communications in mathematical
match
and in computer chemistry

no. 38, October 1998

ISSN 0340-6253

MATCDY (38) 173-180 (1998)

MASSMOL

T. Grüner[1], A. Kerber, R. Laue, M. Meringer[1]

Universität Bayreuth, Mathematisches Institut,
D-95440 Bayreuth

K. Varmuza, W. Werther

Technische Universität Wien, Abteilung Chemometrie,
Getreidemarkt 9/152, A-1060 Wien

Most of the computer programs used for automatic spectra interpretation depend on large spectral databases. The experimental spectrum in question is compared with the entries of the database and the structures of the most similar spectra are given as possible solutions for the structure elucidation problem. This method has several limitations:

- The quality of the database spectra restricts the elucidation ratios of database dependent interpretation programs. Even good experimental spectra may lead to wrong results if the reference spectra are erroneous.
- If the spectrum of a query substance is not included in the library a reasonable result can only be expected if structures are in the database that are very similar to the unknown. For example, database search

[1]supported by DFG under KE 201/16-1

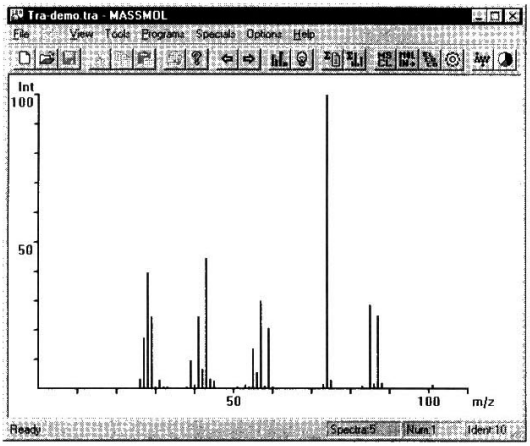
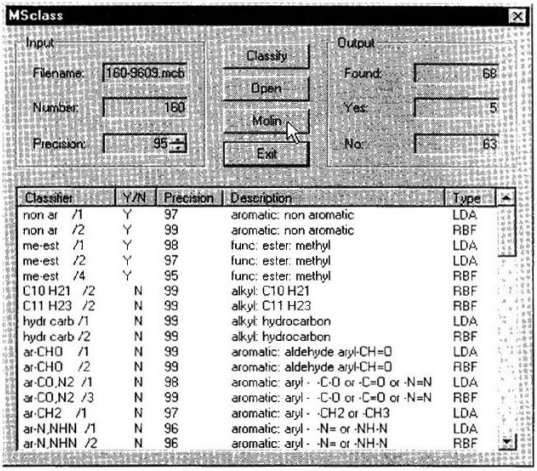


Figure 1: The MASSMOL main window.

seems to be inappropriate for the purposes of structure elucidation in the context of combinatorial chemistry, because many substances are synthesized for the first time.

Because of the above mentioned limitations of database searching, the authors' aim was to develop interpretation algorithms that are database independent, use mass spectra and can be coupled with the structure generator MOLGEN 4.0 (cf. [5]). We therefore implemented the program system MASSMOL, which is mainly adapted for low resolution electron impact mass spectroscopy. MASSMOL includes the following components:

MSclass (mass spectra classification)
predicts the presence or absence of substructures from low resolution mass spectra by multivariate classification methods [4]. In the current



Classifier	Y/N	Precision	Description	Type
non ar /1	Y	97	aromatic: non aromatic	LDA
non ar /2	Y	99	aromatic: non aromatic	RBF
me-est /1	Y	98	func: ester: methyl	LDA
me-est /2	Y	97	func: ester: methyl	LDA
me-est /4	Y	95	func: ester: methyl	RBF
C10 H21 /2	N	99	alkyl: C10 H21	RBF
C11 H23 /2	N	99	alkyl: C11 H23	RBF
hydr carb /1	N	99	alkyl: hydrocarbon	LDA
hydr carb /2	N	99	alkyl: hydrocarbon	RBF
ar-CHO /1	N	99	aromatic: aldehyde aryl-CH=O	LDA
ar-CHO /2	N	99	aromatic: aldehyde aryl-CH=O	RBF
ar-CO,N2 /1	N	98	aromatic: aryl - -C-O or -C=O or -N=N	LDA
ar-CO,N2 /3	N	99	aromatic: aryl - -C-O or -C=O or -N=N	RBF
ar-CH2 /1	N	97	aromatic: aryl - -CH2 or -CH3	LDA
ar-N,NHN /1	N	96	aromatic: aryl - -N= or -NH-N	LDA
ar-N,NHN /2	N	96	aromatic: aryl - -N= or -NH-N	RBF

