Journal of Chromatography, 112 (1975) 573–580
© Elsevier Scientific Publishing Company, Amsterdam — Printed in The Netherlands

CHROM. 8505

THE DEVELOPMENT OF A COMPUTER-ASSISTED SEARCH FOR ANOMALOUS COMPOUNDS (CASAC)

E. JELLUM, P. HELLAND and L. ELDJARN
Institute of Clinical Biochemistry, Rikshospitalet, University of Oslo, Oslo (Norway)
and
U. MARKWARDT and J. MARHÖFER
Varian-MAT GmbH, Bremen (G.F.R.)

SUMMARY

An automated system for the recognition of anomalies in multi-compound mixtures is described. The mixtures to be investigated are injected into a combined gas chromatograph–mass spectrometer and low resolution spectra are acquired by repetitive scanning. Using an on-line computer, all of these spectra are compared with a pre-recorded file of spectra obtained by identical analyses of a "normal" mixture. The matching procedure has been designed to allow for differences in retention times. The CASAC program calculates and plots the degree of coincidence and in this way determines whether the sample spectrum contains fewer or more fragments than the corresponding library spectrum. The system has been applied to studies on pathological urine samples.

INTRODUCTION

There is at present a rapidly growing interest in multi-component analyses of biological fluids and tissues. It has been shown by many workers that such analyses may give detailed information about the metabolic situation in the cells and in the body, and that such knowledge may be of great diagnostic, biochemical and clinical-chemical value (see, *e.g.*, refs. 1 and 2).

The combined gas chromatography–mass spectrometry–computer system (GC–MS–COM) seems at present to be the most powerful tool for carrying out multi-component analyses and such systems are now in operation in many biomedical laboratories throughout the world. The analytical procedures adopted by most of these laboratories include visual evaluation of the various gas chromatograms followed by mass spectral identification of all major or unexpected GC peaks. Although this approach has proved fruitful, particularly if combined with computer identification of the mass spectra, it is nevertheless obvious that much valuable information may be overlooked and lost by this method. This is particularly true if the disorder causes only small changes in the metabolic profiles, or if the abnormal compound is hidden (and therefore neglected) under a normal GC peak.

According to the literature, the main efforts have so far been devoted to obtaining rapid identification of the selected GC peaks, *e.g.*, by library search routines or computer interpretation of mass spectra, and many of the systems now in operation have become extremely powerful. However, our experience, after analyzing 2000–3000 patient samples by GC–MS–COM methods, is that although it still may be very difficult to identify a given GC peak, the major problem is rather the correct evaluation of the GC profiles, *e.g.*, which should be considered "normal" (taking dietary variations, drug intakes, and the many artifacts and pitfalls into consideration) and which are pathological, *i.e.*, contain abnormal metabolites or lack normal metabolites. Although much work remains to be carried out before these problems will be solved, it is encouraging that some laboratories[3–5] are now beginning to emphasize the need for more automated identification of the human metabolites, *e.g.*, as in the recently published method of Sweeley *et al.*[5].

In this paper, we describe the first steps of a joint attempt aimed at producing an automatic procedure for detecting abnormalities in any complex mixture. The system (CASAC, Computer-Assisted Search for Anomalous Compounds) utilizes retention times (indirectly) and is based on computer analyses of mass spectral data recorded by repetitive scanning. These data are matched against a "normal" file made up of spectra recorded from a normal sample (*e.g.*, pooled urine) in an identical manner. Pathological spectra selected by the on-line system may subsequently be identified by using more traditional library search methods and comprehensive (or dedicated) files of reference spectra.

EXPERIMENTAL

Apparatus

The combined GC–MS–COM instrument consisted of a Varian 1400 gas chromatograph fitted with an 8 ft. × 1/4 in. glass column filled with 10% OV-17, a molecular separator of the glass frit type, a Varian CH7 low-resolution mass spectrometer and a SpectroSystem 100 MS consisting of a 12K on-line computing system, with one magnetic tape station and one magnetic disc station (large dual discs, moveable head), and a graphic display unit (Varian-MAT, Bremen, G.F.R.). With this system, the data acquisition, data conversion and the CASAC program were performed. Off-line library searches were performed on a central computer (CDC-6600) via a display terminal and a modem connected to an ordinary telephone network, as described previously[6].

