

A Combined Gas Chromatography-Mass Spectrometry-Computer System for the Analysis of Volatile Components of Foods

Charles Merritt, Jr.,* Donald H. Robertson, John F. Cavagnaro,¹ Richard A. Graham, and Thomas L. Nichols

The great complexity in the composition of the volatile components of foods has required the development of elaborate systems of analysis employing the combination of gas chromatography and mass spectrometry for separation and identification. In some laboratories a multiplicity of such analysis systems is employed and a computer becomes a mandatory adjunct in order to process the copious amounts of

data generated. The utilization of a computer for acquiring and processing data from several gc-mass spectrometry analysis systems is described with special emphasis on data encoding techniques and procedures for component identification by reference to precoded data files. Examples are given to illustrate the application to typical qualitative and quantitative analyses of volatile food components.

14 NOV 1974

One of the primary objectives in studies of flavor chemistry is to establish the chemical composition of the volatile constituents that contribute to the aroma or odor of foods. The role of combined gc-mass spectral methods to perform these analyses is well known. Moreover, the application of computer processing to gc-mass spectral data is likewise becoming well established. Several dedicated gc-mass spectrometry-computer systems are available commercially and a number of configurations for larger computer systems for both on-line acquisition and off-line processing have been described in the literature (Henneberg *et al.*, 1972; Hites and Biemann, 1967, 1968; Knock *et al.*, 1970; Smith *et al.*, 1971; and Venkataraghavan *et al.*, 1970).

The size of a given system and the extent of its use are, of course, largely dictated by the requirements and resources of the user. There is a need, however, to develop means of simplifying the computer systems, reducing their size, enhancing their efficiency, and in particular increasing their utilization by decreasing the cost. This presentation described some approaches which have been devised in our laboratory to achieve some of these objectives.

Although compound identification from mass spectra is possible from basic principles, the identification of unknown compounds from their mass spectra by automatic data processing assumes the existence of a file of data for a large number of known compounds, and the ability to search the contents of that file in a manner which provides component identification. The conventional approach uses tables of mass *vs.* intensity values that constitute digitized mass spectra. By comparing the unknown to each known spectrum in the library file, and computing a matching index for each trial, it is possible to achieve identification from the best match of the unknown to a known spectrum in the library file.

Because of the large number of spectra to be searched in a gc-mass spectral analysis system, the computer configuration required to execute searches of data for unknown component identification is usually quite large, and the time required for retrieval is slow. Moreover, there is a need in many laboratories to be able to perform computer searches on smaller computers that lack adequate peripheral storage facilities for large data files.

In order to overcome this limitation our laboratory has been concerned with the encoding of spectra in compressed data files. Three different approaches to the construction of such files may be summarized as follows: (1) calculation of entropy function; (2) calculation of divergence function; (3) selected binary encoding.

The first two approaches to the classification of mass spectral data, namely the calculation of the Khinchine entropy function and of divergence values, are derived from set theory, and are based on expressions of the statistical distribution of peaks in a mass spectrum. These as well as the selected binary coding method which is described below all reduce the mass spectrum to a single valued number which is diagnostic for the compound.

The entropy function is expressed as

$$\eta = - \sum_i^n p_i \log p_i$$

and is calculated by summing the product of the individual ion abundances and their respective logarithms. p represents the ion abundance in terms of the per cent of total ionization of the molecule, or, in another sense, the probability of occurrence of that ion fragment in the spectrum.

Mass spectra are converted by this relationship to a single valued number and in this way a data file can be constructed consisting of these numbers. An example is seen in the portion of such a file tabulated for a group of aliphatic hydrocarbons in Table I. These compounds have been selected to show the typical variation in the entropy value which is expected for the variation in the degree of unsaturation in the molecule. In search of a file of precalculated Khinchine values, a matching index is used to establish the correspondence of the value for an unknown compound with the library value. In early work with Khinchine values, the construction of the reference file was limited to about 200-300 compounds normally encountered in the analysis of volatile components of foods, and in most cases application of the search and retrieval scheme led to single, unequivocal identification of the unknown. In subsequent tests of this procedure with large libraries—*e.g.*, the "Atlas of Mass Spectral Data" (Stenhagen *et al.*, 1969) containing about 7000 compounds—considerable overlap was observed and in many instances there were several compounds which had the same or very close entropy values. Although this may be due in part to the poor quality of the spectral data which exists in most large library files, it has not been possible as yet to purify these data. For libraries of selected compounds, however, the method has been shown to be highly efficacious.

