IMPROVING THE EFFICIENCY OF SMALL GAS CHROMATOGRAPHIC-MASS SPECTROMETRIC DATA SYSTEMS BY MEANS OF SIMPLE ALGORITHMS

J. T. CLERC and M. KUTTER

Department of Organic Chemistry, Swiss Federal Institute of Technology, Universitätstrasse 16, CH-8006 Zürich (Switzerland)

M. REINHARD

Institute for Water Resources and Water Pollution Control, Swiss Federal Institute of Technology (EAWAG), CH-8600 Dübendorf (Switzerland)

and

R. SCHWARZENBACH

Department of Organic Chemistry, Swiss Federal Institute of Technology, Universitätstrasse 16, CH-8006 Zürich (Switzerland)

(First received October 10th, 1975; revised manuscript received January 14th, 1976)

SUMMARY

Computer methods are presented for the automatic identification of compound classes in gas chromatographic–mass spectrometric analyses. A simple approach is illustrated by the combination of mass chromatograms. The detection of polycyclic aromatic hydrocarbons is discussed in more detail in order to exemplify the application of a specialized algorithm.

INTRODUCTION

When complex mixtures are analyzed by means of gas chromatography–mass spectrometry (GC–MS), the amount of data produced is generally so large that the analyst is unable to process the data economically by conventional means, even if the GC–MS instruments are equipped with a computer data system. Usually, the analyst has to inspect most, if not all, of the mass spectra recorded, even when his interest is limited to members of particular classes of compounds or to single compounds. A partial interpretation of the mass spectra of all of the separated components must be made in order to decide whether the compound in question is relevant to the problem. Spectra of compounds found to be potentially relevant are then interpreted more carefully. The correct selection of the spectra which are to be processed further necessitates experience in the interpretation of mass spectra. Although the selection procedure is often quite simple, it consumes an excessive proportion of the analyst's time. Computer methods for performing this task are presented in this paper.

COMBINATION OF MASS CHROMATOGRAMS

An extremely useful tool for the processing of series of repetitively scanned mass spectra is a graph of the (absolute) abundances of selected fragment ions *versus* time. This technique is not limited to GC–MS applications (for a short review see ref. 1). In mass chromatography the abundances of fragment ions are plotted *versus* retention time. In simple cases, mass chromatograms may be used directly to identify mass spectra of potentially relevant compounds. However, the recall and precision generally do not even meet moderately high standards. Better results can be expected if two or more mass chromatograms are combined by means of simple mathematical operations. The selection of the mass chromatograms as well as the method of combination has to be matched to the particular problems (*cf.* refs. 2 and 3). However, acceptable results may often be obtained with sub-optimal parameters. This method will be demonstrated in the following examples.

Identification of compound groups by combination of mass chromatograms

In a study of the chlorination of polluted water[4], all of the aliphatic compounds containing a CCl_3 group had to be detected. A mass chromatogram for $m/e = 117$ corresponding to the CCl_3 fragment seems to be appropriate. However, mass spectra of alkylated benzenes with alkyl chains of three or more carbon atoms also have a peak at this mass number, thus incorrectly indicating the presence of a CCl_3 group. As alkyl-substituted benzenes also exhibit a strong peak at $m/e = 91$, which is absent from the spectra of the chlorinated aliphatic compounds expected in this study, a graph of the positive difference between the intensities at $m/e = 117$ and at $m/e = 91$ eliminates all of the interference from the aromatic compounds.

Fig. 1 shows the mass chromatograms for $m/e = 117$ (A) and for $m/e = 91$ (B), together with their positive intensity difference $A - \text{Min}(A, B)$. Obviously, the combination of the two simple mass chromatograms greatly reduces the number of spectra which require manual checking. However, the suggested procedure for the detection of aliphatic compounds containing a CCl_3 group is not free of error. Any aliphatic compound which has a CCl_3 group and which does not exhibit a peak at $m/e = 117$ in its mass spectrum, or which shows a peak of higher intensity at $m/e = 91$ than at $m/e = 117$, will be lost. In addition, any compound which does not contain a CCl_3 group, but which exhibits more intensity at $m/e = 117$ than at $m/e = 91$, will be picked out erroneously. An example of this is compound 5 in Fig. 1, which does not contain chlorine but does contain nitrogen, and has a molecular mass of 117. Of course, appropriate combinations with a third chromatogram, using for example a peak at $m/e = 119$ (^{37}Cl isotope peak of $m/e = 117$) or at $m/e = 47$ (CCl^+), would easily remove such errors.