Urine samples

The "normal" pooled urine sample was obtained by mixing freshly voided morning urine from 25 healthy persons of both sexes and of age 10–40 years. Great care was taken that none of these persons had taken any drugs for a long period, and that none of them had unusual dietary habits. The pooled urine was divided into smaller portions and stored frozen at −20°. Abnormal urine samples were selected from our deep-frozen supply of specimens from patients.

Methods

Organic acids were extracted from the urine with diethyl ether. The urine sample (6 ml) was acidified to pH 1–2 with 6 *N* hydrochloric acid and extracted three times with two volumes of freshly distilled diethyl ether. The extracts were combined and dried overnight over anhydrous sodium sulphate. Diazomethane was passed into the ethereal solution and the methyl esters were concentrated under a stream of nitrogen before injection into the gas chromatograph and mass spectrometer. Great care was exercised so that the separation conditions during gas chromatography were as identical as possible from the library run to sample runs.

Mass spectra were also recorded under as identical conditions as possible and by using repetitive scanning (scan time about 2 sec from mass 20 to mass 400; return time 2 sec).

CASAC computer program

The program contains three main routines: one for automatic generation of the "normal" library, one for matching spectra from a sample against this library, and one module for displaying the results of the comparison.

Library generation. After the real-time data collection during the GC–MS run of the normal sample has been completed, the data are mass-converted and stored on the magnetic disc as complete mass spectra. These are subsequently used to generate the "normal" library. The latter procedure may involve reduction of the spectra, *e.g.*, abbreviation to the two largest peaks for every 14 mass units, or three peaks for every 20 mass units, or two peaks for every 10 mass units. The library can also be made up so as to consist of complete spectra, which we used for the work described here, with or without exclusion of certain masses (*e.g.*, 0–40), and one also has the option of setting threshold values so as to eliminate low-intensity peaks (noise and background). In general, this particular program has been designed to be as flexible as possible, so that almost any type of library structure can be produced[7]. The time taken to convert a given set of mass spectral data into any type of library is 10 min per 100 spectra.

Matching procedure. The unknown mixture (*e.g.*, a patient's urine) is first analyzed by GC–MS under conditions identical with those used for the normal sample (same column, same gas flow-rate, same temperature program, same time for starting the repetitive scanning, etc.). The mass spectral data from the patient's sample is then stored on the magnetic disc. Each spectrum of this sample is automatically being reduced to the same format as given for the library, and each reduced spectrum is compared with a set of corresponding library spectra defined by a "window", as illustrated in Fig. 1. The more reproducible the absolute retention time is, the more narrow the limits of the window can be set. Our experience so far indicates that if an unknown spectrum is matched against ±10 scan numbers, as shown in Fig. 1, one can be fairly certain of not missing any compounds.

The equation used to calculate the degree of coincidence (C) is

$$C = \frac{N_{LS}{}^2}{N_S N_L} \cdot 1000$$

where N_{LS} = number of common peaks in the defined mass range, N_L = number of peaks in library spectrum and N_S = number of peaks in sample spectrum.

In order to be able to distinguish between missing and additional compounds

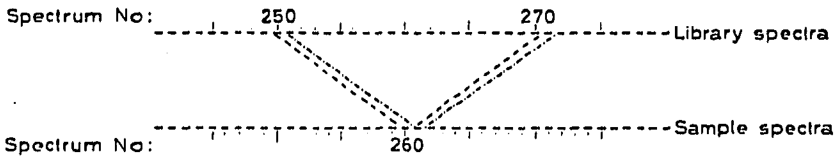

Fig. 1. Principle of the CASAC search. Scan No. 260 from the patient is matched against all library spectra in the region 250–270. The highest degree of coincidence within the "window" is stored and used. Next, scan No. 261 from the sample is compared with the window 251–271 of the library. Comparison and displacement continues in this way until all sample spectra have been matched against the appropriate library windows.

