Journal of Chromatography, 99 (1974) 507–517

CHROM. 7645

RAPID COMPUTERIZED IDENTIFICATION OF COMPOUNDS IN COMPLEX BIOLOGICAL MIXTURES BY GAS CHROMATOGRAPHY–MASS SPECTROMETRY

C. C. SWEELEY, N. D. YOUNG, J. F. HOLLAND and S. C. GATES
Department of Biochemistry, Michigan State University, East Lansing, Mich. 48824 (U.S.A.)

SUMMARY

A new method has been developed for rapid identification of compounds in complex mixtures by a gas chromatograph–mass spectrometer–computer system. The method uses a gas chromatographic retention index in conjunction with a small set of discriminating ions to identify each compound. The retention index predicts to within 1% the location of a time-dependent "window" within which a specific compound may appear, while the accuracy of the peak identification is dependent primarily upon the specificity of the ion set chosen. Optimum specificity is obtained by selecting ions which not only are characteristic of the compound to be identified, but which also differentiate it from other compounds likely to elute at similar retention times. Utilizing the data files created by repetitive scanning during a single elution by gas–liquid chromatography, the present algorithm permits the automated identification of approximately eight compounds per minute.

Examples of the identification of acidic urinary metabolites are given; however, the method is sufficiently general to be applied to any group of volatile compounds.

INTRODUCTION

Considerable effort has been devoted to the development of instrument systems for multicomponent analyses of complex mixtures of biological origin. The concept of the "metabolic profile" of a living organism, first suggested by Horning and Horning[1], could be applied in a variety of biochemical and clinical chemical situations if a practical instrument system were devised for simultaneous identification of the components of such mixtures and estimation of their abundance.

With adequate prepurification of particular classes of compounds, liquid chromatography (LC) and gas chromatography (GC) have each had notable successes in a wide variety of applications. However, extension to the total analysis of all components in a biological fluid has been difficult to achieve using either LC or GC procedures. Some of the factors contributing to this problem have been: the very large number of compounds in a typical sample, difficulties in quantitative determination of compounds with diverse structures and large differences in concentration, the inability to resolve all of the compounds adequately, the presence of a substantial number of

components of unidentified structure, instrumental variability and the sheer bulk of data that must be reduced before meaningful information can be deduced from the analysis. In addition, there is a wide range of variation in the composition of biological samples from one subject to another and even in different samples from the same subject, making it difficult to establish normal values[2]. A final complication when dealing with human subjects is the presence of numerous diet- and drug-dependent variations which are very difficult to control with typical subject groups[3].

The combined gas chromatograph–mas spectrometer–computer instrument system (GC–MS–COM) seems to offer a greater potential than GC alone for the determination of metabolic profiles, and recent progress has been reported in the use of this system for the analysis of various physiological fluids for several classes of metabolites[3], steroids[4,5], drugs[6,7], and other specific types of compounds. Several approaches have been used for the identification of the GC peaks; most have depended upon library search routines[3,5,7,8], spectral interpretation[4,6], or selected ion monitoring (SIM) techniques[9,10]. (For a review of recent papers, see ref. 11.) Recent work by Nau and Biemann[12] and in this laboratory[13] suggests that GC retention data can be a significant additional factor in locating components in complex mixtures; for example, Fig. 1 illustrates the necessity of such an approach in differentiating isomers with similar structures and mass spectral patterns.

This report describes a general procedure which utilizes computer analysis of mass chromatography data[14] to obtain retention indices[15] and selected ion intensities for the qualitative analyses of the volatile trimethylsilyl (TMS) derivatives of urinary

