Pure & Appl. Chem., Vol. 49, pp. 1817 -1825. Pergamon Press, 1977. Printed in Great Britain.

STORAGE AND RETRIEVAL OF MASS SPECTRAL INFORMATION

Michael Ed. Hohn, Michael J. Humberston and Geoffrey Eglinton

Organic Geochemistry Unit, School of Chemistry, The University,
Bristol BS8 1TS.

Abstract - Computer handling of mass spectra serves two main purposes: the interpretation of the occasional, problematic mass spectrum, and the identification of the large number of spectra generated in the gas-chromatographic-mass spectrometric (GC-MS) analysis of complex natural and synthetic mixtures. Methods available fall into the three categories of library search, artificial intelligence, and learning machine. Optional procedures for coding, abbreviating and filtering a library of spectra minimize time and storage requirements. Newer techniques make increasing use of probability and information theory in accessing files of mass spectral information.

INTRODUCTION

The inspection of the large libraries of mass spectra now available has become a task for computers rather than manual procedures. Furthermore, the advent of computerised GC-MS systems equipped with facilities for rapid scanning has led to the acquisition of many spectra per analysis, though several spectra may be those of a single compound. For example, a scanning cycle of 2 sec. generates 1,800 spectra in a one hour GC-MS run, and 14,400 spectra by the end of an eight-hour day.

Although itself partially responsible for the volume of mass spectral data, computerised data handling offers the best means for dealing with this surfeit. Problems commonly associated with GC-MS analyses include overlapping GC peaks, short elution intervals, a wide range of peak sizes, large numbers of compounds in complex mixtures, and variable background signals. Deconvoluting spectra of overlapping components can produce spectra that are presumed to correspond to those of the pure compounds (Ref.1, 2). Methods are also available for correcting the skewing of m/e intensities that results from scanning the leading or trailing edge of the GC peak (Ref.2). Consideration of background signals allows the elimination of background spectra from further study, and also allows subtraction of these signals from the spectra of minor components (Ref.2). All of these spectral enhancement methods are part of the data storage and retrieval process in the sense that they reduce the number of spectra requiring identification, and improve the quality of the acquired spectra, thus increasing the likelihood of future retrieval success.

The structure of a spectral library and the arrangement and nature of the data placed in it should relate to the motivation for retrieval. The library can act as a repository for information in the form of full or partial mass spectra which may be accessed in response to a specific request. Questions one might ask of such a library are: What is the mass spectrum of a specific compound?; What are the characteristic fragmentations of a class of compounds?; What m/e can be used to discriminate between one group of compounds and all others? The first question could precede a manual identification of an unknown mass spectrum; the second could precede a study of fragmentation mechanisms; and the third, the monitoring of a specific group of compounds by mass fragmentography (Ref.3).

The library can also act as a source of the spectra needed to confirm the suspected presence of a particular component in a sample. In a reverse search (Ref.4, 5), library spectra are compared to the unknown spectrum, thereby indicating the presence of a component even if the unknown spectrum is that of a mixture of compounds.

The greatest interest in storage and retrieval of mass spectral data lies in the identification of the spectrum of a single, unknown, chemical compound. The spectroscopist may be interested in ascertaining only the gross structural features of the unknown, but more often one wishes to match it with a specific reference compound. Failing a specific identification, as will happen when that reference spectrum is not in the library, the retrieval method should either suggest a possible structure or list reference compounds of structure similar to the unknown. Naegeli and Clerc (Ref.6) list other desirable features of such

retrieved spectra; acceptable results with slightly impure samples; allowances for instrumental error; a search strategy tailored to the problem at hand; and relative ease of utilisation by the chemist.

The discussion which follows deals mainly with the methods available for the identification of an unknown by detailed processing of its mass spectrum as these methods hold the greatest interest at present. However, reverse search and feature extraction are discussed where appropriate. In keeping with the format of this symposium, methodological descriptions will center around prominent examples. An attempt will be made to synthesise current aspects of the topic rather than review the literature. For a review, see Ref.7. The three approaches used in identifying an unknown spectrum comprise library search, artificial intelligence, and learning machine.