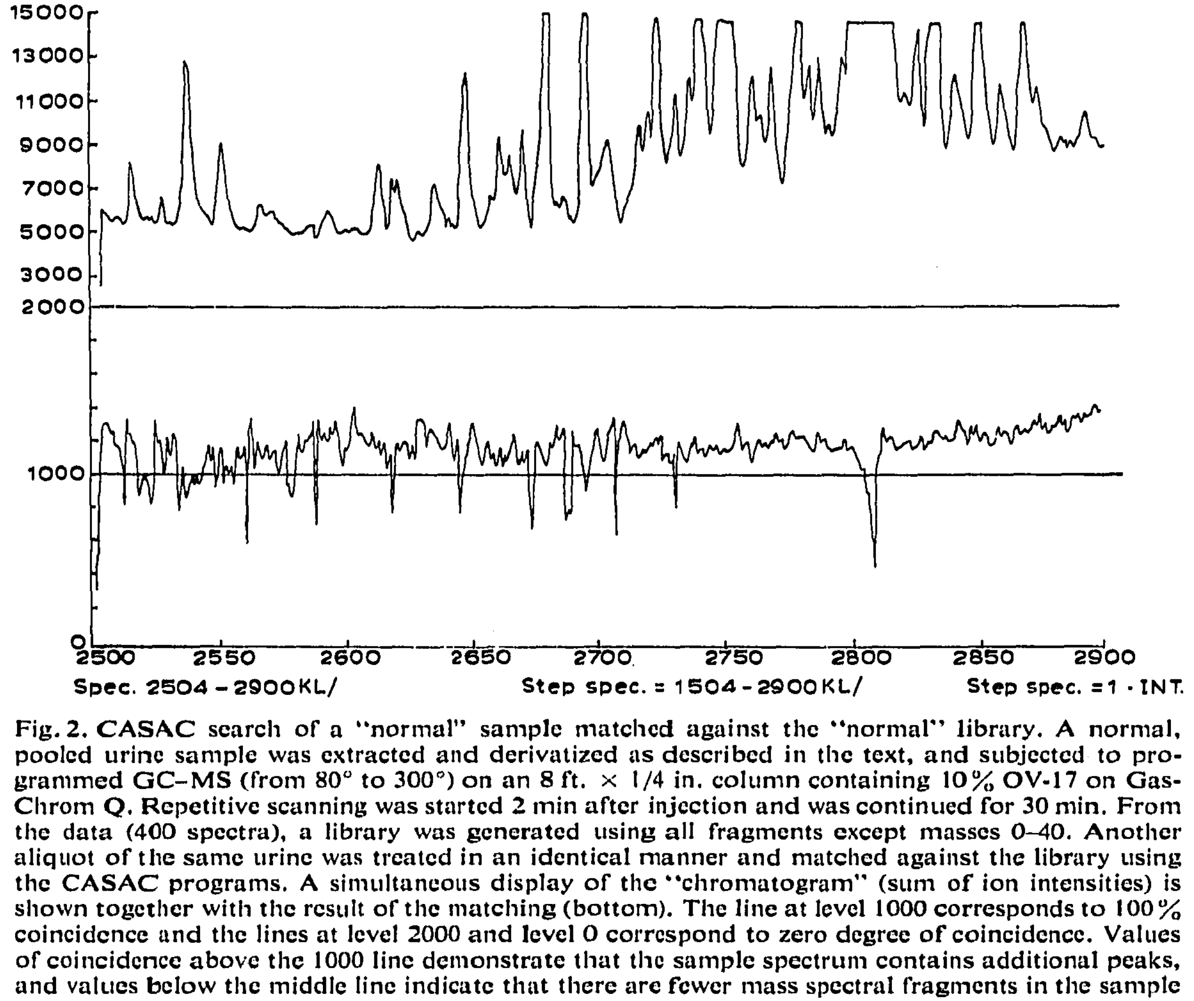

Fig. 2. CASAC search of a "normal" sample matched against the "normal" library. A normal, pooled urine sample was extracted and derivatized as described in the text, and subjected to programmed GC–MS (from 80° to 300°) on an 8 ft. × 1/4 in. column containing 10% OV-17 on Gas-Chrom Q. Repetitive scanning was started 2 min after injection and was continued for 30 min. From the data (400 spectra), a library was generated using all fragments except masses 0–40. Another aliquot of the same urine was treated in an identical manner and matched against the library using the CASAC programs. A simultaneous display of the "chromatogram" (sum of ion intensities) is shown together with the result of the matching (bottom). The line at level 1000 corresponds to 100% coincidence and the lines at level 2000 and level 0 correspond to zero degree of coincidence. Values of coincidence above the 1000 line demonstrate that the sample spectrum contains additional peaks, and values below the middle line indicate that there are fewer mass spectral fragments in the sample than in the library. In this particular demonstration, the overall degree of coincidence was *ca.* 75–80% (ocrresponding to the numbers 750–800 and 1250–1200). The only significant deviation is for some reason under the large hippuric acid peak, and indicates that at this particular position the sample spectra contain fewer fragments than the library.

in the sample, the program also determines a value $X = -1$ for $N_S > N_L$ and $X = +1$ for $N_S < N_L$. Then another coincidence value C^* is calculated:

$$C^* = \begin{cases} C \text{ for } X = +1 \\ 2000 - C \text{ for } X = -1. \end{cases}$$

The maximum value of C^* obtained within the given window is then searched for, selected and added to the sample spectrum as if it was a peak with mass 999 and of intensity given by the value of C^*. By means of this "trick", a mass chromatogram (constructed from the sum of ion intensities) and the coincidence value ("mass" 999) could be plotted simultaneously with existing software as shown in Fig. 2.