In cases where the entropy values for more than one compound may be too close to provide unambiguous iden-

Pioneering Research Laboratory, United States Army Natick Laboratories, Natick, Massachusetts 01760.

¹ Present address: Department of Chemistry, University of California, Berkeley, Calif.

Table I. Khinchine Entropy Values for Mass Spectra of Selected Hydrocarbons

Compound	Khinchine function
<i>n</i> -Butane	0.927
2-Methylpropene	0.812
<i>trans</i> -2-Butene	1.042
3-Methyl-1,2-butadiene	1.216
1,3,5-Hexatriene	1.340
1,5-Hexadiyne	1.605
3-Heptyne	1.379

Table II. Divergence Values of the Mass Spectra of Selected Hydrocarbons^a

Comparison compd	Divergence, J
<i>n</i> -Hex-1-ene	4.227
2-Methylpent-2-ene	7.780
Cyclohexane	10.024
3-Hexyne	11.830
2-Methylpentene	12.505
1-Hexyne	18.716
2-Hexyne	25.809

^a The reference compound for the series is hexane.

tification, the divergence function is used to resolve the ambiguity (Farbman *et al.*, 1973). The divergence function is expressed as

$$J(1,2) = N_1 \sum_i^n (p_{1i} - p_i) \ln \frac{p_{1i}}{p_i} + N_2 \sum_i^n (p_{2i} - p_i) \ln \frac{p_{2i}}{p_i}$$

where

$$p_i = (p_{1i} + p_{2i})/2$$

and N_1 and N_2 are the respective total abundances of ions in each compound.

This function is similar in nature to the entropy function, since the calculation is based on the evaluation of the summation of the products of the ion abundances and their logarithms, but in this case the equation is derived to establish a comparison of the values for two compounds, specifically where p_{1i} and p_{2i} are the ion abundances expressed as the per cent of total ionization for each of the compounds for which the divergence J is calculated. In practice, it has been found convenient to refer the calculation of divergence of a given compound in the aliphatic hydrocarbon series to that of the normal alkane of the same carbon number. Thus, as seen in Table II the divergence values are listed for several C₆ hydrocarbons referred to *n*-hexane. In a like manner, C₄ or C₅ compounds would be referred to *n*-butane or *n*-pentane, C₇ and C₈ to *n*-heptane and *n*-octane, and so forth. Thus, in a library file of divergence values, a group of subsets is established corresponding to the values for the compounds having the same carbon number. This greatly reduces the number of values to be searched and correspondingly the time to execute the search. For example, in the case of two compounds having close entropy values such as hexene and methylpentene, it is seen that divergence values are sufficiently different to provide unambiguous identification.

The calculations required to encode spectra as entropy, and particularly divergence functions, are somewhat lengthy to perform in a small computer without the aid of a hardware arithmetic unit. Moreover, the variability in the values of relative ion abundances with variations in mass spectrometer design and operation produces considerable uncertainty in the reliability of diagnostics based on measurement of spectral intensity factors. For these

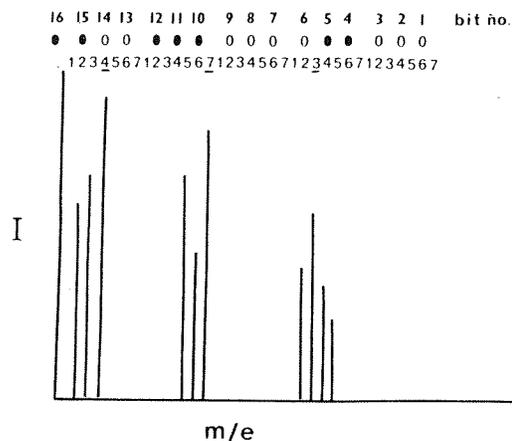


Figure 1. Representation of binary encoding of selected peaks in a mass spectrum.

Table III. Example of Octal Coding

Mass ranges	m/e to be encoded	Position of m/e in octet	Binary code	Octal code
23-29	24	2	010	2
30-36	32	3	011	3
37-43	43	7	111	7
44-50	0	0	000	0

reasons, there has recently been conceived and tested a codification procedure for use with low resolution mass spectral data banks which allows compression of the library file through selective binary coding of characteristic peaks and use of variable length logical records.

