JOURNAL OF MASS SPECTROMETRY, VOL. 32, 438–443 (1997)

Automated Extraction of Pure Mass Spectra from Gas Chromatographic/Mass Spectrometric Data†

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

An algorithm is described that extracts pure mass spectra from gas chromatographic/mass spectrometric (GC/MS) data. It is based on backfolding, a method described previously to enhance chromatographic resolution in GC/MS data. The ability to extract pure mass spectra was evaluated with both simulated and real GC/MS data and the algorithm was compared with two other methods described recently. It is shown that the algorithm presented gives good results, even when the chromatographic resolution is poor and the spectra are very similar. No *a priori* knowledge concerning the composition of the data is required. © **1997 by John Wiley & Sons, Ltd.**

J. Mass Spectrom. **32**, 438–443 (1997)
No. of Figures: 3 No. of Tables: 3 No. of Refs: 24

KEYWORDS: gas chromatography/mass spectrometry; deconvolution; spectrum clean-up; backfolding

INTRODUCTION

Samples are analyzed using gas chromatography/mass spectrometry (GC/MS) to identify and, if desired, to quantify individual components. When a complex mixture is analyzed the identification is often hampered by insufficient chromatographic resolution and/or by a relatively high background. A number of authors have reported on methods to obtain pure spectra from GC/MS data. These methods are based on comparison with library spectra,[1–5] regression,[6–8] principal component analysis[9–13] and reconstruction from mass chromatograms.[14–17] In one of these methods, called differential GC/MS,[18,19] ion abundances in a scan are subtracted from those in the subsequent scan, mass by mass. Positive and negative results are stored in separate sets of differentiated data. Backfolding is an algorithm in which these separate sets of differential data are recombined.[20] In the data thus obtained, background is eliminated and the chromatographic resolution is improved. This two-step algorithm, differentiation followed by recombination, can be repeated several times.

The effect of backfolding on the chromatography is shown in Fig. 1. Starting from the simulated GC/MS data[21] at the top, the backfolding algorithm converges in six cycles. The chromatographic resolution has increased significantly. In this paper, an algorithm is

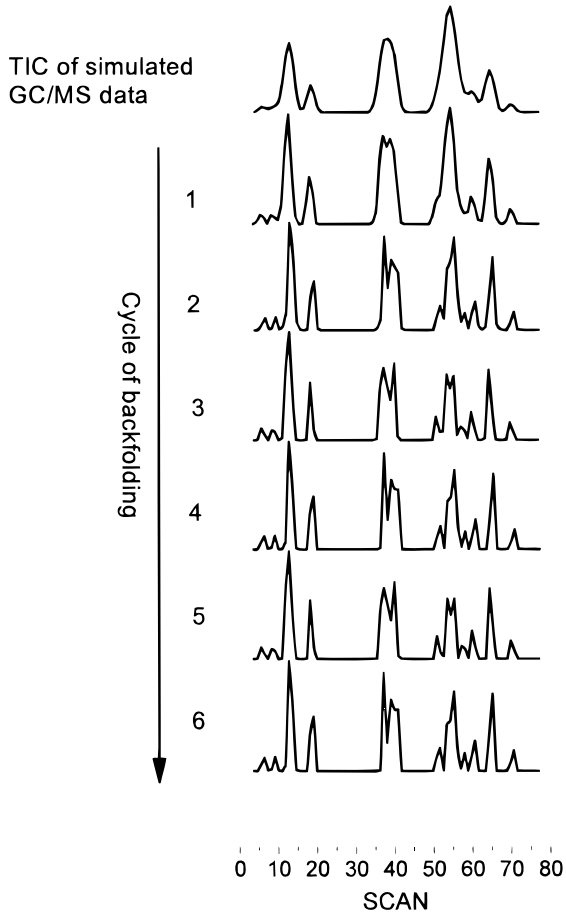

Figure 1. The backfolding process improves chromatographic resolution. When repeated it converges rapidly.

† NIOZ Contribution No. 3014.
* Correspondence to: W.G. Pool.

CCC 1076–5174/97/040438–06 $17.50
© 1997 by John Wiley & Sons, Ltd.