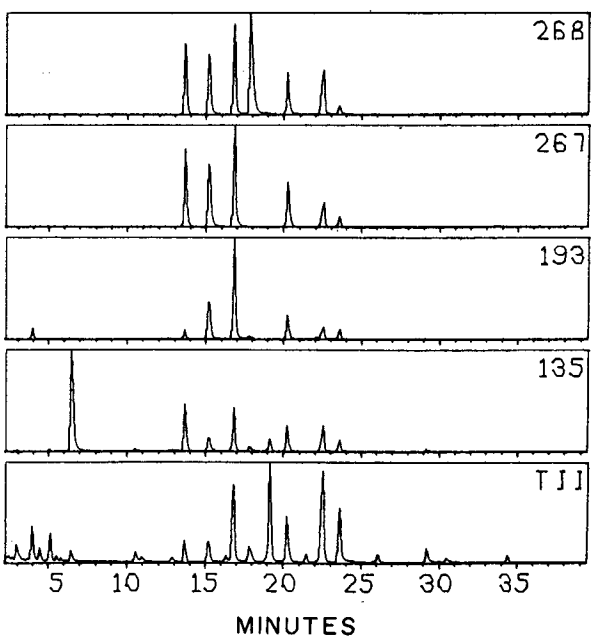

Fig. 1. Mass chromatograms of the artificial mixture of compounds listed in Table I. Ions characteristic of the TMS derivatives of *o*-, *m*-, and *p*-hydroxybenzoic acids include 135, 193, 267 and 268. The major peaks containing these ions are difficult to differentiate unless a retention index or other retention time indicator is used. Conditions are given in Methods. TII = total ion intensity.

acids. The method is generally applicable to any compound for which retention indices and differentiating ions can be assigned for use in the data reduction routines.

EXPERIMENTAL

Materials

Pure aromatic acids were the generous gift of Dr. Clyde Williams, with the exception of α-hydroxyisovaleric, β-hydroxy-β-methylglutaric, homogentisic, and α-ketoadipic acids, which were purchased from Sigma (St. Louis, Mo., U.S.A.). Both reagent-grade methanol (Mallinckrodt, St. Louis, Mo., U.S.A.) and pyridine (Baker, Phillipsburg, N.J., U.S.A.) were redistilled and stored under anhydrous conditions. Bis(trimethylsilyl)trifluoroacetamide (BSTFA) with 1% trimethylchlorosilane (Pierce, Rockford, Ill., U.S.A. or Regis, Morton Grove, Ill., U.S.A.) was stored desiccated at 4° until used. Urine samples were collected in glass containers and stored at −20° until used. Other analytical-grade reagents were purchased from Mallinckrodt.

Methods

Organic acids were recovered from human urine as follows. After the addition of an internal standard, 3 ml of urine were adjusted to pH 1–2 and saturated with NaCl. The acids and neutral fraction were extracted with three portions of ethyl acetate–diethyl ether (1:1). The combined extracts (final volume approximately four times that of the urine) were washed with 0.5 volume of a saturated saline solution (pH 2) and dried over anhydrous Na_2SO_4. After decanting, the solvents were removed *in vacuo* with a Buchler rotary evaporator and the residue transferred in a small volume of redistilled anhydrous methyl alcohol to a conical centrifuge tube. The solvent was removed by evaporation under a stream of nitrogen gas, and the residue was dissolved in dry redistilled pyridine. A mixture of eight straight-chain hydrocarbons from C_{10} to C_{24} was added and the total volume adjusted so that the substances to be analyzed were 50- to 100-fold more concentrated than in the starting urine. Replaceable hydrogens of the urinary metabolites were reacted with the TMS donor (BSTFA) in a sealed silanized capillary tube at 82° for 1 h, after which the reaction mixture was stored at 4° for subsequent analysis. The TMS derivatives appear to be stable for several months under these conditions.

Samples were analyzed with an LKB 9000 mass spectrometer, using a 3.7-m × 2 mm coiled glass column packed with 1% SE-30 for separation of the TMS derivatives[1]. The column was programmed from a starting temperature of 90° to 230° at 4°/min, during which mass spectra were recorded at 6-sec intervals from *m/e* 10–500 under computer control[13]. Raw data from each scan were acquired and processed by an 8K PDP-8/I computer (Digital Equipment, Maynard, Mass., U.S.A.) with two 32K disks; the computer was interfaced to the LKB 9000 through a multiple gain amplifier that provided a linear dynamic range of ion intensities from 0 to 508,000 (ref. 16), and a Hall effect amplifier system for automatic calibration over an extended mass range. Spectra were recorded on magnetic tape for subsequent analysis by the program MSSMET on the PDP-8/I, as discussed below.

Computer program MSSMET

The program for mass spectral metabolite identification (MSSMET) contains

three discrete sections: location of compounds used as retention index references, location of the internal standard, and identification of peaks from a library file of compounds likely to be found in the sample.

