JOURNAL OF MASS SPECTROMETRY, VOL. 32, 1253–1257 (1997)

Automated Processing of GC/MS Data: Quantification of the Signals of Individual Components

Wim G. Pool,[1]* Leo R. M. Maas,[1] Jan W. de Leeuw[1] and Bastiaan van de Graaf[2]
[1] Netherlands Institute for Sea Research (NIOZ), Postbus 59, NL-1790 AB Den Burg, Netherlands
[2] Laboratory of Organic Chemistry and Catalysis, Technical University Delft, Julianalaan 136, NL-2628 BL Delft, Netherlands

An algorithm is described to quantify the signals of components in GC/MS data. It is an extension of the backfolding algorithm described recently. [W. G. Pool, B. van de Graaf and J. W. de Leeuw, *J. Mass. Spectrom.* 31, 509 (1996); 32, 438 (1997)]. The method is evaluated on both simulated and real GC/MS data. The results indicate that the method performs quite well, even in cases of components with highly similar spectra and with severe coelution. © **1997 John Wiley & Sons, Ltd.**

J. Mass Spectrom. 32, 1253–1257 (1997)
No. of Figures: 5 No. of Tables: 3 No. of Refs: 15

KEYWORDS: deconvolution; GC/MS; backfolding; spectrum clean-up; quantification

INTRODUCTION

Data sets obtained with GC/MS contain hundreds of spectra. Generally with complex samples even the chromatographic resolution of a high-quality capillary column is not sufficient to completely separate all the components. As a result, several components can contribute to the same spectrum. This obviously hampers the identification and quantification of individual components. Several mathematical routines have been developed to assist in the analysis of GC/MS data.[1–8] Recently a new two-step algorithm called backfolding was introduced.[9,10] In the first step, ion intensities in a scan are subtracted, mass by mass, from those in the next scan and positive and negative results are stored in separate sets of differentiated data.[11,12] In the second step these separate sets of differential data are recombined.[9] In the data thus obtained, background is eliminated and chromatographic resolution is improved. This two-step algorithm is repeated until no further improvements are observed. It has been shown that pure component spectra can be obtained automatically from the backfolded GC/MS data.[10] In the present paper an extension to the backfolding algorithm is described that quantifies the signal of the individual components detected.

THEORY

The intensities in the data matrix **D** obtained with GC/MS are first corrected (unskewed)[13] for the changes in concentration during each scan:

$$\mathbf{D} \rightarrow \mathbf{D}_u \tag{1}$$

The unskewed data can be written as the matrix product

$$\mathbf{D}_u = \mathbf{CFS} \tag{2}$$

where **C** is a chromatogram matrix with all chromatograms normalized to unit areas in its columns, **S** is a spectrum matrix with all spectra normalized to unit intensities in its row and **F** is a diagonal matrix with quantitative factors for the components. When backfolding is applied, new matrices are formed according to

$$\mathbf{D}_u \rightarrow \mathbf{B}_1 \rightarrow \mathbf{B}_2 \rightarrow \rightarrow \rightarrow \mathbf{B}_n \tag{3}$$

in which $\mathbf{B}_n$ is the backfolded data set obtained after n cycles of the backfolding algorithm. In analogy with Eqn (2), one can write

$$\mathbf{B}_n = \mathbf{C}_n \mathbf{F}_n \mathbf{S} \tag{4}$$

in which $\mathbf{C}_n$ and $\mathbf{F}_n$ are chromatograms and quantitative factors respectively after n cycles of the backfolding process. Component spectra are not influenced by backfolding, so **S** is not indexed. An algorithm to derive **S** has been described previously.[10] As shown in the Appendix, the backfolding algorithm is not quantitative and therefore $\mathbf{F}_n$ does not reflect the concentration of the components in the sample analysed.