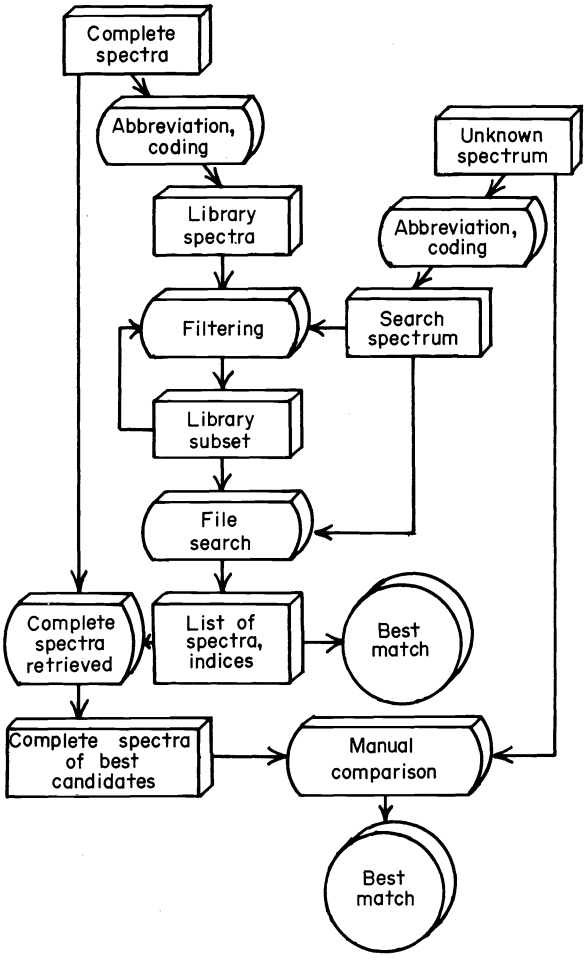

Fig.1 General scheme of library search.

LIBRARY SEARCH

Most library search systems are similar in the steps used to create the library and in the subsequent retrieval of reference spectra (Fig.1). Individual systems minimize requirements for storage and for search time in different ways, particular steps being combined or omitted.

Library creation begins with a set of complete reference spectra, which have been coded and abbreviated before being stored in the library. The coding and abbreviation schemes must be applied uniformly to all spectra, reference and unknown.

The actual search begins with a filter or a sequence of filters that limit the file search to a subset of the complete library. The search involves calculation of a similarity index between each reference compound and the unknown; the resultant indices allow ranking the candidate identifications according to their goodness of fit with the unknown. The chemist has the option to stop at this point and to accept the best fitting reference as identical

with the unknown if the similarity is sufficiently great. Alternatively, one can retrieve the complete candidate spectra if these have been stored. A manual inspection could then follow.

Creation of a binary code entails the selection of a lower threshold for m/e peak size after normalisation, and recording for each m/e value a "1" if a peak is present, a "0" if otherwise (Ref. 8, 9). This code allows an obvious saving of storage capacity by expressing a number of positions in a single binary word, and provides a means of comparing spectra through binary similarity indices (Ref.10). Tabulating matching peaks between an unknown and a reference spectrum corresponds to a logical AND operation on the binary codes; tabulating mismatches corresponds to a logical EXCLUSIVE OR (Ref.8). The binary code is thus well suited for calculations on the digital computer, permitting very rapid file searches. A binary code is also appropriate for recording and comparing derived spectral features such as ion series and spectral moments, and data additional to mass spectral information (Ref.6). Spectral abbreviation usually accompanies binary coding of discrete m/e positions.

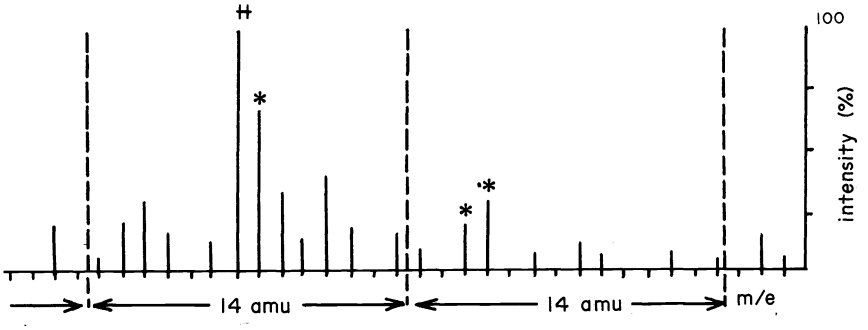

Fig.2 Spectrum abbreviation = the "2 most intense peaks per 14 amu window" rule.