The coding procedure is illustrated in Figure 1. A hypothetical mass spectrum is shown with a representation of a 16 bit computer word at the top. Selective binary coding of characteristic peaks is accomplished by arbitrarily dividing the mass range of interest into multiple groups of seven. The number corresponding to the spectrum peak in each group having the highest intensity is then encoded as a three bit binary number. Thus the fourth peak is encoded in the first grouping, the seventh peak in the second, and so on; zero is used to denote the absence of a peak within the grouping, thereby giving a total of eight possible values, hence the term "octal coding" by which this scheme has been designated in prior descriptions (Robertson *et al.*, 1972). The procedure is quite similar to a method proposed independently by Grotch (1970) in which the spectrum is divided in groups of 14 mass units instead of 7. There are several advantages and disadvantages of grouping of 7 vs. 14, but it is not relevant to discuss those aspects in this presentation. The decision to adopt a grouping of seven in our laboratory was inherently pragmatic since it is based on the word size of our computer.

Representation of an octal number within the computer requires three bits; thus, in a 16-bit machine such as the Hewlett-Packard 2116B used in setting up this system, five octal characters can be stored in each computer word with one bit left over. Thereby a single computer word is capable of storing information which covers a range of 35 atomic mass units. Compounds requiring a greater range of mass units to be encoded require an additional number of computer words. As many are used as are needed to encode the spectrum. The last word is then designated by setting a flag in the 16th bit. A further illustration of the octal coding scheme is seen in Table III.

If consideration is given to the m/e values which occur most often in the spectra of organic compounds, a series of octal ranges beginning with the group of seven masses,

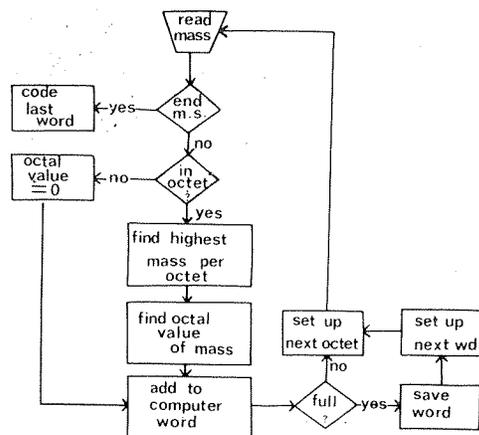


Figure 2. Flow chart for on-line encoding of mass spectra.

Table IV. Subsets of Octally Coded Spectra

No. of words	Subset
One	Ethane 161000
Two	Benzene 040371, 160710
Three	2,3-Dimethyl-3-pentanol 052725, 025252, 102000

23-29, serves to provide greater diagnostic capability for this method of coding. Subsequent mass ranges would be 30-36, 37-43, etc. The first octad, containing masses 12-14-15-16-17-18-19, was included in the original codification procedures, but when it was learned that no additional information was gleaned from using these m/e values, the entire octad was dropped from consideration.

The efficiency of a coding procedure is reduced by the need to use "0" for coding, *i.e.*, coding which leads to a large number of zeros. If one codes in octads (or some larger sized grouping) there is more likelihood of a peak appearing and thus, according to the basic principles of information theory, provides more efficient transmission of information. In general, the encoding and retrieval of data from a number system such as this are predicated primarily on the principle of simple manipulation of the numbers. By arranging the mass ranges so that certain ions fall characteristically in particular octads, it is also possible to develop qualitative information content as well that may relate to the functional group structure of the molecule. However, presentation of the details of this aspect must be deferred at this time.

The octal code for a mass spectrum may be readily obtained from digitized mass and intensity data acquired on-line and stored for subsequent processing. A flow chart of our program for encoding the mass spectrum into its corresponding octal code words is shown in Figure 2. In our laboratory a normal sequence of data processing would involve the following on-line operations: (1) conversion of the analog mass spectrometer output to digital format, *i.e.*, mass and intensity data; (2) condensation of the data to octal format by means of this routine (Figure 2) which determines the most intense peak in each octal grouping and provides the binary equivalent as the "spectrum" for which the data library is searched; (3) when binary "1" is sensed in the 16th bit, the number of words to encode the "spectrum" is known; thus it is not necessary to search the entire library, but only the subset or sublibrary collection of spectra which require that number of words for coding. With this selective coding technique, it is possible to divide the total file into a series of subfiles based on the number of computer-words necessary to selectively code

Table V. Structure of Library Data File

2	No. of words to encode spectrum
36	No. of compds encoded
3040	Code for MeSH
4000	Label for MeSH
3	Label for MeSH
2570	Code for propenal
0600	Label for propenal
4	Label for propenal

the spectrum to its highest observed m/e . Some examples are seen in Table IV.

This particular organizational form of the data library file appears to be of special utility in fully automated gas chromatographic-mass spectrometric analysis systems, since the highest observed m/e is a quantity readily extracted during the data reduction process. In actual use, such file organization implies prefiltering of the data tables, since only those subfiles having the same number of words as the unknown must be searched.