Received 4 July 1996
Accepted 6 January 1997

presented which permits the automatic extraction of pure spectra from the backfolded GC/MS data without *a priori* knowledge of the sample that is analyzed. The algorithm described was tested and compared with two other proposed deconvolution methods[8,17] using both simulated and real GC/MS data.

THEORY AND DESCRIPTION OF THE ALGORITHM

The data matrix **D** contains the raw GC/MS data, scans represented by rows, masses by columns. The intensities in **D** are first corrected (unskewed)[22] for the changes in concentration during the measurements of each mass spectrum:

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

The unskewed data can be represented as a product of three matrices:

$$\mathbf{D_u} = \mathbf{CFS} \tag{2}$$

where **C** is a matrix containing pure component chromatograms in its columns normalized to unit area and **S** contains component spectra in its rows normalized to unit intensity. **F** is a diagonal matrix with factors that quantify each component. When backfolding is applied, new matrices are formed:

$$\mathbf{D_u} \rightarrow \mathbf{B_1} \rightarrow \mathbf{B_2} \rightarrow \rightarrow \rightarrow \mathbf{B_n} \tag{3}$$

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

$$\mathbf{B_n} = \mathbf{C_n F_n S} \tag{4}$$

where $\mathbf{C_n}$ and $\mathbf{F_n}$ are chromatograms and quantitative factors, respectively, after n cycles of the backfolding process. Component spectra are not influenced by backfolding, therefore **S** is not indexed. The matrix product $\mathbf{F_1 S}$ is extracted from the data by use of Eqn (4) with $n = 1$ after a suitable approximation **A** for $\mathbf{C_1}$ has been constructed:

$$\mathbf{F_1 S} = (\mathbf{A^T A})^{-1} \mathbf{A^T B_1} \tag{5}$$

The pure mass spectra (**S**) are obtained by normalization of $\mathbf{F_1 S}$. Equation (4) is preferred over Eqn (2) because in $\mathbf{B_1}$ background is eliminated and the chromatographic resolution has been enhanced.[20] Equation (4) with $n = 1$ is used because at this stage of the backfolding process peak shapes can still be approximated with Gaussian profiles and satellite peaks are insignificant.[20]

The construction of **A** proceeds in two stages: (i) detection of components and (ii) construction of a chromatogram for each of the detected components.

Detection of components

The algorithm for component detection uses the properties (see Appendix) of the matrices $\mathbf{B_n}$ at convergence: (a) the spectrum of each component is present in two rows of $\mathbf{B_{n-1}}$ and in two rows of $\mathbf{B_n}$; (b) the backfolding process is alternating ($\mathbf{B_n} = \mathbf{B_{n-2}}$ and $\mathbf{B_{n+1}} = \mathbf{B_{n-1}}$); and (c) either in $\mathbf{B_n}$ or in $\mathbf{B_{n-1}}$ the ratio of the intensities in two successive spectra exceeds $(1 + \sqrt{2})$.

Apparently, the information in $\mathbf{B_n}$ and $\mathbf{B_{n-1}}$ is redundant and a data reduction of 75% is possible without loss of information. This data reduction is carried out by looking in $\mathbf{B_n}$ and $\mathbf{B_{n-1}}$ for non-zero elements $b_{i,j}$ and $b_{(i+1),j}$ whose ratio exceeds $(1 + \sqrt{2})$. Of these elements only the highest values are saved and all other elements are zeroed. After the data reduction $\mathbf{B_n}$ and $\mathbf{B_{n-1}}$ contain complementary information. These matrices are combined in **R**, which has twice as many rows as $\mathbf{B_n}$. When in a row of **R** both the sum of the elements (representing the total ion current) and the number of non-zero elements (mass peaks) exceeds set minimum values, a component is considered to be present.