Location of retention index standards. The computer routine for metabolite location has options for the use of absolute retention times, retention times relative to a single internal standard, or retention indices relative to a homologous series of hydrocarbons or other standards coinjected with the sample. Analyses based on retention indices have proven to be the most reliable and were used throughout this study.

After the real-time data collection during the gas–liquid chromatographic (GLC) elution of the sample is complete and the data have been stored on magnetic tape, off-line analysis using this program begins with the operator inputting a defined retention index and a predicted retention time for each retention index standard. In the current version of the program, this input may be from either a high-speed paper tape reader or the teletype keyboard. Using the predicted retention time for the standard, the computer calculates a time-dependent "window" or search region in the magnetic tape file of mass spectra within which it must find the compound. The size of the window may be changed by the operator if desired.

The operator then inputs a "designate" ion which has been previously selected as the ion most likely to differentiate a given compound from its neighboring compounds and hence is used as the principal criterion of the presence or absence of the compound. A set of confirming ions are inputted in order of decreasing expected intensity and always includes the designate ion. For each of the confirming ions, those portions of the mass chromatograms that fall within the window are then collected from the magnetic tape file. Using a peak-finding routine, the computer locates all peaks of the designate ion within the search region, and performs a search for peaks of each of the confirming ions within one scan of the designate ion. The actual retention time of each compound is calculated from the position of the highest intensity of the designate ion. The highest intensity of each of the confirming ions that peaks within one scan of the designate ion peak is recorded for calculation of a percent match, as given below.

Finally, the areas of each of the peaks of the designate ion found are calculated within computer-set limits and the peak with the largest area is assumed to be the retention index standard. The retention time (scan number) of that peak is assigned the retention index defined for it in the input for that standard and stored for later use.

This procedure is repeated for each of the reference standards until the computer has a complete table of retention indices with the retention time (scan number) corresponding to each. Based on these standards, the appropriate retention index search window or the retention index for any compound found can easily be calculated by a linear interpolation of the indices of the flanking hydrocarbons or by extrapolation from the indices of the two nearest hydrocarbons if the compound has a lower or higher retention time than any member of the set of hydrocarbons.

Location of internal standard. The quantitative internal standard is located using the same algorithm as for the retention index standards, with the exception that location of the search window is based on a previously determined retention index for that compound rather than an estimated retention time. The most intense peak of

the designate ion within the search window is assumed to represent the internal standard.

Identification of compounds from library. A predicted retention index, designate ion and set of confirming ions are inputted for each compound in the library file. The same algorithm for peak location is followed, and the same data calculated by the computer. In addition, the computer calculates the ratio of the area of the designate ion of each peak found to that of the designate ion of the internal standard.

Percent match. The percent match of each peak found to be a possible candidate for a particular compound is calculated based on the actual order of the maximum intensities for the confirming ions for that compound. Using the formula of Knock *et al.*[8], the percent match, p, is:

$$p = \frac{1}{n^2} \sum_{k=1}^{n'} n - | i_k - j_k |$$

where n is the number of confirming ions searched, n' is the number of confirming ions found, and i and j are the order of the maximum intensities of the confirming ions in the sets searched and found, respectively.

Computer printout. At the end of the calculations for each compound, a finished set of data is displayed on a scope display terminal (Tektronix, Model CRT 4010) and copied, if desired, using a thermal copier (Tektronix, Model 4610). The results are displayed consecutively for each compound in the file as illustrated in Fig. 2 and include the percent match between the input set and the found set, the designate ion

INPUT

```
1400
12:00
71
57,71,85

1600
17:00
71
57,71,85


1800
22:00
71
57,71,85

1
1
1472
179
179,147,180,253

META-HYDROXYBENZOIC
1558
267
267,193,282,223,268
PARA-HYDROXYBENZOIC
1621
267
267,193,223,282,269
BETA-HYDROXY-BETA-METHYLGLUTARIC
1619
363
247,115,363,109
```