As previously indicated, the subfiles, as constructed in this work, use the 16th bit of the word to signify the end of the logical record. The word immediately following the end of the logical record thereby contains an integer pointer to a separate file containing the alphanumeric characters of the compound name. The division into subfiles of variable logical record length and creation of a name file were designed to make maximum use of random access mass storage devices. For example, if full binary representation is used to code the spectra of both ethane and tridecane, it is found that sixteen 16-bit words of uniform logical record length are required. However, the use of variable-length logical records for the same pair of compounds requires one and five 16-bit words, respectively, to identify the reference spectra, *i.e.*, ethane and tridecane. An example of the construction of a subset file is illustrated in Table V. The name of the file is designated simply as 2 corresponding to the number of computer words for each compound. The number of compounds in the subset and then the code word(s) for each compound are listed. Instead of listing a name for each compound in the subset which is searched, an index number or label is given for each compound name so that the names can be stored in a separate library and can be called up later without encumbering the computer during a search.

For each unknown compound being searched for in the library, a matching index² is calculated; the five best matches, in decreasing order of goodness of match, are printed. This feature is expected to be most useful in future cases when much expanded library files are being searched and the possibility exists for the same matching index to be calculated for more than one compound.

In addition to the utility of construction of subfiles for more efficient retrieval, the octal coding system is found to provide a damping effect on variations in spectral characteristics due to its relative insensitivity to errors in digitization. For example, errors which may occur in the initial codification of an unknown spectrum, due to a spurious signal or the additive effect of impurities upon peak intensities, have insignificant influence upon correct iden-

² The matching index is calculated in the following way: INDEX = XOR + 2nNZW - 2AND, where XOR is logical exclusive OR and AND is logical AND. XOR = +1 if the *i*th octal window of the unknown disagrees with the *i*th octal window of a library spectrum, and either or both windows are nonzero. XOR = 0 if both windows agree. 2nNZW is twice the number of nonzero windows in the code for the unknown spectrum. AND = +1 if both windows agree and both are nonzero; otherwise AND = 0.

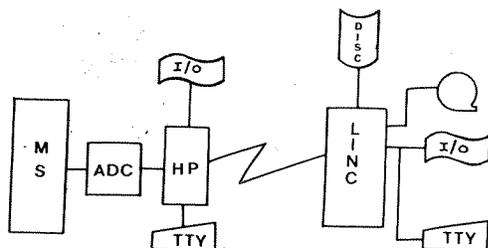


Figure 3. Schematic diagram of computer system.

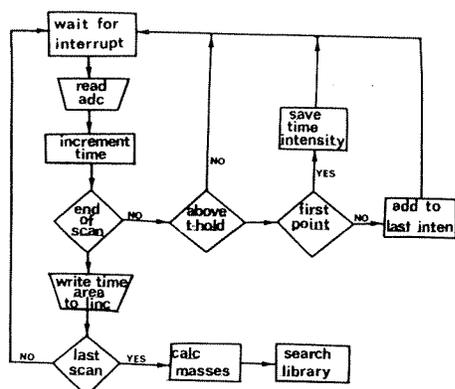


Figure 4. Data flow chart for combined system for acquisition and processing of mass spectra.

tification of the compound in question. Likewise, in the case where the spectrum of an unknown compound is incorrectly coded because of variation in relative intensity values due to mass spectrometer instabilities, it is shown that correct identification may result even in cases where a coding error occurs in each word of the spectrum.

This system of encoding is particularly useful for gc-mass spectrometry because only the most intense ions need to appear in the spectrum and trace amounts can be identified even though the less abundant peaks are absent from the spectrum.

The practical application of octal coding to the processing of analytical data for flavor studies is illustrated below. The processing of mass spectral data, however, is constrained by the limitations of the capabilities of the computer system.

A schematic diagram of the equipment available for current studies is shown in Figure 3. The system used for data acquisition is a Hewlett-Packard 2116B computer which is coupled to a gas chromatograph-mass spectrometer analysis system. The mass spectrometer is a magnetic deflection type capable of providing 1-sec spectrum scans which can be digitized by the computer in real time. The system has only 8K of 16-bit core and its I/O structure includes only paper tape read and punch and a teletype. It is thus not possible to output in real time, or to store large volumes of acquired data for subsequent processing. A LINC-8 computer, however, equipped with a disk storage device resides in a nearby laboratory. The LINC-8 is an 8K 12-bit machine having in addition to the disk a paper tape I/O, a teleprinter, and block addressable magnetic tape. By some special output, formatting data from the Hewlett-Packard Computer may be transmitted to the LINC-8 for disk storage of the acquired data. Subsequent processing is then accomplished in the LINC-8.