Abbreviation serves to minimise the redundancy within a given mass spectrum stored in the library. Recording only the n most intense peaks in each spectrum leads to unsatisfactory results because these peaks often fall into a restricted, undiagnostic mass range. A widely used technique records the n most intense peaks in each non-overlapping interval of m amu's (Ref.11, 12). Commonly, n is set equal to 2 and m to 14, and the first interval is set to begin at m/e 6 to avoid splitting recurring peak clusters among adjacent intervals (Fig.2). This procedure usually ensures that the molecular ion - if present - will not be deleted from the abbreviated spectrum. An abbreviated spectrum may be recorded either as peak positions, or as positions and intensities.

Presearch techniques - called here "filters" - permit rapid rejection of spectra that are obviously not identical to the unknown. Dromey (13) describes a Series Displacement Index (SDI), an intensity-weighted measure of the displacement of spectral peaks from a reference ion series, in this case, that in the spectra of the alkenes. The SDI was found to characterise molecular class and hence allows a program to direct the search to a subset in the library. Discrete m/e information can also function as a filter. In a "key ion" strategy, the unknown spectrum is searched for ions or combinations of ions known to be characteristic of a library subset. Carried further, the use of key ions would describe a tree and could obviate the need for a spectral library (Ref.14). A filter based on the molecular ion has only limited usefulness because of the difficulty in obtaining a confident identification of the ion, and the decreased ability of the file search to suggest homologues if the unknown is not represented in the library (Ref.12).

Use of a key ion filter recognises the differential information content of m/e values, and reflects the experience of the spectrometrist. Following this approach, one can devise a similarity index such that comparisons between spectra are weighted according to the information content of each spectral feature. The weightings can derive from the spectrometrist's experience (Ref.6) or from study of the spectral library. McLafferty's Probability Based Matching system (Ref.5, 15) rests upon a tabulation of the number of reference spectra containing a peak for each m/e value in turn. In the subsequent calculation of a similarity index between any two spectra, each term - corresponding to a single m/e value - is weighted according to the data in the "uniqueness table". Most significance is given to matches at relatively unique m/e values. The similarity index also takes into account the rarity of particular ranges of intensity at each m/e position. In addition to probabili-

stic measures of similarity, one can devise indices as distance measures (Ref.16).

ARTIFICIAL INTELLIGENCE

"Artificial intelligence" is used here to describe programs that purport to imitate the mass spectrometrist's interpretation techniques. Mass spectral features are not treated as quantities to be compared and contrasted between spectra, but rather as the results of fragmentation processes. From the observed spectral patterns, the programs attempt to infer the fragmentation processes that these patterns reflect, and reconstruct the unknown structure. The usual storage and retrieval methods access a library of mass spectral features - a much condensed "library" of weights in the case of the learning machine described below - whereas artificial intelligence methods access a library of fragmentation rules. A brief description of the Stanford system is presented below (Ref.17).

LEARNING MACHINES

Learning machines offer a method of spectral identification that minimizes storage and time requirements, but most published applications lack the resolution needed to attach a specific name to an unknown. Development of a learning machine initially requires a large set of spectra and generous quantities of time on a large computer. Briefly, "training" of a learning machine begins with the chemist dividing the set of spectra among two or more categories, each characterised by a structural feature of interest. For each pair-wise combination of categories, a weighting vector is calculated; vector multiplication between this vector and the vector of m/e values of an unknown mass spectrum yields a scalar. The magnitude or sign of the scalar indicates to which category the spectrum best belongs (Ref.7, 18). A sequence of such decisions would be used to determine type and number of functional groups, number of carbons and presence of heterocyclic rings (Ref.19, 20).

Because the learning machine comprises an empirical derivation of spectral features that discriminate between specified categories, it could find application in constructing a search filter, based for example, on ions or indices. Such an application of the learning machine would be most appropriate when the chemist is faced with the task of creating a search system on a poorly-studied group of compounds, or one with which he has had little experience.