In the same study, monochlorinated C_3-alkylated benzenes were of interest. The reconstructed total ion-current chromatogram (TIC) of the product mixture is given in Fig. 2. The graph of the positive intensity difference between peaks at $m/e = 154$ (molecular ion) and $m/e = 153$ ($M^+ - 1$, absent in the compounds of interest) is presented in Fig. 3A. An analogous graph for the peaks at $m/e = 156$ (isotope peak of the molecular ion) and at $m/e = 155$ is given in Fig. 3B. Both graphs fail to give adequate precision. However, their product, obtained by multiplication of corresponding amplitude values, clearly indicates the compound of interest. A comparison
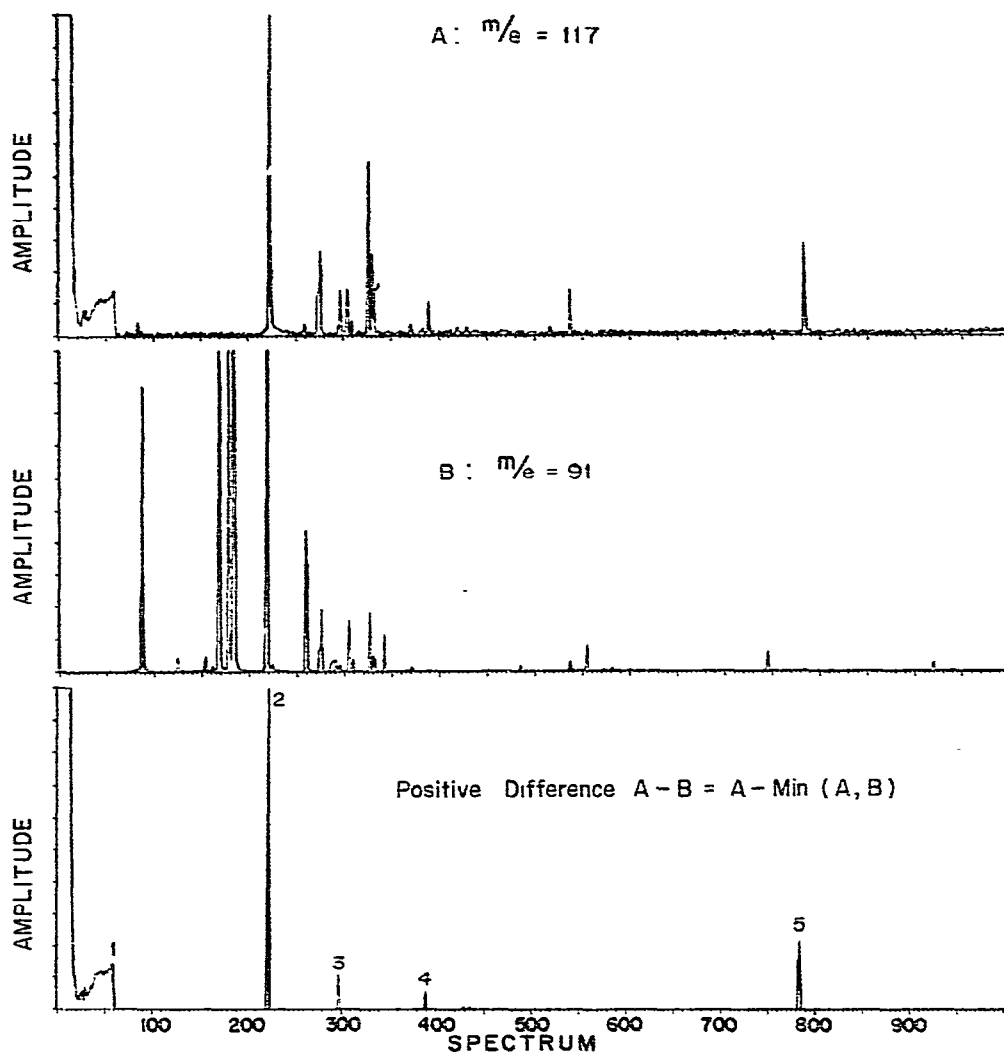
Fig. 1. Normalized mass chromatograms for $m/e = 117$ (A), $m/e = 91$ (B) and the positive intensity difference A–B. 1 = CCl$_4$, CHCl$_3$; 2 = CH$_3$COCCl$_3$; 3 = C$_2$H$_5$CCl$_2$COC$_2$H$_5$; 4 = unknown; 5 = C$_8$H$_7$N.

