lected, processed, and stored by a front-end computer. Selected files can then be sent upstream to the 11/40 system for processing, library searching, and long-term storage on the magnetic tape. Several display terminals located in user laboratories throughout the university can call into the system via a serially multiplexed phone-line modem. These devices enable the processing and output-

8. Stanford University Medical Center

T. C. Rindfleisch, D. H. Smith, W. J. Yeager, M. W. Achenbach, and A. Wegmann, Department of Genetics, Stanford University, Stanford, California

GLOSSARY

CLEANUP	Computer program to extract representative spectra for eluting components from GC/MS data.
DENDRAL	Set of computer programs using artificial intelligence techniques to assist in the elucidation of molecular structures.
FRAGMENTOGRAM	Ion current profile at a particular mass as a function of time or MS scan number.
Gaussian	Peak profile of the form $\exp(-x^2/2s^2)$.
GC	Gas chromatography.
GC profile	A collection of data relating to a GC/MS experiment, including the resolved spectrum for each elution component, the relative retention indexes for each component, the relative concentration for each component, and a name for each component.
HISLIB	Computer program for the historical quantitative comparison of GC/MS analyses of complex mixtures.
HRMS	High-resolution mass spectrometry.
LRMS	Low-resolution mass spectrometry.
MS	Mass spectrometry.
PKU	Phenylketonuria.
RRI	Relative retention index—elution time for a mixture component from a gas chromatograph, normalized relative to those for internal hydrocarbon standards.
TIC	Total ion current—instantaneous total ion current from a mass spectrometer source. "Areal" TIC is the time integral of the TIC resulting from the elution of a particular mixture component.
TIMSEK	Computer program to detect internal hydrocarbon standards and to compute relative retention indexes for other elution components in a GC/MS run.

a. Introduction

Since the pioneering instrumentation and applications work reported by many institutions in the first edition of this book (1), there have been numerous improvements in gas chromatography/mass spectrometry (GC/MS) methodology. These include, for example, improved ionization techniques, selected ion monitoring, and GC retention indexes (2, 3). Such advances in the power of GC/MS systems have been accompanied by significant improvements in the quality and ease of operation of available GC/MS instrumentation and by the more routine use of these analytic techniques for the study of biomedical and natural product samples. Indeed, screening of populations of samples has been contemplated and in some cases begun in areas such as metabolic studies, evaluations of environmental quality, and environmental impact of chemical agents. Recent applications in our own laboratories in the Stanford University Departments of Genetics and Chemistry have included qualitative and quantitative investigations of urinary metabolites from premature infants (4–6), sterols extracted from marine organisms (7–10), environmental samples (11, 12) and the relative amino acid composition of carbonaceous chondrites (13). Such applications, even on a small scale, result in prodigious amounts of data that are unmanageable for chemists to analyze thoroughly and effectively without computer assistance. Computer support is needed at all levels of the processing from raw data acquisition through spectral interpretation and structure determination.

The importance of the digital computer for data analysis in GC/MS applications was, of course, foreseen long ago. However, since the efforts documented in the first volume of this book, significant developments in the technology and economics of computer hardware and software have taken place (14) that have had a profound impact on the organization of computer support for laboratory instrumentation and on the effectiveness of program tools available. These trends have provided the key impetus of our GC/MS system development work over the past 6 years. We have focused on developing an integrated and reliably automated set of computer programs to assist routine laboratory use of GC/MS. Programs have been written to help with several phases of GC/MS data processing, including:

1. Acquisition and automatic reduction of low- and high-resolution mass spectral data acquired in GC/MS experiments.
2. Extraction of high-quality, low-resolution mass spectra of GC-eluted components free from background, GC column bleed, and interference from closely eluting neighbors.
3. Quantitation, library identification, and comparison of GC/MS profiles with results of earlier experiments.
4. Interpretation of mass spectral information consistent with other spectrometric and structural constraints to elucidate unknown molecular structures (see Chapter 7).