Jurs (20) utilises a learning machine that simultaneously calculates the weighting vectors and drops from consideration m/e values with low discriminatory power between the specified categories. Starting with a training set of 300 spectra, a total of 14 different values of m/e were found to correlate with the presence of oxygen in the parent molecule. (The highest m/e value found among the spectra was 195.) A rapid multiplication and addition of these 14 terms could successfully predict oxygen presence or absence in 90% of 330 test spectra. Note that this calculation requires fewer terms than an ion series calculation in a library presearch routine.

DESCRIPTION OF ENTIRE SYSTEMS

The foregoing discussion has stressed the computational elements of the three identification methods in relation to the task required and the important constraints of time and storage capacity (Table 1). The organisation of the library also determines time and storage requirements, but a full treatment of the topic of file structure lies beyond the scope of this paper. As an example of the optional range of structures, one can construct a spectral data base so as to allow direct access to spectra containing a given key ion through a list of such ions and a series of pointers to respective library subsets. A library can be so "inverted" on a calculated key, e.g. the Series Displacement Index (Ref.13).

TABLE 1 Options available

Task	Constraint	Methods
Recognition of class or partial structure OR Feature retrieval	Limited storage and time	Learning machine
	"Unlimited" storage and time	File search Artificial intelligence
Identify or confirm specific compound	Limited storage and time	File search with filters, abbreviation
	"Unlimited" storage and time	Complete file search

From these general considerations, the discussion will now turn to specific retrieval systems described in the literature. Presentation of each will be brief, intended to point out the distinctive features and the particular procedures involved in their implementation.

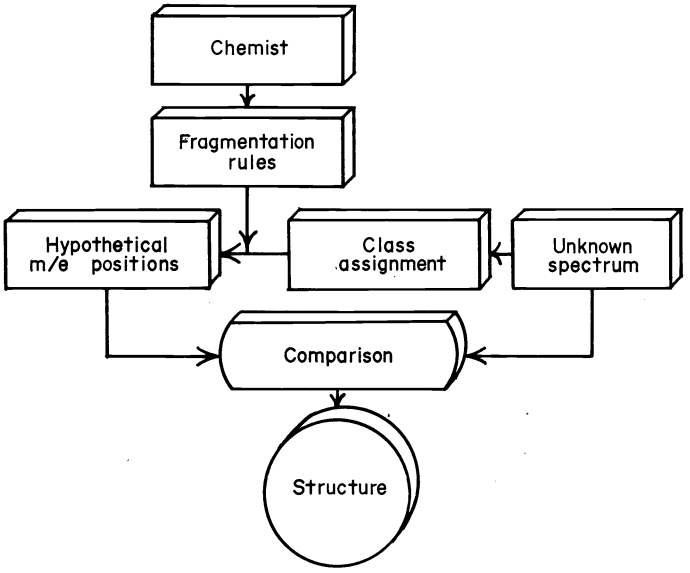

Fig.3 General procedure in heuristic, DENDRAL-like programs for mass spectral interpretation.

The DENDRAL program of the Stanford group (Ref.17) represents a collection of fragmentation rules; implementation of this and similar programs (Ref.21) begins with the chemist specifying the fragmentation pathways and rules characteristic of a given group of compounds (Fig.3). Confronted with an unknown, the DENDRAL-like programs may use the molecular ion to assign the unknown to a compound class. The program then accesses the characteristic fragmentation rules of that class and selects those consistent with the observed mass spectrum. The heuristic aspect of the program allows extension to any class once the chemist provides the fragmentation rules. Both high resolution (INTSUM, (Ref.21)) and low resolution (DENDRAL, (Ref.17)) mass spectral data afford interpretation.

Clerc (6) describes a library search method that minimizes the time required to compute the similarity indices. At the time of library creation, selected spectral features such as ion series and significant peaks are stored as a binary code, and empirically ranked according to discriminatory power. In comparing an unknown with a reference spectrum, the terms in the similarity index are computed from the most discriminating feature to the least; calculation of each term leads to a comparison of the cumulative matches and mismatches with a given threshold. If the deviation between the spectra falls above the threshold, the calculation of similarity between the two spectra is abandoned before compilation of all the terms, and the search moves on to the next reference spectrum. This algorithm acts as a filter, with the difference that the processes of filtration and the calculation of similarity occur simultaneously. Output comprises an ordered list of compounds in the reference library that best fit the unknown, together with the respective similarity indices. The system is intended for use with any compound class, and can also include data other than mass spectral features.