OUTPUT

```
H.C. FOUND AT PEAK   2 1400   11:54   12:00
  22      448  11:00   11:00   12:00
 100    10278  11:54   11:54   12:00
  22      412  12:48   12:48   12:00

H.C. FOUND AT PEAK   3 1600   17:18   17:00
  44      636  16:12   16:12   17:00
  22      296  16:48   16:48   17:00
 100    15488  17:18   17:18   17:00
  55     1410  17:43   17:43   17:00

H.C. FOUND AT PEAK   2 1800   22:20   22:00
 100      786  21:14   21:14   22:00
 100    22778  22:20   22:20   22:00

VOLUME OF URINE, I.S.=+0.1000000E+01 +0.1000000E+01

I.S. FOUND AT PEAK   1   13:54 1474 1472
 100    24156  13:54 1474 1472
  56     1408  14:36 1500 1472

META-HYDROXYBENZOIC
 100 +0.4115747E+01  16:12 1559 1558


PARA-HYDROXYBENZOIC
 100 +0.6105315E+01  17:43 1617 1621

BETA-HYDROXY-BETA-METHYLGLUTARIC
 100 +0.1385990E+01  17:43 1617 1619
  12 +0.1622784E-01  18:37 1653 1619
```

Fig. 2. Typical input and output of computer program MSSMET. Entries are described in the text.

intensity ratio for the compound and internal standard, the retention time of the peak, and the experimentally observed and predicted retention indices. Essentially the same format is used for the hydrocarbons and the internal standard with the exception that absolute peak areas are given instead of areas relative to that of the internal standard ion.

RESULTS

Precision of retention indices

One of the most significant features of this method for qualitative analysis is the precision observed in the automated calculation of retention indices, as shown in Table I. In no case was the variation between the mean value and an experimental value more than 0.5%. When the reference compounds listed in Table I were added to urine, the observed retention indices were essentially the same as those calculated by the computer for the simple mixture, indicating that the presence of numerous substances of similar chromatographic retention behaviour had little or no effect on the retention index or the ability of the computer to find the peaks, even where the compound was incompletely resolved by the chromatographic column, as illustrated in Fig. 3. The precision of the retention index appears to be related primarily to the scan time interval; over 95% of the retention indices were found within one scan (approximately four retention index units) of the mean value for that compound. The data of Nau and Biemann[12] also indicate a precision in this range.

Absolute retention times were also reasonably reproducible from one run to another since the hydrocarbon peaks were generally within $\pm$ 60 sec of an average retention time over a prolonged period of use of the system. This factor facilitated the finding of the hydrocarbons in any particular analysis by enabling use of the actual average retention time to be expected for each member. Difficulty arose only when the amount of a hydrocarbon was not sufficient to provide unambiguous peaks for certain members of the confirming set.

TABLE I

COMPUTER-CALCULATED RETENTION INDICES OF SOME ACIDIC COMPOUNDS

No.	*Compound**	*Mean index*	*Standard deviation*	*Maximum deviation*	*Index in urine***
1	α-Hydroxyisovaleric	1166	2	2	1167
2	Benzoic	1226	2	3	1229
3	Mandelic	1472	2	3	1473
4	*o*-Hydroxybenzoic	1501	2	7	1504
5	*m*-Hydroxybenzoic	1558	2	4	1558
6	*p*-Hydroxybenzoic	1621	2	7	1620
7	β-Hydroxy-β-methylglutaric	1619	2	5	1620
8	α-Ketoadipic	1711	1	2	1710
9	Vanillic	1755	2	2	1753
10	Homogentisic	1843	2	4	1847
11	Caffeic	2149	3	7	2152

* Analyzed as TMS derivative under conditions given in the text.

** Three ml of urine to which 20 μg of each of the compounds was added before extraction.

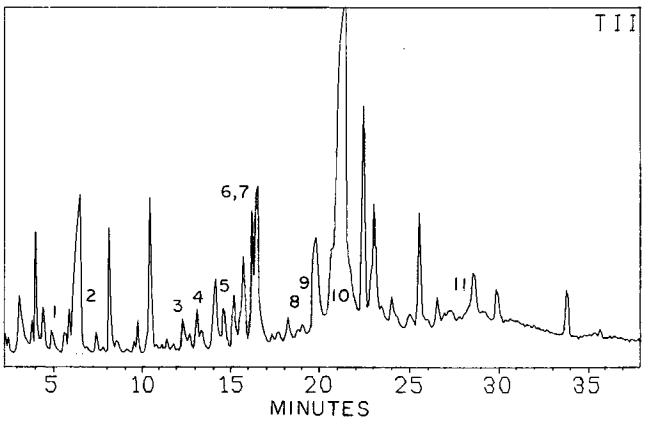

Fig. 3. Location of standard compounds added to a urine sample. Shaded peaks indicate positive match to library file. Numbers refer to the compounds listed in Table I.