In principle the matrix product **CF** can be computed by

$$\mathbf{CF} = \mathbf{D}_u \mathbf{S}^+ \tag{5}$$

where $\mathbf{S}^+$ denotes the Moore–Penrose generalized inverse of **S**. The calculation of a generalized inverse requires the inversion of the matrix product ($\mathbf{SS}^T$). This will be a problem when spectra (rows) in **S** are similar. In such a case ($\mathbf{SS}^T$) is close to singular and application

* *Correspondence to*: W. G. Pool, Netherlands Institute for Sea Research (NIOZ), Postbus 59, NL-1790 AB Den Burg, Netherlands. E-mail: pool@nioz.nl

CCC 1076–5174/97/111253–05 $17.50

Received 6 May 1997
Accepted 15 August 1997

of Eqn (5) will result in distorted chromatograms (see Results and Discussion).

The problem of singularity is bypassed when the computation of $\mathbf{S}^+$ is performed for the spectra in **S** separately. The subsequent multiplication of the spectrum's generalized inverse with the unskewed data matrix $\mathbf{D}_u$, however, still results in poor chromatograms when components with similar spectra are present. Such chromatograms show multiple maxima and require additional routines to extract the relevant information.

The problem with similar spectra can be circumvented almost completely by computing $\mathbf{S}^+$ for the spectra in **S** separately and by using the approximations of the chromatograms after one cycle of the backfolding algorithm (**A**) as a window:

$$f_i = (\mathbf{A}_i)^+ \mathbf{D}_u (\mathbf{S}_i)^+ \quad (6)$$

In Eqn (6) the scalar f_i represents the quantitative factor for component *i*. The chromatograms after one cycle of backfolding (**A**) were chosen because they combine enhanced chromatographic resolution with peak profiles that resemble the original ones.

EXPERIMENTAL

Simulations (data sets D_1 and D_2) were performed with the computer program described previously.[14] This program produces realistic data from sample characteristics (concentrations, library spectra, chromatographic profiles) and the operation conditions of the GC (column bleeding) and the MS (scan characteristics and data acquisition parameters).

Data set D_1 (Table 1) contains four n-alkanes with different peak widths. It serves to show that backfolding as such is not quantitative (see Appendix). D_2 (Table 2, Fig. 1) consists of five simulation experiments in which the chromatographic resolution between four components was varied. To check the sensitivity of the algorithm, each simulation experiment was repeated 10 times. This will produce slightly different data sets, because the simulation program uses a random generator to calculate various contributions to the noise.[14] The data sets of D_2 were created to study the effect of increasing chromatographic overlap for components with similar spectra.

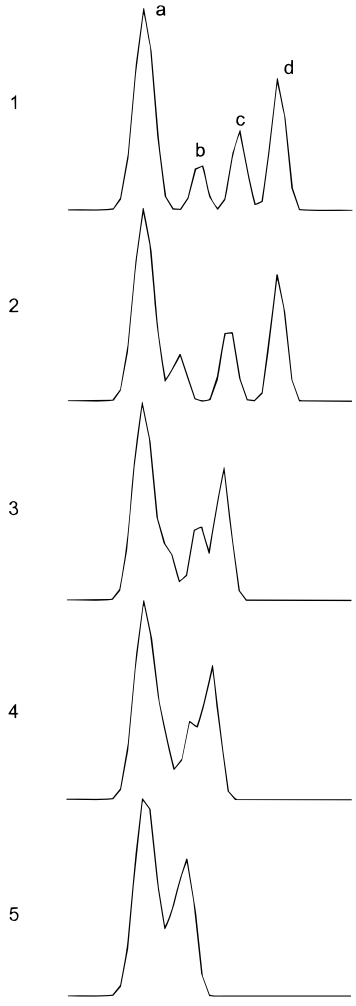

Figure 1. TIC traces of five simulation experiments in D_2 (Table 2).