The Self-Training Interpretive and Retrieval System (STIRS, (Ref.22)) confronts the problem of identification when the unknown compound is not represented in the reference library. Fundamentally a library search method, STIRS compares the unknown with a reference spectrum through calculation of eleven "Match Factors" between the spectra. Each Match Factor utilises an individual spectral feature - e.g. ion series, primary neutral loss, or characteristic ion - that is known to correlate with structural attributes. Comparison of an unknown spectrum with a reference spectrum of the same compound will yield high values for the Match Factors. If the unknown is not represented in the reference library, then a list of the best matches accompanies each Match Factor; if a number of reference spectra in a given list share a structural attribute, the program indicates that the unknown also possesses this attribute. Manual inspection of these suggested structural features usually allows derivation of the unknown structure.

The Probability Based Matching system (PBM, (Ref.5, 15)) exploits the advantages of a probabilistic similarity index and a reverse search in the analysis of mixtures. The similarity index described above expresses the "confidence" that a given spectrum contains that of a particular reference compound. The so-called "forward search" found in most retrieval systems tacitly assumes that the unknown spectrum represents a reasonably pure, single compound, with allowances made for slight impurities and instrument error. This assumption is not made with a reverse search and any reference spectrum will be matched to the unknown as long as the unknown spectrum contains the peaks in the reference spectrum. Peaks left unaccounted for in the unknown spectrum are assumed to represent another component of the mixture; similarly taken into consideration is the augmentation of peaks in the unknown spectrum by ions due to any other components. The PBM system reduces the need for complete chromatographic separation and is suited to the direct analysis of simple mixtures. The output consists of an ordered list of compounds, the confidence of identification, and the respective contribution to the mixture. The method is not intended for detecting compounds unrepresented in the reference library; these problems are left to STIRS or other systems.

The Mass Spectral Search System (MSSS) offers a number of library retrieval methods, including retrieval on the basis of molecular weight, formulae, MSDC class terms and combinations of these (Ref.23). Identification methods presently include the Biemann approach, a straightforward search of a library of spectra abbreviated by the "2 most intense peaks per 14 amu" rule. The STIRS and PBM methods are to be implemented shortly (at the time of this writing). Using a library of 39,500 spectra representing 28,000 different compounds stored on large, structured disc files, the service is implemented on the Cyphernet system and accessed interactively in the United States and Western Europe by telephone. Costs include a $300 annual fee, a $9 fee per search, and additional fees according to retrieval options utilised. The system is not suited to the identification of large numbers of spectra, but offers the advantages of a large spectral library that is continually updated.

As a final example of a mass spectral retrieval system, the INTERP program in use at Bristol was designed with two aspects particularly in mind: the chemist often knows the chemical class of his unknown; and within that class, inspection of only a few, distinctive ions often allows structural or specific identification (Ref.14, 24). The chemist implements the method for a given class of compounds by providing a hierarchy of categories and tests for determining membership at each node in the tree (Fig.4). Identification of an unknown proceeds by working through the hierarchy, applying the tests - an average of 4 at each node, and continuing until no further tests are satisfied or a terminal node is reached. The program outputs the specific compound or group of compounds, corresponding to the node which best matches with the unknown, the number of matching peaks for the reference and unknown (i.e. the number of tests satisfied) and the ion current accounted for.

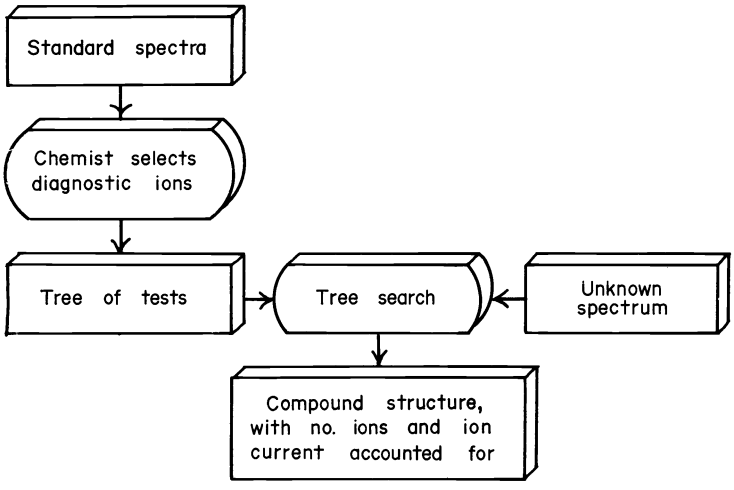

Fig.4 Steps in the identification of an unknown by the Bristol INTERP program.