Construction of individual chromatograms

Rows in **R** represent the first approximation of the component spectra. In the absence of chromatographic overlap these approximations are very good.[20] The spectra are inspected for m/z values that are unique *in the time window* of the chromatographic peak in $\mathbf{B_1}$. This time window, known from the applied shift in de the backfolding process (see Appendix), equals the width at the base of the peak in $\mathbf{B_1}$. The sum of intensities of the selected m/z values across the time window of the component in $\mathbf{B_1}$ gives the chromatographic profile of the component. The profile is then characterized by the position and the width at half-height of the Gaussian peak that is fitted through its three most intense points. At this stage further data reduction is sometimes necessary: (a) components with a very large peak width are considered to be background and are therefore deleted; and (b) components for which the peak profiles almost completely overlap (i.e. peak maxima less then 0.5 scan apart) are combined in one spectrum. The peak characteristics of the remaining components are used to construct **A** column by column.

EXPERIMENTAL

The algorithm described above is carried out by a program written in Pascal (available from the authors on request). It has been tested on and compared with other deconvolution methods on three sets of GC/MS data. One set (D_1), which contains 15 components in various quantities, is obtained via simulation.[21] The total ion current (TIC) of this data set is given in Fig. 1. The two other sets (D_2 and D_3) represent measurements of sediment extracts containing alkanes and hopanes. For both samples the part of the measurements that contains the signals of n-hentriacontane, (22S)- and (22R)-17α,21β(H)-homohopanes and gammacerane (see Fig. 2 for structures) was selected. Although both samples contain the same components, this is not directly clear from the TIC traces (see top traces in Fig. 2). The differences in chromatographic resolution are major; there is a time difference of 6 months between

© 1997 by John Wiley & Sons, Ltd.

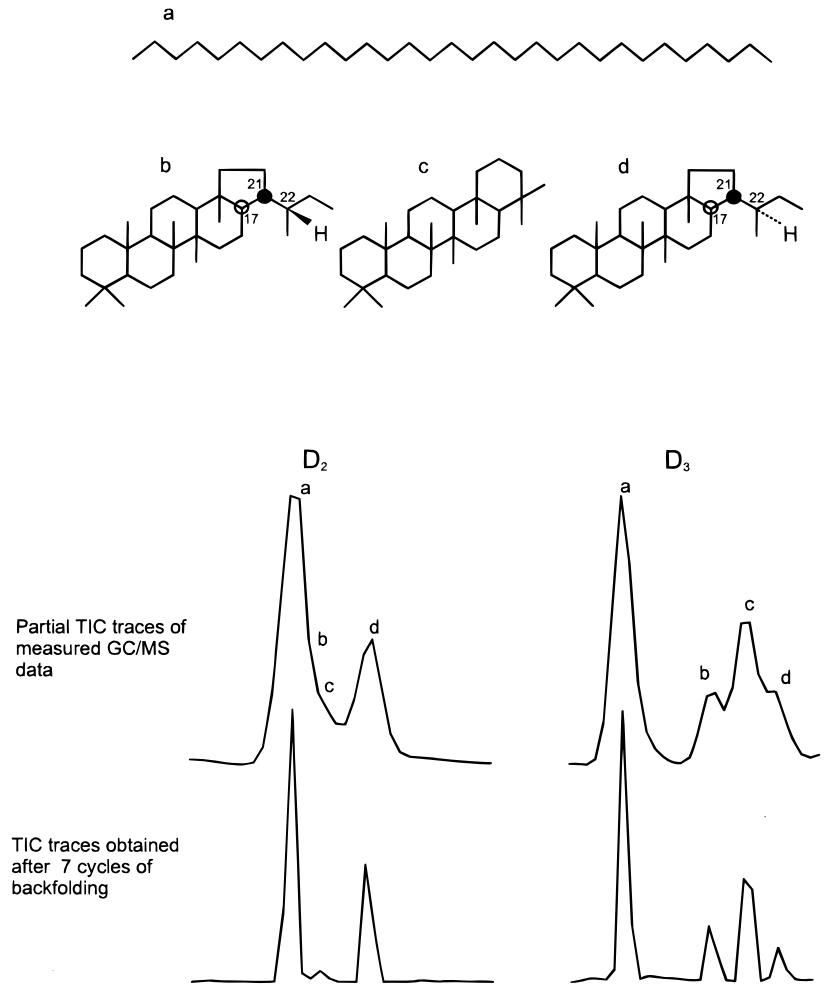

Figure 2. Partial TIC traces of two samples measured by GC/MS before (top trace) and after backfolding (bottom trace). The positions of n-hentriacontane, (22S)-17α,21β(H)-homohopane, gammacerane and (22R)-17α,21β(H)-homohopane are indicated as a, b, c and d, respectively. The structures of the components are shown at the top.

the measurements of the samples and during that time another column was installed.