A value of C^* in the vicinity of 1000 (= 100%) shows a high degree of coincidence, whereas values closer to zero indicate poor coincidence and that the sample spectrum is lacking peaks that are present in the library spectrum. A value closer to 2000 also demonstrates poor coincidence, but means that the sample spectrum contains more peaks than the library spectrum. The time taken to match sample spectra against the library is 3 min per 100 spectra (for a window size of 20 scan numbers). A typical GC-MS run produces about 400 spectra.

The results of the matching are displayed on an oscilloscope screen. The output may either be as shown in Fig. 2, where the chromatogram and the matching results appear simultaneously, or the chromatogram can be omitted. In both instances the scales can be expanded or compressed at will.

RESULTS

Library structure

The CASAC program has been designed to select regions in complex gas chromatograms where differences exist between sample and "normals". The result of such a comparison depends upon many factors, of which the structure and organization of the "normal" library is one. Many permutations are possible on generation of the library file, and so far we have tested only a few. A library organization that appears useful when one is dealing with the methyl esters of urinary organic acids (complete spectra except for omission of masses 0–40) may not necessarily be a good solution when one is dealing with other derivatives or other classes of compounds.

Reproducibility

In order to test the reproducibility of the system, a library was generated from the methyl esters of the normal pooled urine. When the actual mass spectrum used to generate the library is matched against itself, as expected 100% coincidence occurs throughout (Fig. 3a). If the "window" is deliberately set in a wrong manner (matching started 25 scan numbers displaced), the results are chaotic, as demonstrated in Fig. 3b.

If another aliquot of the same "library" urine is extracted, derivatized and subjected to the same GC-MS analyses, and then matched against the library, one would expect a high degree of coincidence throughout the chromatogram. In practice, however, there were considerable fluctuations, usually around 75–80% coincidence,