with Fig. 2 illustrates the degree of data reduction obtained. The described approach may be applied to any set of compounds which give rise to specific fragment ions at specified mass numbers.

With the exception of the multiplication of the two difference chromatograms, which was made off-line at the computing centre of ETH-Zürich, all calculations and measurements were performed on a Finnigan Model 6000 Interactive Data System. The gas chromatograph was equipped with a Grob-type injector[5] and a 50-m OV-101 glass capillary column (H. Jaeggi, Trogen, Switzerland).
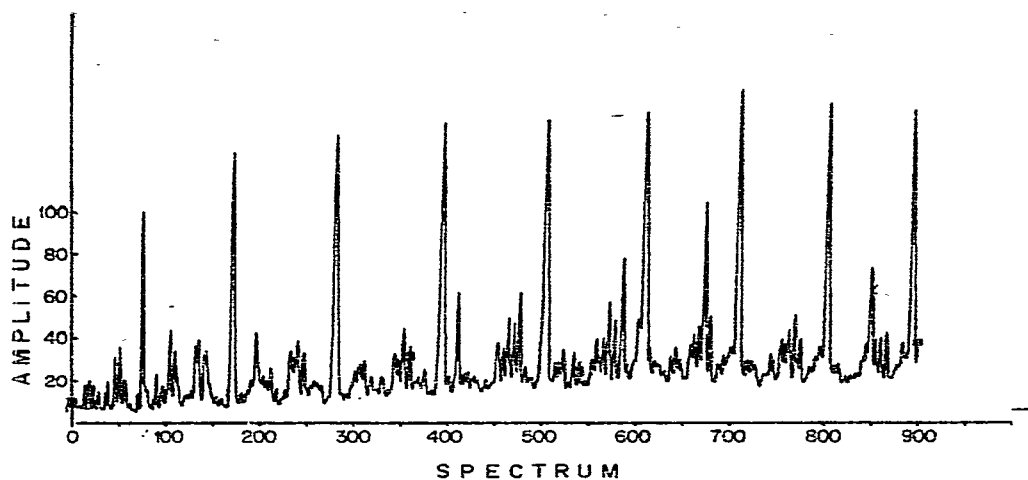
Fig. 2. Reconstructed TIC chromatogram of chlorinated diesel fuel.

SPECIALIZED ALGORITHMS

There are many classes of compounds which cannot be adequately described by simple combinations of fragment abundances. If the analysis of compounds belonging to such a class constitutes an important part of the routine work, it becomes economically feasible to develop special algorithms for the detection of these compounds. The algorithms use more complex features of the mass spectra in order to identify the compounds of interest with enhanced selectivity and precision. For example, ion series may be used for the identification of the class of relatively simple homologous compounds[6-8]. The use of more complex features permits the identification of certain structural elements[9]. If library spectra for all of the compounds of interest are available, reversed file search[10] may be applied.

Special algorithms have to be optimized for best recall rather than for best precision, as it is of primary importance not to overlook any compound of interest. If a few other compounds are picked out erroneously, they may be easily identified either manually or automatically by means of a library search procedure[6]. Furthermore, the algorithm must be relatively simple so that it may be applied on the mini computers which are used in GC–MS data systems.