All three data sets contain 70 eV spectra from 800 to 50 Da with a cycle time of 1.6 s at a resolution of 1000. D_1 was simulated with the computer program described previously.[21] This program produces realistic data from sample characteristics (concentrations, library spectra, chromatographic profiles) and the operating conditions of the gas chromatography (column bleeding) and the mass spectrometer (scan characteristics and data acquisition parameters). D_2 and D_3 were measured on a VG Autospec Ultima mass spectrometer coupled to an HP Series II gas chromatograph equipped with a fused-silica capillary column coated with CP Sil 5 CB.

Computer code was also written for two other deconvolution methods with which the algorithm described in this paper was compared. The first of these is that described by Colby.[17] This method is an extension of the Biller–Biemann[15] procedure and calculates peak centroids for all mass chromatograms. The intensities of the peak centroids that fall within a window of 0.1 scan are summed to form a deconvoluted TIC (DTIC) trace. The maxima in this trace are used to extract the spectra.[17] The second method, advocated by Karjalainen,[8] is alternating regression (AR). This iterative method starts with a spectrum matrix filled with random positive numbers. The generalized inverse of this matrix is multiplied with the data matrix to obtain a chromatogram matrix. After data reduction to positive unimodal chromatographic profiles the generalized inverse of this matrix is premultiplied with the data matrix to obtain an improved spectrum matrix. Negative intensities are removed from this matrix after which it is used again to calculate a chromatogram matrix. This process is repeated until it converges.[8]

The deconvolution methods that were tested all suffer from skewing. Therefore, both simulated and measured GC/MS data were unskewed[22] before they were subjected to a deconvolution method.

The quality of the spectra obtained was evaluated using a similarity index (SI), which is calculated by

$$SI = 999 \times \frac{\left(\sum_j \{m_j \sqrt{ra_j\, pa_j}\}\right)^2}{\sum_j (m_j ra_j) \sum_j (m_j pa_j)} \qquad (6)$$

where ra_j is the relative abundance of mass j in the spectrum obtained, pa_j is the relative abundance of mass j of the library spectrum and m is the mass. A similarity of

© 1997 by John Wiley & Sons, Ltd.

Table 1. Averaged similarity indices and maximum differences in retention time obtained with the deconvolution methods on simulated data (Fig. 1)

Calculation of spectra using	Average similarity index	Maximum difference in retention time (s)
Backfolded data	940	0.04
Peak centroids	899	0.03
AR	889	0.12

zero indicates that no common peaks were observed and 999 indicates totally identical spectra.

RESULTS AND DISCUSSION

The simulated GC/MS data (D_1; Fig. 1) provided the opportunity to test the computer codes. It also served as a first test on the methods in this paper. Note that, com-