Figure 4 shows a chart of the data flow in the combined Hewlett-Packard-LINC-8 system. The mass spectrum signal above threshold is digitized and, using centroid computations, is reduced in the Hewlett-Packard to time and intensity values which are transmitted to the LINC in real

Table VI. Composition of Limited Data Library by Function Groups

Alkanes	28	Esters	39
Alkenes	17	Acids	8
Alkynes	8	Amines	11
Alkadienes	4	Cyclic	16
Alkanols	22	Oxy	13
Alkanals	11	Halogens	4
Alkanones	18	Thiols	5
		Thiaalkanes	8

Table VII. Composition of Word Subfiles in Limited Data Library

No. of words	1	2	3	4	5	6	7	8
No. of compds	9	36	65	59	22	15	7	6
Total				221				

time and stored on the disk. At the conclusion of a chromatographic run, the spectra are successively retrieved from the disk, mass and intensity are calculated and converted to octal code, and the library is searched. An interim printout can be provided if required by the analyst if, for instance, the spectrum is found not to be in the library and the digitized data are needed for another type of search.

The current system is being used to expedite the interpretation of data obtained from large numbers of samples required to be analyzed by gc-mass spectrometry. In a study of the wholesomeness of irradiated beef the study in 1 year may require identification of the components corresponding to 150,000 spectra. Fortunately, the composition of many of the components can be anticipated and thus a limited library file for searching can be established. Some of the expected compounds in beef as well as other current food items of interest such as strawberries or fish are listed in our current library as shown in Table VI. In all, the library contains about 200 compounds. Using subfiles, however, the number of compounds to be searched in most cases is usually quite limited. Some examples of subfiles from the limited library are shown in Table VII. The top row shows the number of computer words needed to encode the spectrum and the number below corresponds to the number of compounds in the subfile.

An example of a computer search, as executed, to identify a gc peak from an analysis of the volatiles isolated from strawberries is illustrated in Tables VIII-X. A printout of the digitized mass spectrum for scan no. 16, corresponding to one of the peaks in the chromatograms, is shown in Table VIII. The presence of peaks in the spectrum corresponding to masses 18, 28, and 32 can be attributed to background peaks such as water and oxygen and nitrogen from the small amounts of air present. If this spectrum is coded according to the selected mass scheme without eliminating the background, key peaks in the spectrum such as 27 or 31 would not be encoded. It should also be noted that the highest observed mass is 89 which is uneven and probably denotes that the parent mass, which is expected to have an even value, is missing from the spectrum.

An example of a typical computer printout of a library search is depicted in Table IX. In the search routine a method for interaction by the spectroscopist has been established so that one of a number of options may be selected before instituting the search. These options, 1-9, are seen at the beginning of the search routine. In the case depicted here option 4 to delete specific masses is selected, and masses 18, 28, and 32 are eliminated from the spectrum.

- Henneberg, D., Casper, K., Ziegler, E., Weimann, B., *Angew. Chem., Int. Ed. Engl.* 11, 347 (1972).
- Hites, R. A., Biemann, K., *Anal. Chem.* 39, 965 (1967).
- Hites, R. A., Biemann, K., *Advan. Mass Spectrom.*, 37 (1968).
- Knock, B. A., Smith, I. C., Wright, D. E., Ridley, R. G., Kelly, W., *Anal. Chem.* 42, 1516 (1970).
- Robertson, D. H., Cavagnaro, J. F., Holz, J. B., Merritt, C., Jr., Proceedings of the 20th Annual Conference on Mass Spectrometry, and Allied Topics, American Society for Mass Spectrometry, Dallas, Tex., 1972, Paper No. R3.
- Smith, D. H., Olsen, R. W., Walls, F. C., Burlingame, A. L., *Anal. Chem.* 43, 1796 (1971).
- Stenhagen, E., Abrahamsson, S., McLafferty, F. W., Ed., "Atlas of Mass Spectral Data," Wiley/Interscience, New York, N. Y., 1969.
- Venkataraghavan, R., Klimowski, R. J., McLafferty, F. W., *Accounts Chem. Res.* 3, 158 (1970).

Received for review December 5, 1973. Accepted April 26, 1974. Presented at the Symposium on Computers in Flavor Chemistry sponsored jointly by the Agricultural and Food Chemistry and Analytical Chemistry Divisions, 166th National Meeting of the American Chemical Society, Chicago, Ill., Aug 1973.