Figure 2: The MSclass window shows the detailed classification results.

version there exist 160 classifiers for 86 different substructures or more general structural properties.

ElCoCo (elemental composition computation)
computes candidates for the molecular formula by recognition of interleaving isotope patterns in low resolution mass spectra [1].

Molin (molgen input converter)
serves as an interface between MSclass, ElCoCo and MOLGEN (cf. [3]) and allows the user's interaction during the elucidation process.

MOLGEN (molecular structure generator)
is able to compute in an efficient and redundancy free way all the connectivity-isomers that correspond to the molecular formulae pro-

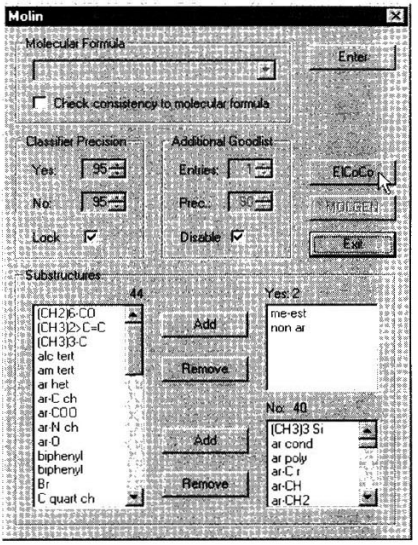


Figure 3: The Molin dialog allows user interaction for the choice of structural restrictions.

posed by ElCoCo and which fulfill the substructure restrictions obtained from MSclass, prepared by Molin as graphical input.

Figures 1-6 provide an example. They show a typical molecular structure elucidation using MASSMOL.

1. Figure 1 shows the main window with the mass spectrum of an unknown compound.

2. On the next screenshot we see the classification results of MSclass: There were 160 classifiers applied and 68 answers found with precision

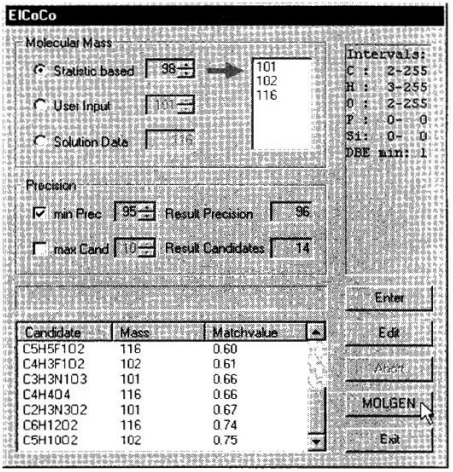


Figure 4: At the ElCoCo dialog the user can modify the precision parameters for the computation of most likely elemental compositions.

at least 95%. For instance we have three classifiers that indicate the functional group 'methyl ester' with precision 95%, 97% and 98%.

3. Figure 3 shows the Molin dialog. Here the results of the classifier are already transformed into substructure information. For our example we obtained two prescribed and 40 forbidden substructures that are displayed in the according listboxes. At this stage the user can edit these lists or enter the molecular formula for a consistency test. If the molecular formula is unknown, we proceed with the computation of good candidates.