Description of an algorithm for the detection of polycyclic aromatic hydrocarbons

As a model case, a special algorithm has been developed for the detection of polycyclic aromatic hydrocarbons through analysis of their low-resolution mass spectra. This class of compound has been selected for two reasons. Polycyclic aromatic hydrocarbons are of paramount interest in pollution analysis[3], and their identification by means of simple methods (*e.g.*, mass chromatography or reversed file search[10] is extremely difficult, if not impossible.

The mass numbers of important peaks are not suitable characteristics for the spectra of polycyclic aromatic hydrocarbons, since they vary widely for different species of the group. Also, their modulo 14- and modulo 13-spectra[9], as well as the common aromatic ion series[11], are not characteristic. However, in the spectra of
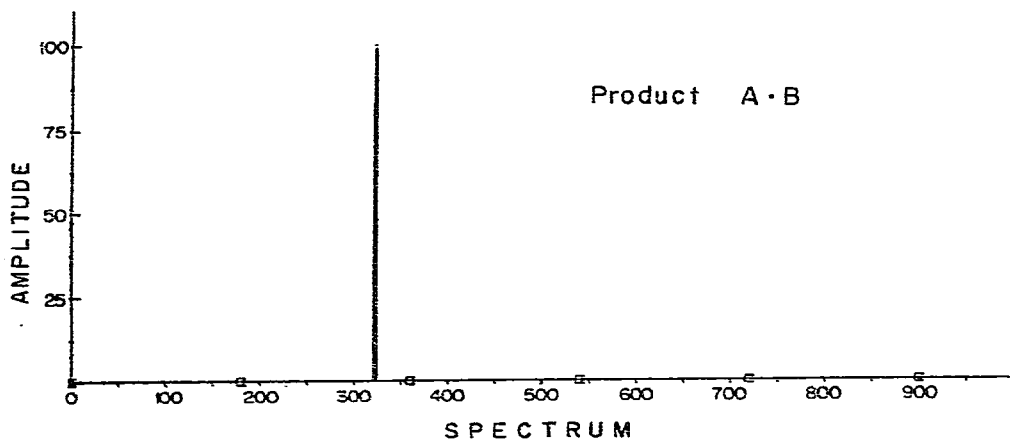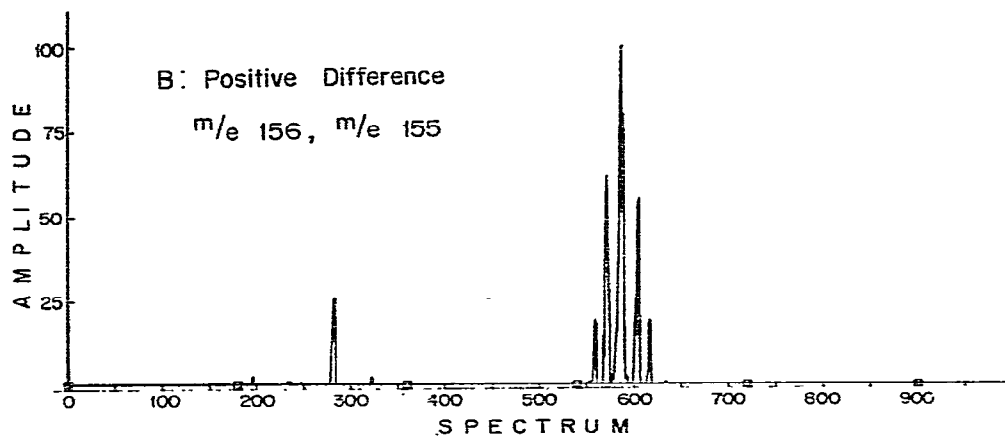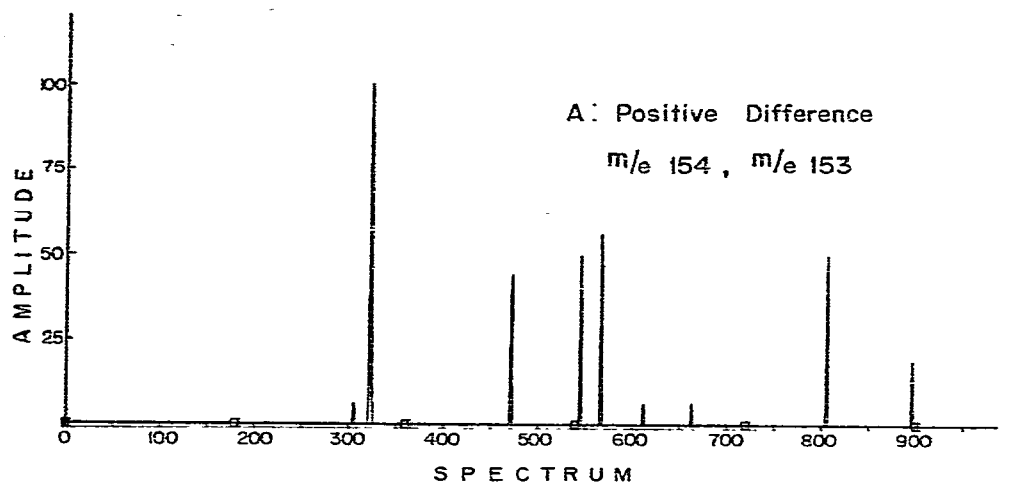
Fig. 3. Positive intensity difference between mass chromatograms for $m/e = 154$ and $m/e = 153$ (A), between mass chromatograms for $m/e = 156$ and $m/e = 155$ (B) and for the product A·B.