CONCLUSIONS

A mass spectrum is the single most definitive data set readily obtainable in seconds from microquantities (ca. 10^{-9}g) of a compound. Analyses of many biological, medical and geochemical samples presently based on GC data alone would benefit from the greater confidence

in identification possible by GC-MS. The inevitable increase in the use of mass spectrometers as routine analytical tools must be accompanied by parallel developments in the ease, speed and certainty of computer-based identifications.

Developments in storage and retrieval systems have led to a diversification of methods. While on the one hand this diversification requires the chemist to choose among a baffling array of alternatives - forward search and reverse search; sequential files and trees; complete file search and inverted-file search with keys - it also means the application of computer-based systems to a range of problems, a closer tying-in of a specific technique to the problem at hand, and the incorporation of optional search and interpretation strategies into the same program. For instance, forward search and reverse search techniques are available for analysis of pure compounds and mixtures, respectively. Developing a given technique in the context of a specific problem permits effective use of the computational facilities and allows one to place more confidence in the information it retrieves.

TABLE 2. Systems for the computerised identification of mass spectra

Method	Computer	Computation time per unknown	Applications*
DENDRAL (Ref.17)	PDP-6	4-5 min.	Occasional
Clerc's method (Ref.6)	CDC6400/6500	10 sec.	GC-MS
STIRS (Ref.22)	PDP-9	20 min.	Occasional
PBM (Ref.5)	PDP-11/45	2 sec.	GC-MS
MSSS (Ref.23)	-	15 min.	Occasional
INTERP (Ref.14)	PDP-8/e	0.25 sec.	GC-MS

* For meaning of "Occasional" and "GC-MS" see section "Conclusions". These evaluations are those of the present authors.

The need to identify an occasional, problematical mass spectrum versus a large number of spectra that may be easy to recognise individually constitutes a fundamental dichotomy in the application of mass spectral information systems. Methods appropriate to the former situation generally require powerful computation and large library facilities, whereas methods satisfying the second need minimize storage and time requirements and are well-suited to the treatment of GC-MS data (Table 2). In many applications of mass spectrometry, an analysis involves a known class of compounds, either as a single compound or a mixture. This fact permits use of a small, laboratory computer for the identifications, and the choice between real-time or delayed processing as only a small, specialised library needs to be accessed. Unusual and unexpected spectra could be left to more detailed treatment by large file search or interpretive programs, probably on a main-frame computer or a network service.

Because of the uncountable millions of possible organic compounds, present libraries containing 10^4 spectra are clearly inadequate for any but the most commonly-studied compounds. The storage and organisational problems increase with high resolution MS and ionisation procedures that generate markedly different spectra, e.g. chemical ionisation and field desorption. The use of a single, small library or a set of specialised libraries accessed through a key may be mandatory in laboratories where the nature of the compounds, the type of spectrum, or the number of spectra generated precludes use of a service such as MSSS. Depending upon the reasons for creating a local data base, a choice exists between the chemist obtaining commercial and published collections of spectra, or producing the data himself. The first option offers the advantage that a sizeable library can be constructed with a minimal outlay of time and manpower (Table 3). Generating one's own data allows personal control over the validity of the spectra, the relevance of the collection to work underway in the laboratory, and the instrumental conditions during spectral acquisition.

The diversification of mass spectral storage and retrieval methods has not only meant a wider range of basic methods available, but has resulted in a refinement of methods as well, particularly in the development of "smart" techniques in library searching. Early search strategies treated each piece of data of a given type as though equivalent in information. Thus, a complete search of a library compared all of the reference spectra with the unknown, rather than utilising a filter to limit the search to a subset. Similarly, all peaks were weighted equally regardless of m/e value. Information theory suggests that m/e values should be weighted by their information content when calculating similarity between two spectra. "Smart" library search systems - employing filters, keys, inverted files, and m/e weighting schemes - provide savings in time and storage needs, can give more reliable retrievals, and indicate uncertainty in an identification. The future will see increased use of these techniques and greater attention paid to library organisation, requiring the close cooperation of

analytical chemists, information scientists and software specialists.