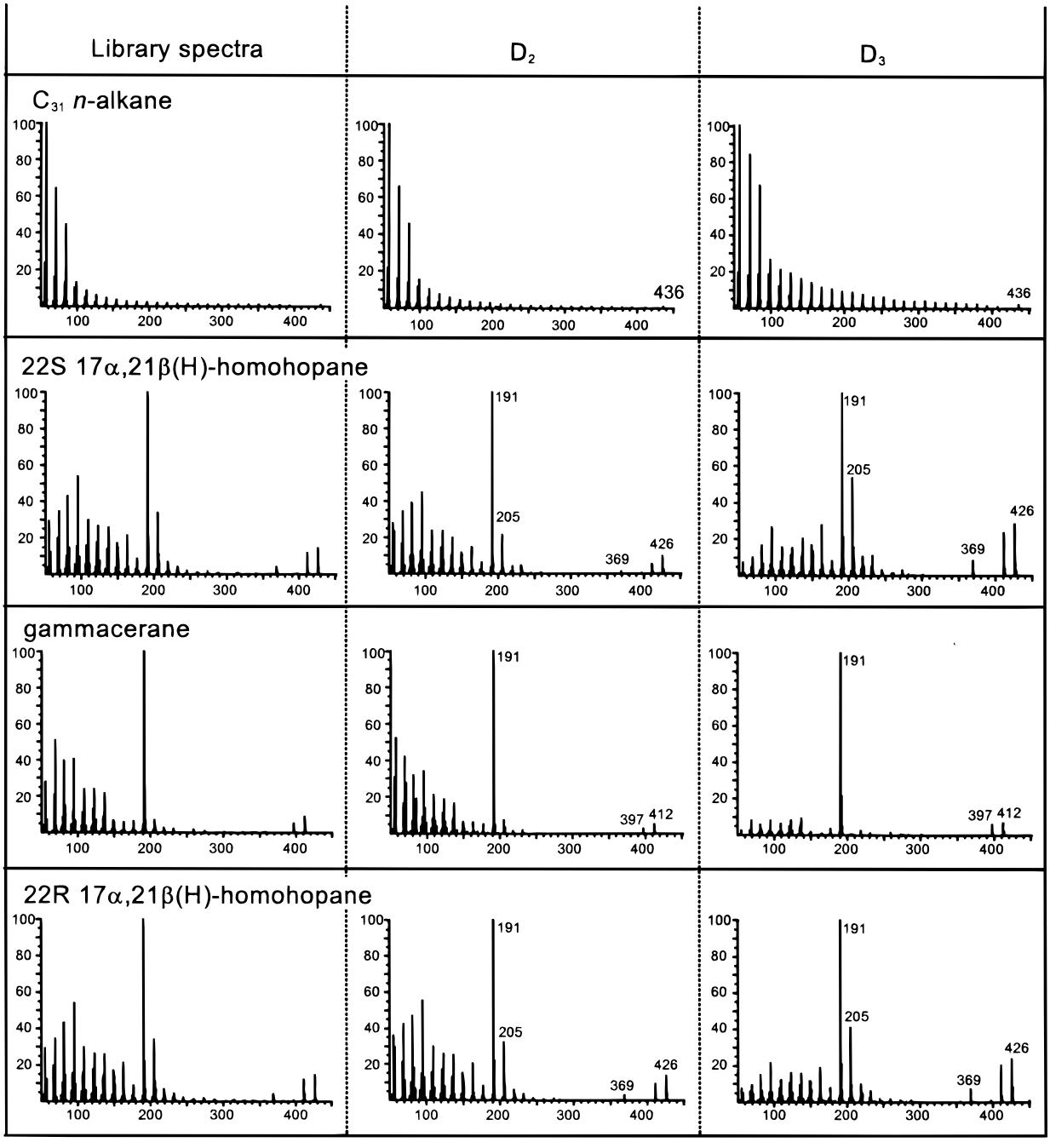

Figure 3. Library spectra of the four components present in the selected part of the TIC traces of the samples (left column) and the spectra obtained from D_2 (middle column) and D_3 (right column) with the algorithm described.

© 1997 by John Wiley & Sons, Ltd.

Table 2. Similarity indices calculated for the spectra (Fig. 3) obtained from D_2 and D_3 (Fig. 2)

Data set	Calculation of spectra using	C_{31} *n*-alkane	(22*S*)-Homohopane	Gammacerane	(22*R*)-Homohopane
D_2	Backfolded data	964	853	515	969
	Peak centroids	950	355	327	945
	AR	681	136	867	959
D_3	Backfolded data	887	894	737	900
	Peak centroids	891	832	888	179
	AR	890	844	876	514

pared with measured data, simulated data are eminently suited to test the accuracies in obtaining chromatographic parameters such as retention time. In Table 1 the average SI of the detected components is shown, together with the maximum difference between the calculated retention times and the retention times used in the simulation. The results of the algorithm described in this paper are very close to those obtained by the method based on peak centroids by Colby.[17]

Data sets D_2 and D_3 present a more stringent test for any deconvolution method. In fact these samples were chosen as a worst case scenario: the spectra of three of the components present in the data are very similar and there is a severe chromatographic overlap.