virtually all of the polycyclic aromatic hydrocarbons certain mass differences between significant peaks occur with a much higher frequency than in the spectra of other classes of compounds. In addition, an important part of the total ion current is concentrated around the base peak and around the peak at half its mass number. The described algorithm uses these two characteristics.

First, the significant peaks in the mass spectrum are identified. A peak is considered significant if it has no peak of higher intensity within the range given by its mass number $\pm$ 6 a.m.u. Thus, significant peaks correspond to local maxima in intensity. Only significant peaks are retained. Their intensity is encoded into two bits. The base peak receives code 11, peaks with intensities of less than 10% of that of the base peak receive code 01, and all of the other peaks receive code 10. All further processing is carried out with this reduced mass spectrum. A pseudo total ion current is then calculated by arbitrarily assigning intensities of 3, 30 and 100 to peaks with intensity codes 01, 10 and 11 respectively.

The frequencies of occurrence of certain mass differences are then counted. Whenever two peaks occur at mass numbers differing by 11, 13, 24, 26, 37, 39, 52, 63, 65, 74, 76, 87 or 89, the register of a counter is increased by one. The number of peaks at these mass numbers is also added to the value of the register. If after processing the reduced spectrum the value of the register exceeds 32, the compound is definitively marked as a polycyclic aromatic hydrocarbon. If the value of the register is less than 8, the compound is discarded. Spectra which exhibit values of between 8 and 32 are processed further.

In these spectra, all of the peaks at low mass numbers expected for hydrocarbons are cleared. The corresponding mass numbers are given in Table I. Further, all of the low-intensity peaks (*i.e.*, those with intensity code 01) between *m/e* 100 and the mass number of the base peak, as well as all of the peaks at certain selected mass numbers related to the mass number of the base peak, are deleted (Table II). For unsubstituted polycyclic aromatic hydrocarbons the spectrum should now be virtually empty. This is also true if the nucleus is substituted by short aliphatic chains. The pseudo total ion current for the remaining spectrum is now calculated and expressed as a percentage of the original pseudo total ion current. This number is subtracted from the value of the register and the result is then increased by 35. The number thus produced is re-tested as before. If it exceeds 32, the compound in question is considered as a polycyclic aromatic hydrocarbon.