Two series of GC/MS measurements (D_3 and D_4) were performed on an HP Series II gas chromatograph (capillary column CP Sil 5 CB, 25 m × 0.32 mm, film thickness 0.12 μm) coupled to a VG Autospec Ultima mass spectrometer. The first series of measurements (D_3) consisted of four samples (12.5, 25, 50 and 100 ng μl^{-1}), each containing six components in equal concentrations. These samples were measured (mass range

Table 1. Parameters used in simulation experiment D_1

n-Alkane carbon number	Retention time (s)	Quantity (ng)	Peak width at half-height 2σ (s)
29	15	10	2
30	50	10	3
31	90	10	4
32	130	10	5

Table 2. Parameters used in simulation experiments D_2 with components *n*-hentriacontane (a), 22*S* 17α,21β(*H*)-homohopane (b), gammacerane (c) and 22*R* 17α,21β(*H*)-homohopane (d)

Component	Quantity (ng)	Peak width at half-height (s)	Retention time in data set (s) 1	2	3	4	5
a	6.0	4	18	18	18	18	18
b	1.2	3	29	25	23	22	20
c	1.8	3	37	35	29	27	24
d	3.0	3	45	45	34	31	27

500–50 Da, cycle time 1.4 s) to test the linearity of the calculated signals with respect to the signals in the raw data.

The six samples of series D_4 were analysed (mass range 800–50 Da, cycle time 1.6 s) previously in our laboratory to study the effect of maturation on the release of gammacerane from high-molecular-weight sedimentary matter.[15] The data sets of these samples were subjected to the algorithm described above to compare its application with previous quantitative experiments.

Before the backfolding algorithm was applied, both the simulated and real GC/MS data were unskewed.[13]

RESULTS AND DISCUSSION

Simulated data

In the Appendix it is shown that the ratio of the signals in the raw data over the signals in the backfolded data (A_g/A_b) is a linear function of the peak width (σ). In Fig. 2 it is illustrated that the ratio A_g/A_b for the components in D_1 (□) is close to the theoretical line. The deviation from the theoretical line for narrow peaks results from the small number of mass spectra that can be measured across a mass peak. Quantification with Eqn (6) is almost independent of peak width (○). Note that the variation in peak widths in D_1 is generally not encountered in real samples.

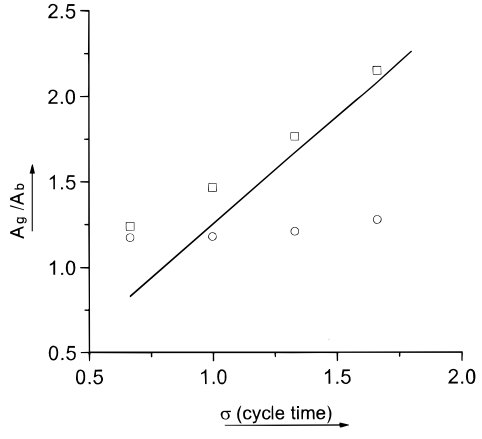

Figure 2. Quantification of components in simulated data set D_1 (Table 1). The ratio of the signal in the raw data (A_g) over that in B_1 (A_b) is given against the peak width parameter σ (□). The ratios obtained with Eqn (6) are also given (○). The line shows the theoretical slope (see Appendix).

The data sets of D_2 (Table 2) contain very similar component spectra. The spectra of the two homohopanes are practically equal; the similarity between the spectra of gammacerane and the homohopanes is high. As illustrated in Fig. 3, this results in distorted chromatographic profiles when Eqn (5) is applied. More than one maximum is found for component 'd' and negative values are obtained for component 'b'. As shown in Table 3, the application of Eqn (6) on these data sets gives, with the exception of data sets 4 and 5, results which are in good agreement with the input of the simulation. When the chromatographic overlap increases, quantification of the signals is less reliable: the average values of the replicate experiments deviate more from the input used in the simulation and the standard deviations increase. In particular, the quantification of the homohopane spectra, which are practically identical, becomes difficult as soon as their peaks overlap. In that case they are overestimated with respect to the other components.