Identification of compounds in urine

Two samples of urine, taken at a six-month interval from the same individual, were used as standards. Each sample was analyzed by the GC–MS–COM system using repetitive scanning, and the data set was then examined in detail to select a designate ion and from three to eight confirming ions for each of the major peaks in the total ion intensity plot. These data were then used to re-analyze the same sample of urinary metabolites and were modified until the computer had found all peaks with a 100% match of the designate and confirming ions and the experimental retention index was within the precision limits of the predicted value. The data from the two samples were then combined and used to find peaks in a third sample, which was a pooled urine from several subjects including a less than 5% contribution from the individual who furnished the original reference samples. Any peak in the pooled urine sample was considered "found" if it met both of two arbitrary criteria: confirming ions with a 75% or greater match with those from the reference urine, and retention index within five units of that obtained with the reference sample.

Using the data obtained from the two reference samples as the library file, virtually all of the major peaks in the pooled urine sample were "found" except hippurate, the largest component in the sample. The peaks found are illustrated by the shading in the total ion intensity plot in Fig. 4. Overall, 49% of the peaks in the library were detected in the pooled urine.

A further test of the specificity of the method was provided by assigning incorrect retention indices to the hydrocarbons. $C_{12}H_{26}$ was assigned a retention index of 1240 instead of 1200, $C_{14}H_{30}$ was given a value of 1440, etc. Under these conditions, only 8% of the compounds in the library were found in the pooled urine. Approximately half of these peaks was from isomers of the library compounds or otherwise highly similar compounds, while the other falsely identified peaks contained the confirming ions at very low intensities.

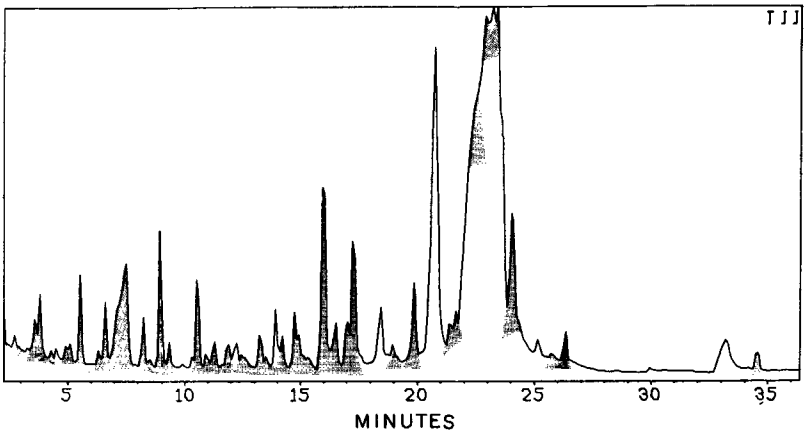

Fig. 4. Peaks found in pooled urine sample. Shaded areas indicate a positive match with a compound in one or both of two standard urine samples. Details in the text.

DISCUSSION

Nau and Biemann[12] recently described a computer program for the automatic assignment of retention indices of all compounds in a complex mixture, and suggested that the use of retention index along with spectral interpretation and/or search systems might provide an appropriate set of data for the identification of multi-component mixtures. The present version of our GC–MS–COM system, which is an extension of concepts described earlier[13], has incorporated the use of retention indices to limit the region within which the computer searches for the presence of characteristic ion sets which provide compound identification. This approach increases the accuracy of identification and greatly decreases the time involved in the processing of data from several hundred mass spectra.

The accuracy obtained in the identification of peaks ultimately depends on two factors: correct assignment of retention indices and selection of highly differentiating sets of confirming ions. The first of these factors is clearly the most easily accommodated in a computer program, and our system has been demonstrated to be capable of accuracy to within ten retention index units in all assignments. Appropriate selection of ions to accommodate the second prerequisite is more difficult to make on an empirical basis. Some of the criteria which should govern the selection of confirming ions include:

(a) Avoidance of characteristic ions which do not differentiate the compound from other substances in the mixture with approximately the same chromatographic properties. For the TMS derivatives used in this study, this suggested that ions of any of the following categories should be avoided: *m/e* 0–60, 73–77 and 147–149. These ions are present in many of the compounds in the mixture; hence peaks of interest containing these ions are often completely obscured by peaks of the same ions from different compounds eluting at a similar time.