TABLE 3. Main sources of mass spectral data

Distributor	Description	Approx. No.of spectra	Notes
1. MSDC	Full spectra collection (tape)	18,000	1
2. MSDC	Full spectra collection (data sheets)	7,000	1
3. MSDC	Eight peak index - tape	31,000	1
4. MSDC	Eight peak index - book	31,000	1
5. Wiley Inter-science	Registry of mass spectral data - tape	23,000	2
6. Wiley Inter-science	Registry (McLafferty) of mass spectral data - book	18,000	2
7. Heyden	Compilation of mass spectral data (Cornu & Massot)	10,000	
8. NBS	NIH/EPA Collection - Tape	19,000	3
9. U.S. NTIS	Mass spectra of compounds of biological interest - book	2,000	4
10. MSDC	Mass spectra of compounds of biological interest - tape	2,000	1
11. TRC	Complete mass spectra (data sheets)	6,000?	5

Notes

1. Mass Spectrometry Data Centre, AWRE, Aldermaston, Berkshire.
2. Contains about 6,000 spectra from 2.
3. Contains about 3,000 spectra from 1.
4. National Technical Information Service, U.S. Department of Commerce, Springfield, Virginia.
5. Thermodynamics Research Centre, Texas A & M University.

Acknowledgements - The authors are grateful to Dr. A. McCormick for supplying data and useful comments, Mr. N.A.B. Gray for critically reading an early outline of this work, the Leverhulme Trust for the award of a Fellowship (to M.E.H.) and the Natural Environment Research Council for a Studentship (to M.J.H.). Computerised GC-MS studies at Bristol have been supported by the Nuffield Foundation, The Natural Environment Research Council (GR3/1695 and GR3/2420) and the National Aeronautics and Space Administration (Subcontract from NGL 05-003-003).

REFERENCES

1. J.McK. Halket and R.I. Reed, Org.Mass Spectrom. 10, 370-375 (1975).
2. R.G. Dromey et al., Analyt.Chem. 48, 1368-1375 (1976).
3. P.W. Brooks et al., Adv.in Org.Geochem., 1975 (in press).
4. F.P. Abramson, Analyt.Chem. 47, 45-49 (1975).
5. G.M. Pesyna et al., Analyt.Chem. 48, 1362-1368 (1976).
6. P.R. Naegeli and J.T. Clerc, Analyt.Chem. 46, 739A-744A (1974).
7. F.A. Mellon, Computerised data acquisition and interpretation. Mass Spectrometry Vol.3, Specialist Periodical Report, Chem.Soc., R.A.W. Johnstone, reporter, 117-142 (1975).
8. S.L. Grotch, Analyt.Chem. 43, 1362-1370 (1971).
9. S.L. Grotch, Analyt.Chem. 45, 2-12 (1973).
10. N.A.B. Gray, Analyt.Chem. 48, 1420-1421 (1976).
11. B.A. Knock et al., Analyt.Chem. 42, 1516-1520 (1970).

12. H.S. Hertz et al., Analyt.Chem. 43, 681-691 (1971).
13. R.G. Dromey, Analyt.Chem. 48, 1464-1469 (1976).
14. N.A.B. Gray and T.O. Gronneberg, Analyt.Chem. 47, 419-424 (1975).
15. F.W. McLafferty et al., Org.Mass Spectrom. 9, 690-702 (1974).
16. D.H. Smith, Analyt.Chem.44, 536-547 (1972).
17. A.M. Duffield et al., J.Amer.Chem.Soc. 91, 2977-2981 (1972).
18. T.L. Isenhour and P.C. Jurs, Analyt.Chem. 43 (10), 20A-35A (1971).
19. C.F. Bender and B.R. Kowalski, Analyt.Chem. 46, 294-296 (1974).
20. P.C. Jurs, Analyt.Chem. 42, 1633-1638 (1970).
21. D.H. Smith et al., Tetrahedron 29, 3117-3134 (1973).
22. K.-S. Kwock et al., J.Amer.Chem.Soc. 95, 4185-4194 (1973).
23. R.S. Heller et al., J.Chem.Info.Comp.Sci. 16, 176-178 (1976).
24. N.A.B. Gray et al., Anal.Lett. 8, 461-477.