Backfolding applied to D_2 and D_3 results in new TICs as shown at the bottom trace of Fig. 2. The spectra obtained with the algorithm described in this paper are shown in Fig. 3. A numerical comparison of the results compared with the other deconvolution methods is given in Table 2. From both Fig. 3 and Table 2 it is clear that our algorithm produces good quality spectra. For both data sets, all four components are detected and the spectra calculated can be used to identify the components. The relative low similarity index (Table 2) obtained for gammacerane in D_2 is caused by the presence of fragments of n-alkanes. However, the mass peaks generally used to identify this component (*m/z* 412 and 397)[23] are clearly present. The data in Table 2 show that the peak centroids method of Colby[17] does not generate useful spectra of (22S)-17α, 21β(H)-homohopane and gammacerane in D_2 and of (22R)-17α,21β(H)-homohopane in D_3. Apparently this method suffers more from chromatographic overlap and the similarity in the spectra than the algorithm described in this paper.

The performance of AR is not better than that of the peak centroids method. This is not directly seen from the data in Table 2, but an inspection of the mass spectra shows that mass peak intensity ratios, important in the differentiation of the stereoisomers of the homohopanes (*m/z* 191 and 205),[23] are not well reproduced. Both the algorithm described in this paper and the peak centroids method do much better here. It should be noted that AR has some additional disadvantages. At the start, the number of components present in the data is needed as input. When this number is set too high, the algorithm converges slowly and is not reproducible. When the number of components is set too small, convergence is rapid, but the fits remain poor.[24] For D_2 and D_3 the number of components was correctly set to four. However, it was found that the results, especially for D_2, were not reproducible. In three out of the four times that AR was applied to this data set, the spectrum of the second component was completely empty.

CONCLUSIONS

The algorithm described in this paper to extract pure spectra from GC/MS data produces good quality spectra that can be used to identify the components in the sample. The method is able to detect components with very similar spectra, even at low chromatographic resolution. No a priori knowledge concerning the composition of the sample is necessary. The method compares favourably with other deconvolution routines.

REFERENCES

1. E. Jellum, O. Stokke and L. Eldjarn, *Anal. Chem.* **45**, 1099 (1973).
2. R. Reimendal and J. Sjövall, B., *Anal. Chem.* **45**, 1083 (1973).
3. C. C. Sweeley, N. D. Young, J. F. Holland and S. Gates, *J. Chromatogr.* **99**, 507 (1974).
4. H. Nau and K. Biemann, *Anal. Chem.* **46**, 426 (1974).
5. B. E. Blaisdell and C. C. Sweeley, *Anal. Chim. Acta* **117**, 1 (1980).
6. F. J. Knorr, H. R. Thorsheim and J. M. Harris, *Anal. Chem.* **53**, 821 (1981).
7. B. Vandeginste, W. Derks and G. Kateman, *Anal. Chim. Acta* **173**, 253 (1985).
8. E. J. Karjalainen, in *Scientific Computing and Automation*, edited by E. J. Karjalainen, p. 477. Elsevier, Amsterdam (1990).
9. G. L. Ritter, S. R. Lowry and T. L. Isenhour, *Anal. Chem.* **48**, 591 (1976).
10. E. R. Malinowsky, *Anal. Chem.* **49**, 612 (1977).
11. F. J. Knorr and J. H. Futrell, *Anal. Chem.* **51**, 1236 (1979).
12. E. R. Malinowsky, *Anal. Chim. Acta* **134**, 129 (1982).
13. L. Roach and M. Guilhaus, *Org. Mass Spectrom.* **27**, 1071 (1992).
14. R. A. Hites and K. Biemann, *Anal. Chem.* **42**, 855 (1970).
15. J. E. Biller and K. Biemann, *Anal. Lett.* **7**, 515 (1974).
16. R. G. Dromey, M. J. Stefik, T. C. Rindfleisch and A. M. Duffield, *Anal. Chem.* **48**, 1368 (1976).
17. B. N. Colby, *J. Am. Soc. Mass Spectrom.* **3**, 558 (1992).
18. A. Ghosh and R. J. Anderegg, *Anal. Chem.* **61**, 73 (1989).