Real GC/MS data

Four samples were analysed (D_3) to study the linearity of Eqn (6) with respect to the height of the signals in the measured data. As shown in Fig. 4, the quantitative factors obtained with Eqn (6) correlate well with the signal in the raw data. The slope of the regression line is smaller than one, because the chromatograms in the backfolded data (**A**), used in Eqn (6) as an envelope,

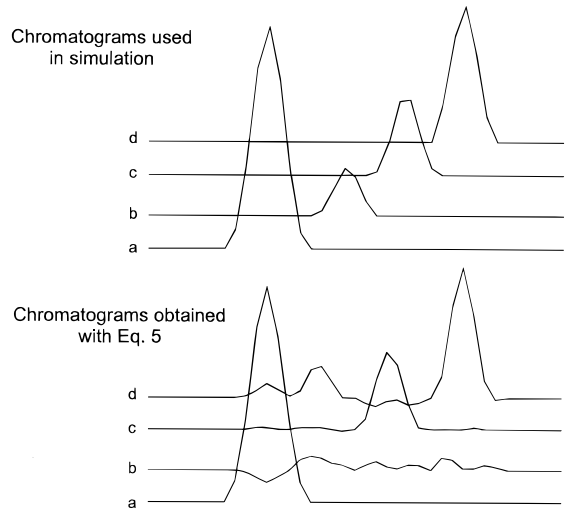

Figure 3. Chromatograms used in simulation of first data set of D_2 (Table 2) compared with those obtained by application of Eqn (5).

Table 3. Analysis of simulation experiments D_2. Comparison of simulated versus calculated compositions, normalized to 100%. The standard deviation is given in parentheses

		Results of backfolding for data set				
Component	Simulation	1	2	3	4	5
a	51.0 (0.1)	51.5 (0.1)	50.0 (0.1)	49.8 (0.3)	46.6 (0.9)	42.7 (1.2)
b	8.8 (0.1)	9.0 (0.0)	8.2 (0.1)	9.4 (0.3)	6.2 (0.2)	12.9 (1.6)
c	15.0 (0.1)	13.0 (0.1)	15.3 (0.1)	15.5 (0.5)	15.7 (1.9)	13.7 (1.3)
d	25.2 (0.1)	26.5 (0.1)	26.5 (0.1)	25.3 (0.3)	31.5 (1.0)	30.8 (0.7)

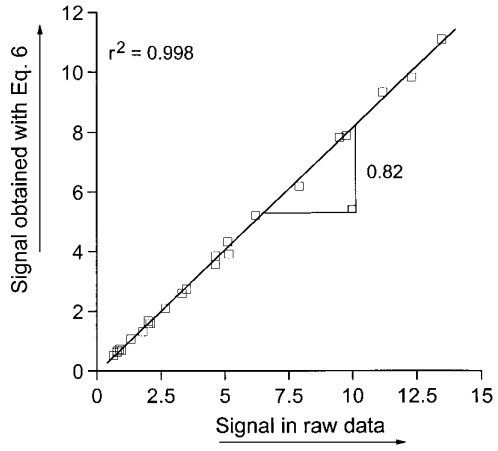

Figure 4. Signals obtained with Eqn (6) as a function of those in the raw data for six components in each of the four samples of D_3.

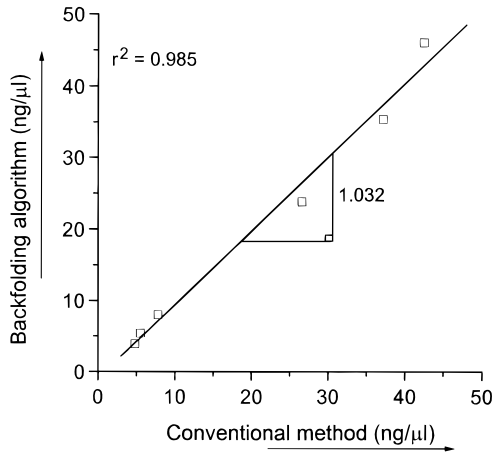

Figure 5. Quantification of gammacerane in six maturation experiments of the same sediment (D_4). Results obtained by application of Eqn (6) are compared with results obtained conventionally.[15]

have smaller peak widths than those in the raw data (**C**). As it should be, this slope is practically equal to the reciprocal values of the ratios A_g/A_b found with Eqn (6) for D_1 (Fig. 2).