Each of these programs has been developed within a design philosophy that emphasizes reliable automation, use of commonly available laboratory computer systems where possible, and implementations that can be relatively easily exported to other laboratories. Our concern for "reliable automation" stems from two issues. First, raw data analysis programs should be "data adaptive" in that they should be able to operate correctly in spite of changes in instrument parameters, such as sensitivity, resolution, scan rate, and noise levels. They should be able to track such changes when feasible and flag for the operator intolerable excursions from normal operation. Second, data analysis programs inevitably encounter ambiguous situations where the outcome of the analysis will be substantially affected by minor changes in a signal threshold or other parameter setting. In using the computer to abstract "significant" information from the voluminous raw GC/MS source, the chemist may be lulled into a false sense of security and faith when using a well-running program. It is important that such programs be designed with a significant degree of introspective ability so that the chemist is properly warned when results are questionable or ambiguous.

We have also recognized the need and, indeed, obligation to construct our analysis software insofar as possible to be sharable with other laboratories. The high cost of software development and the advantages of data sharability among different laboratories make this goal seem obvious. However, considerable thought and effort are required to develop a generally usable program design and to minimize the temptation to use specialized local system features that may produce some increase in efficiency but that impede the export of complex programs. Exportability has been a key design element in each of the systems we have worked on, emphasizing the use of computers of commonly available types and sizes where possible, "standard" languages, and designs accommodating a broad range of instrument configurations and operating conditions. We have assisted the export and installation of our programs in numerous other laboratories.

The following sections give a brief overview of the current GC/MS hardware facilities in use in the Departments of Genetics and Chemistry at Stanford University* and a summary of our work on particular programs useful in the data acquisition and analysis phases of GC/MS applications. In this exposition, we give only a cursory description of the hardware and software portions of our systems that can be considered "standard" or similar to those operating in other laboratories. Rather, we concentrate on those aspects of our systems that we feel are unique and exemplify the automation principles outlined above. Developments of higher-level, structural interpretation programs are described in Chapter 7.

b. Current Configuration of Instrumentation

Our GC/MS laboratory systems have undergone substantial changes in both the areas of instrumentation and computational facilities. We have continued to operate the Finnigan 1015

*This report documents developments only in the Departments of Genetics and Chemistry. There are GC/MS systems also in use in other areas, such as the Departments of Psychiatry, Anesthesiology, and Civil Engineering, exemplifying the growth in the use of these tools over the past years.

low-resolution quadrupole* and the AEI MS-9 high-resolution magnetic instruments reported in the original volume of this book. In addition, we use a GC-coupled, Varian-MAT Model 711, high-resolution, double-focusing system. For the applications work described here, the Finnigan quadrupole and MAT 711 high-resolution systems have been the primary instruments used and serve complementary roles for acquiring low- and high-resolution GC/MS data.

In typical applications, the range of analytic tools from stand-alone GC to GC/LRMS (low-resolution MS) to GC/HRMS (high-resolution MS) is used hierarchically to study biochemical mixtures. This range of tools provides increasingly definitive insights, at the expense of difficulty and sensitivity, into the mixture components. Analyses and identifications that cannot be effected at one level can be studied with more powerful techniques, if warranted by the problem.

Our computation facilities are also organized into hierarchical systems, taking advantage of the ongoing advances in relatively low-cost minicomputers where possible. We use three separate stages of processing divided according to real-time instrument support, post data reduction and analysis, and more sophisticated data interpretation processes. Each machine serves a well-defined set of functions with each stage coupled to the next through data communication facilities. Based on the technology of the early to mid-1970s we use Digital Equipment Corporation PDP-11/20 machines at the real-time instrument interface, PDP-11/45's for data reduction and analysis, and the SUMEX time-shared PDP-10 system† for the more complex interpretation tools discussed in Chapter 7. There are separate computer systems for the Finnigan low-resolution and MAT 711 high-resolution instruments, each consisting of a PDP-11/20 data collection and instrument control machine connected to a PDP-11/45 data reduction machine. This hierarchical organization provides for a useful flexibility and asynchrony between the collection of raw data and subsequent stages in the processing, the latter of which may benefit from chemist interactions and the use of other sources of information for optimum analysis. The advantages of shared software between the systems should be obvious.