© 1997 by John Wiley & Sons, Ltd.

19. A. Ghosh and R. J. Anderegg, *Anal. Chem.* **61**, 2118 (1989).
20. W. G. Pool, B. van de Graaf and J. d. Leeuw, *J. Mass Spectrom.* **31**, 509 (1996).
21. W. G. Pool, B. van de Graaf and J. W. de Leeuw, *Comput. Chem.* **16**, 295 (1992).
22. W. G. Pool, B. van de Graaf and J. W. de Leeuw, *J. Mass Spectrom.* **31**, 213 (1996).
23. K. E. Peters and J. M. Moldowan, *The Biomarker Guide; Interpreting Molecular Fossils in Petroleum and Ancient Sediments*. Prentice Hall, Englewood Cliffs, NJ (1993).
24. E. J. Karjalainen and U. P. Karjalainen, *Anal. Chim. Acta* **250**, 169 (1991).

APPENDIX

The first cycle of the backfolding algorithm can be depicted as

$$\mathbf{D_u} \xrightarrow{\text{differentiation}} \begin{matrix}\mathbf{U}\\ \mathbf{V}\end{matrix} \xrightarrow{\text{shift and recombine}} \mathbf{B_1} \quad \text{(A1)}$$

where **U** is the matrix containing the positive differential signal and **V** the matrix containing the absolute values of the negative differential signal. The elements of $\mathbf{B_1}$ are given by

$$b_{i,j} = u_{(i-s),j} + v_{(i+s),j} \quad \text{(A2)}$$

where s is the applied shift that is calculated from the chromatographic peak widths. The backfolding algorithm results in a decrease in peak width and, therefore, leads to a corresponding reduction in the shift applied in the next cycle. In our application the minimum shift that is allowed ($s = 1$) is generally reached after five or six cycles of backfolding. At this stage most of the signal of a component is concentrated in two rows of $\mathbf{B_n}$ and the backfolding algorithm becomes alternating:

$$\mathbf{B_n} = \mathbf{B_{n-2}} \quad \text{and} \quad \mathbf{B_{n-1}} = \mathbf{B_{n-3}} \quad \text{(A3)}$$

In Table A1 a numerical example is given for one column (m/z value) in $\mathbf{B_{n-1}}$ and $\mathbf{B_n}$ at convergence. The data in this table show that, apart from a constant factor two and a shift by one row, $\mathbf{B_{n-1}}$ is reproduced when $\mathbf{B_n}$ is subjected to another cycle of backfolding.

When in $\mathbf{B_n}$ two successive intensity values for a mass are x and y ($y < x$), then the values in $\mathbf{B_{n+1}}$ ($=\mathbf{B_{n-1}}$) will be $x + y$ and $x - y$. The ratio y/x equals $(x - y)/(x + y)$ only when

$$\frac{x}{y} = \frac{x+y}{x-y} \Rightarrow x^2 - 2xy - y^2 = 0$$

$$\Rightarrow x = y(1+\sqrt{2}), \quad x = y(1-\sqrt{2}) \quad \text{(A4)}$$

Since negative intensities do not exist, only the first solution is applicable. This result means that either in $\mathbf{B_{n-1}}$ or in $\mathbf{B_n}$ the ratio of the intensities in two successive scans exceeds $(1 + \sqrt{2})$.

Table A1. The backfolding algorithm converges to matrices wherein component spectra are present in two rows only; the algorithm then becomes alternating: each matrix contains the sum and differences of the values in the previous matrix

Cycle of Backfolding	Row i	Data b	Differential GC/MS u	v	Shift with $s = 1$ u^s	v^s	Summation $u^s + v^s$
$n-1$	1	0	0	0	0	0	0
	2	0	80	0	0	0	0
	3	80	40	0	80	120	200
	4	120	0	120	40	0	40
	5	0	0	0	0	0	0
	6	0	0	0	0	0	0
n	1	0	0	0	0	0	0
	2	0	200	0	0	160	160
	3	200	0	160	200	40	240
	4	40	0	40	0	0	0
	5	0	0	0	0	0	0
	6	0	0	0	0	0	0

© 1997 by John Wiley & Sons, Ltd.