The quantification by Eqn (6) was also compared with previous measurements in our laboratory[15] (D_4). On the chromatographic column, gammacerane is coeluting with the two stereoisomers of 17α,21β(*H*)-homohopane, exactly as in the simulation experiments (D_2) described above. Conventionally, quantification was performed using both an internal standard and a C_{30} hopane, which has a very similar spectrum to that of gammacerane and is well separated on the column. Using the backfolding algorithm, it was possible to quantify gammacerane directly on the signal of the internal standard. The response factors of gammacerane and the internal standard were determined with a standard solution of 20 ng μl^{-1} per component. In Fig. 5 it can be seen that the concentrations obtained with the backfolding algorithm are very close to the concentrations obtained with the conventional method. However, using the backfolding algorithm, fewer steps are required for quantification and as a consequence the reliability will increase.

CONCLUSION

The backfolding algorithm is extended with a procedure that enables quantification of the signals of detected components. This procedure is straightforward and easy to implement in the deconvolution method. The results presented in this paper show that reliable quantitative data can be obtained.

REFERENCES

1. J. E. Biller and K. Biemann, *Anal. Lett.* **7**, 515 (1974).
2. F. J. Knorr and J. H. Futrell, *Anal. Chem.* **51**, 1236 (1979).
3. S. L. Neal, E. R. Davidson and I. M. Warner, *Anal. Chem.* **62**, 658 (1990).
4. M. J. Fay, A. Proctor, D. P. Hoffmann and D. M. Hercules, *Anal. Chem.* **63**, 1058 (1991)
5. E. J. Karjalainen, in *Scientific Computing and Automation*, edited by E. J. Karjalainen, p. 477. Elsevier, Amsterdam (1990).
6. E. J. Karjalainen and U. P. Karjalainen, *Anal. Chim. Acta* **250**, 169 (1991).
7. T. A. Lee, L. M. Headley and J. K. Hardy, *Anal. Chem.* **63**, 357 (1991).
8. B. N. Colby, *J. Am. Soc. Mass Spectrom.* **3**, 558 (1992).
9. W. G. Pool, B. van de Graaf and J. W. de Leeuw, *J. Mass Spectrom.* **31**, 509 (1996).
10. W. G. Pool, B. van de Graaf and J. W. de Leeuw, *J. Mass Spectrom.* **32**, 438 (1997).
11. A. Ghosh and R. J. Anderegg, *Anal. Chem.* **61**, 73 (1989).
12. A. Ghosh and R. J. Anderegg, *Anal. Chem.* **61**, 2118 (1989).
13. W. G. Pool, B. van de Graaf and J. W. de Leeuw, *J. Mass Spectrom.* **31**, 213 (1996).
14. W. G. Pool, B. van de Graaf and J. W. de Leeuw, *Comput. Chem.* **16**, 295 (1992).
15. J. S. Sinninghe Damsté, F. Kenig, M. P. Koopmans, J. Köster, S. Schouten, J. M. Hayes and J. W. de Leeuw, *Geochim. Cosmochim. Acta* **59**, 1895 (1995).

APPENDIX

If a chromatographic peak is described by a Gaussian function

$$y = \frac{c}{\sigma\sqrt{2\lambda(\pi)}} \exp\left[-\frac{1}{2}\left(\frac{x}{\sigma}\right)^2\right] \quad (7)$$

where c is quantity, σ is half the peak width at half-height (s) and x is time (s), then the area enclosed by the Gaussian profile, A_g, is c.

The area obtained with one pass of the backfolding algorithm, A_b, can be approximated by

$$A_b = 2\int_0^\infty y'\,dx$$
$$= \frac{2c}{\sigma\sqrt{(2\pi)}}\int_0^\infty \frac{x}{\sigma^2}\exp\left[-\frac{1}{2}\left(\frac{x}{\sigma}\right)^2\right]dx \qquad (8)$$

This gives

$$A_b = \frac{2c}{\sigma\sqrt{(2\pi)}} \qquad (9)$$

Combination of Eqns (7) and (9) results in

$$\frac{A_g}{A_b} = \frac{\sigma\sqrt{\pi}}{\sqrt{2}} \qquad (10)$$

In practice σ should be expressed in the cycle time of the mass spectrometer, because the differentiated signal is obtained numerically and not analytically as implicated in Eqn (8).