(b) Preference of high-mass ions for integration. The higher the mass of the ion, the more likely it is to be unique for the compound in a particular retention index region.

(c) Preference for high-intensity ions. The more intense an ion is, the more likely it is that the compound will be detected in complex mixtures of poorly resolved substances or when present at low concentrations.

(d) Avoidance of ions with similar intensities. With the present version of the program, the order of the intensities is important but the ratios are not. Since ions with approximately equal intensities may change in their intensity ranking due to random fluctuations in measurement, inclusion of such ions adversely affects the percent match.

(e) Exclusion of isotope ions. Although several ions in an isotopic cluster can be used, it has been our practice to use other ions if possible.

(f) Choice of number of confirming ions. No correlation of the number of confirming ions to the percent of peaks found in a mixture has been made as yet, although there is some evidence that substances with three or fewer confirming ions tend to be found less accurately.

Once appropriate ions have been selected for a given compound, the order of intensity of those ions is often a highly differentiating indicator of the likelihood that a particular spectrum matches that in a library file. An example of this is shown in Fig. 5, in which several compounds in the urine sample contain all of the confirming ions but only one has the correct set of ions in the proper order. The percent match gives an easily interpreted numerical value to the degree of agreement between library and actual spectra. However, isomeric or otherwise similar compounds may be unresolvable by any factor based on spectra alone, and hence retention indices were always calculated concurrently.

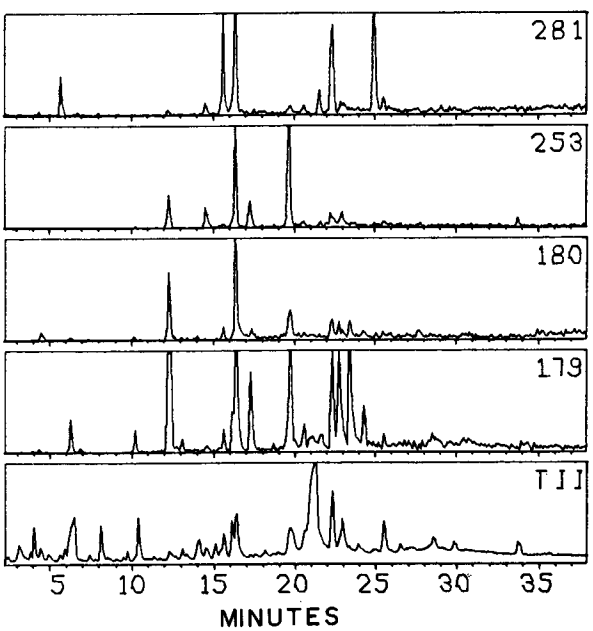

Fig. 5. Stacking of ions in correct order. Only the peak at 12.3 min contains the ions for the mandelic acid derivative in the correct order, 179 > 180 > 253 > 281. Mandelic acid was added to the urine sample before extraction. Ion chromatograms are all normalized to the most intense value of *m/e* 180.

Another important variable affecting the accuracy of the program is the peak size. To give the operator an indication of the size of the designated peak, the ratio of the area of the designate ion of the compound to that of the internal standard is calculated by integration within computer-selected limits. Generally, the lower the ratio, the less certain the assignment of peak identity.

A further factor contributing to the accuracy of peak identification is the use of an adequate number of retention index standards. With the current program, any number of standards up to twenty may be used; however, for many portions of a temperature-programmed analysis the relationship of retention index to retention time is essentially linear[12], and hence fewer standards are required. It has usually proven adequate to mix approximately eight hydrocarbons with the sample for analyses programmed from 90 to 250°. For any of these hydrocarbons, *m/e* 71 is the designate ion usually used, and *m/e* 57, 71 and 85 the set of confirming ions.

Although it generally does not affect the accuracy of peak identification, the size of the search window is also an important variable. The rate at which the computer can output results with the present version of the system varies from three to eight compounds per minute as the size of the search window is changed from 240 to 80 sec. When all of the retention indices of the library compounds are known to be precise, even smaller windows, and hence higher search rates, are possible.