TABLE I

MASS NUMBERS OF PEAKS TO BE DELETED DURING STEP 2

	m/e
Aromatic sequence	37, 38, 39, 50, 51, 52, 63, 64, 65, 76, 77, 78
Aliphatic sequence	27, 28, 29, 41, 42, 43, 55, 56, 57, 69, 70, 71
Sequence related to the mass number of the base peak (meb)	meb/2−1, meb/2, meb−29, meb−15, meb−14, meb

Finally, an elemental composition is assigned to the base peak under the assumption that the corresponding fragment is a hydrocarbon having more than five hydrogen atoms. If a preliminary calculation results in more than 28 carbon

atoms, then the minimum number of hydrogen atoms is set at 18 and the number of carbon atoms is recalculated. Next, the degree of unsaturation for the base-peak fragment is evaluated. If less than six double bonds or rings are indicated, the compound is eliminated.

The parameters used in the described algorithm, as well as the values assigned to them, were selected by first analyzing the mass spectra of 80 selected polycyclic aromatic hydrocarbons and then generalizing the result, using the rules of interpretation of empirical mass spectroscopy.

RESULTS

In order to estimate the possibilities and limitations of the described procedure, the algorithm was applied to a set of 8949 mass spectra taken from the Aldermaston library[12]. The necessary programs were written in Fortran and were run on the CDC 6500 computer at the ETH computing centre. From the 8949 spectra processed, 490 were picked out as polycyclic aromatic hydrocarbons. Of these 490 compounds, 209 are in fact polycyclic aromatic hydrocarbons, some of them being substituted by aliphatic chains. The precision of the algorithm is therefore 209/490 = 0.426. This value may seem rather low at first. However, the enormous reduction in the number of mass spectra to be interpreted manually (from *ca.* 9000 to *ca.* 500) has to be taken into account. Furthermore, in the set of erroneously retrieved spectra, there are 135 polycyclic heteroaromatic compounds and 54 hetero-substituted polycyclic aromatic hydrocarbons. In addition, there are many substituted benzenes, polyphenyls and highly halogenated aromatic compounds. Only eight truly non-aromatic compounds were retrieved.

As stated before, the most important parameter is recall, *i.e.*, the number of polycyclic aromatic hydrocarbons found in relation to the total number of such compounds present in the set. The test collection used in this study contains 237 polycyclic aromatic hydrocarbons, of which 209 have been found by the described procedure. The recall is therefore 209/237 = 0.882. Of the 28 polycyclic aromatic hydrocarbons which were missed, 18 are substituted with long aliphatic chains and five showed incomplete mass spectra or contained obvious errors.

CONCLUSIONS

The results of this feasibility study clearly illustrate the high degree of data reduction attainable with relatively simple algorithms. As such algorithms have to be tailored exactly to the problem at hand, they should be formulated by the user since he is the expert on the relevant compounds. Only then can a selection procedure result in optimum performance. In order to construct algorithms for the automatic detection of various selected classes of compounds on a commercial GC–MS data system, the user must have direct access to all of the acquired data and be able to process them with programs written in a powerful high-level language. We believe that speeding up the development of the appropriate software would be a rewarding activity in this field.

REFERENCES

1 D. Henneberg, B. Weinmann and E. Ziegler, *Chromatographia*, 7 (1974) 483.
2 M. Reinhard, V. Drevenkar and W. Giger, *J. Chromatogr.*, 116 (1976) 43.
3 W. Giger and C. Schaffner, *Advan. Org. Geochem.*, (1975) in press.
4 M. Reinhard, *Thesis*, ETH-Zürich, 1976.
5 K. Grob and G. Grob, *Chromatographia*, 5 (1972) 3.
6 T. O. Grönneberg, N. A. B. Gray and G. Eglinton, *Anal. Chem.*, 47 (1975) 415.
7 D. H. Smith, *Anal. Chem.*, 44 (1972) 536.
8 L. R. Crawford and J. D. Morrison, *Anal. Chem.*, 40 (1968) 1469.
9 J. Franzen, *Chromatographia*, 7 (1974) 518.
10 F. P. Abramson, *Anal. Chem.*, 47 (1975) 45.
11 K. S. Kwok, R. Venkataraghavan and F. W. McLafferty, *J. Amer. Chem. Soc.*, 95 (1973) 4185.
12 Mass Spectral Data Collection, United Kingdom Atomic Energy Authority, Aldermaston.