4. Beginning at the mass of the highest peak of the first peak group, most probable suggestions for the molecular mass are now made. Figure 4

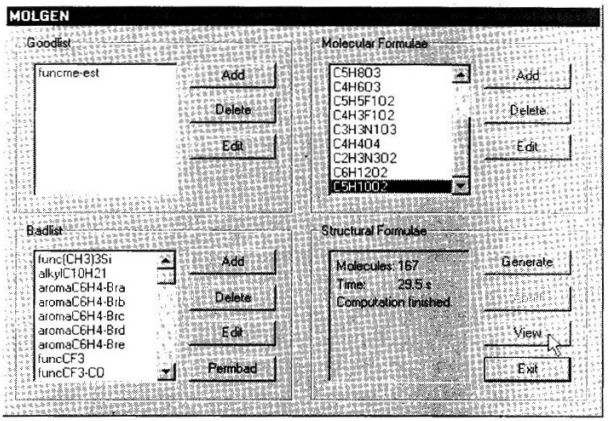


Figure 5: The MOLGEN window shows the input for the structure generator.

shows these suggestions in our example case: 101, 102 and 116.

5. Due to the prescribed substructure 'methyl ester' we get a minimum of two carbon, three hydrogen and three oxygen atoms for the molecular formula. At a precision of 95% we get 14 candidates for the molecular formula. According to these 14 molecular formulae MOLGEN (Figure 5) generates 167 connectivity-isomers that fulfill the structural restrictions.

6. The structural formulae can be displayed in a 2D arrangement (Figure 6). In our example the structure in the upper right corner, the methyl ester of n-pentanoic acid, was the one in question.

Apart from these basic modules needed for a typical elucidation process, MASSMOL offers some additional features that help to determine the molecular formula.

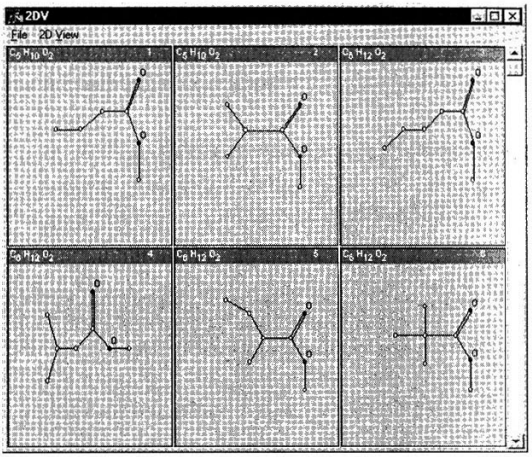


Figure 6: The structure viewer shows 2D arrangements of the generator output.

a) Theoretical isotope pattern and exact molecular weight of a given molecular formula can be computed.

b) All molecular formulae with relative molecular mass in a given interval can be generated.

c) Feature b) is also available for exact atom weights. This option is useful if high resolution data are available.

d) The intensities of the isotope peaks of the molecular ion can be entered in b) and the generated formulae will be ranked by the distance of their theoretical isotope pattern to the experimental data. This can be applied to soft ionisation mass spectra.

e) If atomic emission detector data are available, the molecular formula can be determined very exactly in accordance to the MS data.

Features a)-d) will also be available with full functionality in the free demo version for Windows 95/98/NT which can be found on the home page of MOLGEN

`http://www.mathe2.uni-bayreuth.de/molgen`

in due course.

References

[1] T. Grüner, A. Kerber, R. Laue, M. Liepelt, M. Meringer, K. Varmuza, W. Werther *Bestimmung von Summenformeln aus Massenspektren durch Erkennung überlagerter Isotopenmuster.* MATCH 37, 163-177 (1998).

[2] A. Kerber, R. Laue, T. Grüner, M. Meringer *MOLGEN 4.0.* MATCH 37, 205-208 (1998).

[3] K. Varmuza, P. Penchev, F. Stancl, W. Werther: *Systematic Structure Elucidation of Organic Compounds by Mass Spectra Classification.* Journal of Molecular Structure, Vol. 408/409, 91-96 (1997).

[4] K. Varmuza, W. Werther: *Mass Spectral Classifiers for Supporting Systematic Structure Elucidation.* Journal of Chemical Information and Computer Sciences, Vol. 36, 323-333 (1996).

[5] K. Varmuza, W. Werther, F. Stancl, A. Kerber, R. Laue: *Computer-assisted Structure Elucidation of Organic Compounds, based on Mass Spectra Classification and Exhaustive Isomer Generation.* Software Development in Chemistry, Vol. 10, 303-314 (1996).