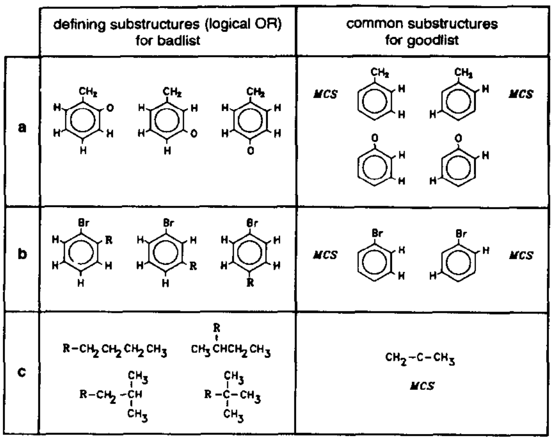

defining substructures (logical OR) for badlist
common substructures for goodlist
a
b
c
MCS
MCS
MCS
MCS
MCS
CH₃
H
O
Br
R
R-CH₂CH₂CH₂CH₃
CH₃CHCH₂CH₃
R-CH₂-CH
R-C-CH₃
CH₂-C-CH₃

Fig. 3. Representing a logical OR combination of substructures by one or several (maximum) common substructures. (a) Classifier for benzene ring, disubstituted by a CH_2-group and an oxygen in ortho-, meta- or para-position. (b) Classifier for ortho-, meta- or para substituted bromobenzenes. (c) Classifier for butyl groups. R, any substructure; MCS, maximum common substructure.

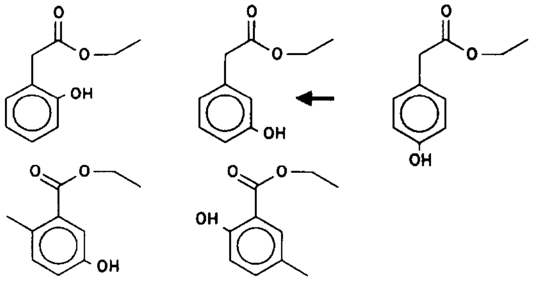


Fig. 4. Candidate structures for $C_{10}H_{12}O_3$ obtained from mass spectrum classification by software MSCLASS and exhaustive isomer generation by software MOLGEN. The correct structure is indicated by an arrow.

size of a substructure is measured by the number of atoms (including hydrogen atoms).

A third category of classification results recognizes presence, absence or even the number of atoms of certain elements in the molecular formula. Because the molecular formula is assumed to be known, these answers can only be used to detect contradictions.

The fourth category of classification results cannot be converted to (a reasonable number of) appropriate substructures. Examples are the recognition of hydrocarbons, or of alkyl-substituted chlorophenols. Such structural restrictions may be considered in the final test of the generated molecular structures.

5. Example

The brutto formula $C_{10}H_{12}O_3$ has more than 3×10^8 constitutional isomers, one of them is benzene–acetic-acid, 3-hydroxy-, ethylester. Classification results from Table 1 have been used as restrictions for isomer generation. The presence of a phenol group and an ethyl ester can be represented by unambiguous substructures in the goodlist. The structural property ''benzene ring, disubstituted by a CH_2-group and an oxygen in ortho-, meta- or para- position'' has been represented by the four substructures as discussed above and shown in Fig. 3(a). Classification answers ''no'' have been considered for the badlist as far as they are relevant to the given brutto formula. Applying these—strictly defined—structural constraints to isomer generation results in only five molecular structures, shown in Fig. 4. Computation time was 1 s for mass spectra classification, 0.5 s for evaluation and transformation of the answers, and 2 s for isomer generation (Pentium, 166 MHz). This example demonstrates that, for small molecules, a systematic and almost complete structure elucidation is possible even if only mass spectral data are available. In general, however, additional spectroscopic data or structure information will be necessary to obtain a reasonable sized list of candidate structures.

Acknowledgements

The authors are grateful to A. Kerber and R. Laue (University of Bayreuth, Germany) for making available their software MOLGEN. P.P. thanks the Austrian Academic Exchange Service for granting a scholarship.

References

[1] R.E. Carhart, D.H. Smith, N.A.B. Gray, J.G. Nourse and C. Djerassi, J. Org. Chem., 46 (1981) 1708.

[2] N.A.B. Gray, Computer-Assisted Structure Elucidation, John Wiley, New York, 1986.

[3] K. Funatsu and S.I. Sasaki, J. Chem. Inf. Comput. Sci., 36 (1996) 190.

[4] M. Will, W. Fachinger and J.R. Richert, J. Chem. Inf. Comput. Sci., 36 (1996) 221.

[5] H. Kalchhauser and W. Robien, J. Chem. Inf. Comput. Sci., 25 (1985) 103.

[6] M.E. Munk and M.S. Madison, J. Chem. Inf. Comput. Sci., 36 (1996) 231.

[7] C. Klawun and C.L. Wilkins, J. Chem. Inf. Comput. Sci., 36 (1996) 249.

[8] K. Varmuza and W. Werther, J. Chem. Inf. Comput. Sci., 36 (1996) 323.

[9] K. Varmuza, F. Stancl, H. Lohninger and W. Werther, Chemomet. Intell. Lab. Syst. (Lab. Aut. Inf. Mgmt), 31 (1995) 225.

[10] K. Varmuza, W. Werther, F. Stancl, A. Kerber and R. Laue, in J. Gasteiger (Ed.), Software Development in Chemistry, Gesellschaft Deutscher Chemiker, Frankfurt am Main, 1996, Vol. 10, pp. 303–314.

[11] MSCLASS: Software for Classification of Mass Spectra. Available from K. Varmuza, Department of Chemometrics, Technical University Vienna, Getreidemarkt 9 4/152, A-1060 Vienna, Austria.

[12] R. Bürgin Schaller, M.E. Munk and E. Pretsch, J. Chem. Inf. Comput. Sci., 36 (1996) 239.

[13] NIST Mass Spectral Database, version 4.0, 1992, National Institute of Standards and Technologoly, Gaithersburg, MD 20899, USA. Available from HD Science Ltd., 4a Bessel Lane, Nottingham NG9 7BX, UK.

[14] H. Scsibrany and K. Varmuza, in C. Jochum (Ed.), Software Development in Chemistry, Gesellschaft Deutscher Chemiker, Frankfurt am Main, 1994, Vol. 8, pp. 235–249.

[15] F.W. McLafferty, Interpretation of Mass Spectra. University Science, Mill Valley, CA, 1980.

[16] C. Benecke, R. Grund, R. Hohberger, A. Kerber, R. Laue and T. Wieland, Anal. Chim. Acta., 314 (1995) 141.

[17] C. Benecke, R. Grund, A. Kerber, R. Laue and T. Wieland, Match, 31 (1994) 229.

[18] MOLGEN, Isomer Generator Software. Available from A. Kerber and R. Laue, University of Bayreuth, Institute for Mathematics II, D-95440 Bayreuth, Germany.