Currently, the major fault in the MSSMET program seems to be that it fails to identify the very large hippurate peak. This compound almost always exceeds the capacity of the lightly loaded SE-30 column, and its retention time at the apex of the peak is therefore variable and dependent on the degree of asymmetry of the peak. In addition, the intensities of the designate and confirming ions may saturate the analog amplifier, leading to flat peaks with which the peak-finding program cannot cope.

In overall terms, it is clear that the MSSMET program provides a method for rapid, automated identification of urinary metabolites and can be applied equally well to other complex mixtures. The algorithms involved in MSSMET represent a departure from other library search programs in the sense that a restricted portion of the data file is searched, and the program utilizes ions that *differentiate* the compound in a certain milieu rather than ions that may be most characteristic of that type of compound. In dealing with complex mixtures, some ions are much more likely to be informative than others, particularly those which are common to many components or even to all components if derived from the functional group used to increase volatility.

An advantage of the program is the relatively small size of the library required. Only three to eight ions and a retention index must be stored for each compound. A file of several hundred compounds is highly compatible with the capabilities of minicomputers currently used in many laboratories. The essential information for these compounds can be stored on paper tape, magnetic tape or disk, depending on the configuration of the computer system. With a large disk and a disk operating system, MSSMET would be considerably faster than that described in this report.

Many of the compounds in the acidic fraction of human urine have not been identified. An added advantage of MSSMET is that unknowns can be found in complex mixtures equally as easily as known compounds, provided the retention index and a set of confirming ions have been selected.

Presently, the program is being expanded to include the use of the peak area of the designate ion to calculate the amount of each compound present in the mixture

relative to that of the internal standard. In cases in which the area of the designate ion contains no contributions from other compounds, it will be possible to catalog urine levels of both known compounds and unknown urinary metabolites in a large number of subjects with a variety of clinical abnormalities. We are encouraged to believe, therefore, that MSSMET provides an opportunity to evaluate the clinical importance of metabolic profiles with an automated single instrument system.

ACKNOWLEDGEMENTS

The authors are grateful to Mr. J. Harten for technical assistance in acquiring mass spectra. The work has been supported, in part, by a research grant (RR-00480) from the Biotechnology Branch, National Institutes of Health.

REFERENCES

1 E. C. Horning and M. G. Horning, *J. Chromatogr. Sci.*, 9 (1971) 129.
2 T. A. Witten, S. P. Levine, J. O. King and S. P. Markey, *Clin. Chem.*, 19 (1973) 586.
3 E. Jellum, O. Stokke and L. Eldjarn, *Anal. Chem.*, 45 (1973) 1099.
4 R. Reimendal and J. B. Sjövall, *Anal. Chem.*, 45 (1973) 1083.
5 J. D. Bary and A. P. Wade, *Anal. Biochem.*, 57 (1974) 27.
6 F. Hutterer, J. Roboz, L. Sarkozi, A. Ruhig and P. Bacchin, *Clin. Chem.*, 17 (1971) 789.
7 C. E. Costello, H. S. Hertz, T. Sakai and K. Biemann, *Clin. Chem.*, 20 (1974) 255.
8 B. A. Knock, I. C. Smith, D. E. Wright, R. G. Radley and W. Kelley, *Anal. Chem.*, 42 (1970) 1516.
9 J. T. Watson, D. R. Pelster, B. J. Sweetman, J. C. Frolich and J. A. Oates, *Anal. Chem.*, 45 (1973) 2071.
10 R. E Summons, W. E. Pereira, W. E. Reynolds, T. C. Rindfleisch and A. M. Duffield, *Anal. Chem.*, 46 (1974) 582.
11 A. L. Burlingame, R. E. Cox and P. J. Derrick, *Anal. Chem.*, 46 (1974) 248R.
12 H. Nau and K. Biemann, *Anal. Chem.*, 46 (1974) 426.
13 C. C. Sweeley, S. Gates and J. F. Holland, in O. A. Mamer, W. J. Mitchell and C. R. Scriver (Editors), *Application of Gas Chromatography–Mass Spectrometry to the Investigation of Human Disease*, McGill University–Montreal Children's Hospital Research Institute, 1974, p. 141.
14 R. A. Hites and K. Biemann, *Anal. Chem.*, 43 (1971) 681.
15 E. Kováts, *Helv. Chim. Acta*, 41 (1958) 1915.
16 C. C. Sweeley, N. D. Young and J. F. Holland, unpublished.