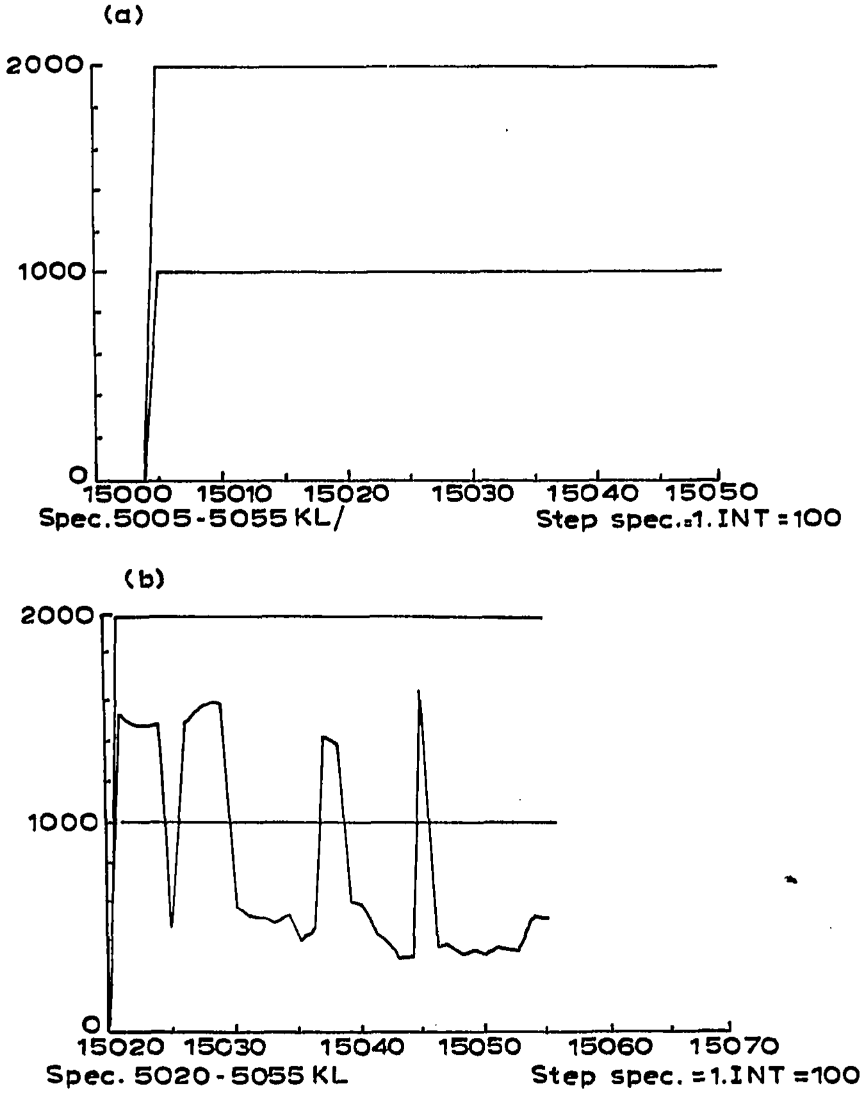

Fig. 3. Effect of wrong "window" setting or incorrect starting points for CASAC analysis. A library was generated from a set of mass spectral data as described in the text. The same mass spectral data were matched against the library using a "window" with ±10 scan numbers, and starting in the correct position (scan No. 10 matched with a library window from 1 to 20). In (a), 100% coincidence throughout is shown. The starting point was then altered so that scan No. 25 of the sample was matched against the same library window as above (scan No. 1–20). The result (b) shows poor coincidence.

and sometimes the sample contained more fragments than the library, whereas with other scan numbers the opposite was true. This leads to a picture shown in Fig. 2. Isothermal GC–MS improved the reproducibility, leading to a degree of coincidence closer to 85% (C^* values around 1150 or 850) (not shown). Special computer listings were produced in order to check whether a window of ±10 scan numbers (corresponding to ±40 sec) was sufficient to compensate for variations in absolute retention times. Even when programming the temperature from 80 to 300° at the rate of 8°/min, it was repeatedly found that such a comparatively narrow search region was indeed acceptable.

Location of abnormalities

In order to demonstrate that the system is capable, in principle, of locating

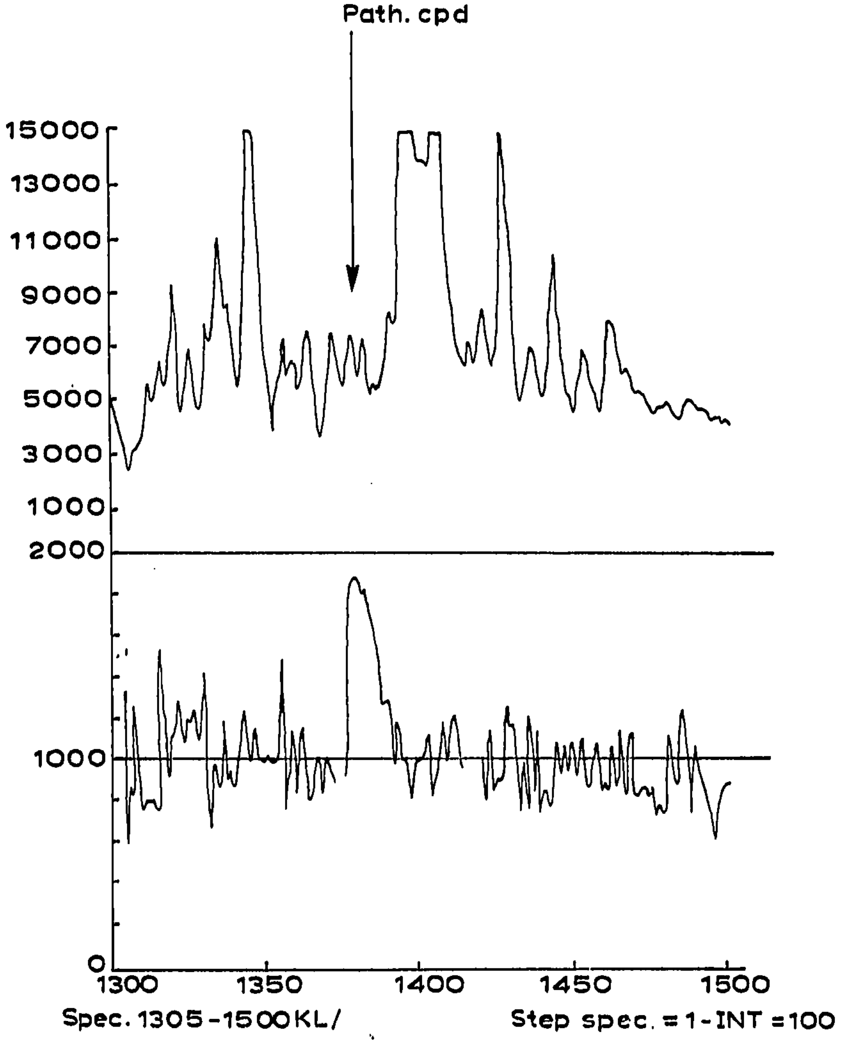

Fig. 4. Localization of an abnormal compound by the CASAC program. A library was generated from the normal, pooled urine sample (organic acid methyl esters), which was separated on an OV-17 column using temperature programming from 80° to 210°. About 200 mass spectra were recorded and converted into a library. A urine sample from a patient was treated in exactly the same way, and automatically matched against the library. Scan number 1375–1385 showed poor coincidence (10–20%) because of the presence of an abnormal compound, which was identified by an off-line library search as a cyclopropanedicarboxylic acid.

anomalies in a sample, we analyzed urines known to contain pathological compounds. The result of such a CASAC search is illustrated in Fig. 4. It can be seen that the degree of matching in the area between scan No. 1375 and scan No. 1385 is very poor (C^* value *ca.* 1800–1900, corresponding to only 20–10% coincidence). The mass spectra in this region were consequently subjected to further investigation, including computer subtraction of the corresponding spectra in the library and off-line computer matching of the "net" mass spectra against our comprehensive file of reference spectra (58,400 entries). The anomalous compound proved to be cyclopropanehexanedioic acid (methyl ester).

DISCUSSION

Nau and Biemann[4] have recently described a computer program for the automatic assignment of retention indices and suggested that the use of retention index and spectral search systems might provide an appropriate set of data for the identification of multi-component mixtures. The system of Sweeley *et al.*[5] also involves the use of GC retention indices, but in conjunction with a small set of discriminating ions to identify each compound in a rapid manner. The CASAC system makes no attempt to identify a given spectrum, but has been designed only to search for differences between a patient's sample and a "normal" reference sample. Once such differences have been indicated by the computer system, a concentrated effort can then be made to identify the anomaly, whether it is the lack of a normal constituent or the presence of an abnormal constituent.

The major difficulty with this first version of the CASAC system is the poor degree of coincidence (*ca.* 75–80%) that is obtained even when one knows for certain that the sample under investigation contains the same compounds as the library. However, despite this drawback, it is our experience that when anomalies exist in a metabolic profile, the degree of coincidence usually decreases well below this normal coincidence level, thus permitting the desired localization of the abnormal region of the chromatograms (*cf.*, Fig. 4). Our experience so far has shown that the comparison of corresponding mass spectra is very sensitive to small differences in the gas chromatograms. We think that this will be further improved by utilizing the absolute peak intensities in relation to internal standards for the calculation of the coincidence, rather than the amount of peaks in common only. In view of the great need for the automatic evaluation of metabolic profiles, we believe that research along the above or similar lines should be continued.

REFERENCES

1 O. A. Mamer, W. J. Mitchell and C. R. Scriver (Editors), *Application of Gas Chromatography and Mass Spectrometry to the Investigation of Human Disease*, McGill University, Montreal Children's Hospital Publication, Montreal, 1974.

2 E. Jellum, in B. J. Mitruka (Editor), *Application of Gas Chromatography in Microbiology and Medicine*, Wiley, New York, 1975, p. 447.

3 S. P. Markey, W. G. Urban, A. J. Keyser and S. I. Goodman, *Advan. Mass Spectrom.*, 6 (1974) 187.

4 H. Nau and K. Biemann, *Anal. Chem.*, 46 (1974) 426.

5 C. C. Sweeley, N. D. Young, J. F. Holland and S. C. Gates, *J. Chromatogr.*, 99 (1974) 507.

6 E. Jellum, O. Stokke and L. Eldjarn, *Anal. Chem.*, 45 (1973) 1099.

7 *Operating Manual for SpectroTheque Programs, SpectroSystem 100 MS*, Varian-MAT, Bremen, 1973.