It should be emphasized that the PDP-11 laboratory computers provide the computational support required for routine application of these GC/MS systems, including the CLEANUP and HISLIB programs described in Sections c and d. The SUMEX facility is used for analyses of unknowns with the DENDRAL programs (Chapter 7), where inputs from several physical methods (e.g., GC/LRMS, GC/HRMS, NMR, or IR) may be combined with other structural constraints to hypothesize and evaluate alternative candidate structures.

c. Resolution of Spectra in GC/MS Profiles (CLEANUP)

i. INTRODUCTION

In biomedical applications of GC/MS, it is often important to be able to systematically isolate and identify minor components in the complex mixtures being analyzed. Because of instrumentation limitations, many of the mass spectra obtained from GC/MS analyses of such mixtures are markedly different from the spectra of the corresponding pure compounds. Differences may be caused by contributions from unresolved neighboring components due to incomplete GC separation and also from GC septum and column bleed and from background of the mass spectrometer itself. These extraneous contributions may severely distort the relative abundances of ions in the mass spectrum and contribute peaks that are not characteristic of the component being examined. Such spurious ion contributions compromise manual or automated (e.g., library search) interpretation of the data.

A program called CLEANUP has been developed for systematically extracting representative mass spectra ("resolved" spectra) of mixture components from GC/LRMS data. CLEANUP uses peak models derived directly from the raw data to locate individual eluted components and to produce spectra with relative ion abundances properly assigned among overlapping components in the eluate. Accurate spectral amplitudes are obtained by correcting for background and neighboring component contributions. By these data-adaptive corrections, components that are eluted within less than two spectral scan times of each

* The Finnigan instrument and data system have been relocated to the Rockefeller University in New York as part of Professor Lederberg's recent move there from Stanford. We are continuing to maintain the programs at Stanford used in that system and to assist in their export to other laboratories.

† SUMEX is a national computational resource funded by the NIH Biotechnology Resources Program to support a community of projects applying artificial intelligence techniques to biomedical research. Its facilities have replaced the ACME system which was used for much of the GC/MS work we reported earlier.

other can be detected and their mass spectra well resolved. One can routinely and reliably extract component spectra of high quality from GC/MS runs that enable more definitive library matching, easier human interpretation of unknowns, and even the addition of extracted spectra to a library as authentic spectra. The following illustrates the results obtainable by these methods and briefly summarizes the procedures used. A more detailed description of the method as well as a comparison with methods proposed by other authors can be found in the literature (15–18).

In a previous publication describing the CLEANUP program (15), we illustrated the ability of the program to extract, from raw data, spectra that compare very favorably with library spectra of authentic compounds. Such high-quality spectra can be obtained even in the presence of severe interferences from background and overlapping components. We present three additional examples here, selected because they illustrate the operation of CLEANUP for compounds relevant to other parts of this presentation.

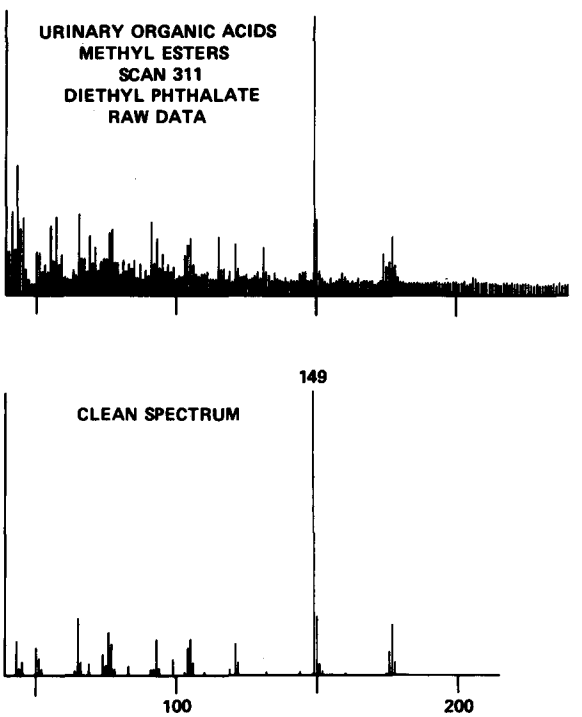

Figure 3-26. Result from CLEANUP for the first of two low-level, closely eluting (1.7 scans apart) components in the GC/MS analysis of the methyl ester derivatives of urinary organic acids from a patient suffering from phenylketonuria (PKU). This figure shows the raw (upper) and corrected (lower) spectra corresponding to scan 311. The extracted spectrum compares favorably to the authentic library spectrum of diethyl phthalate.

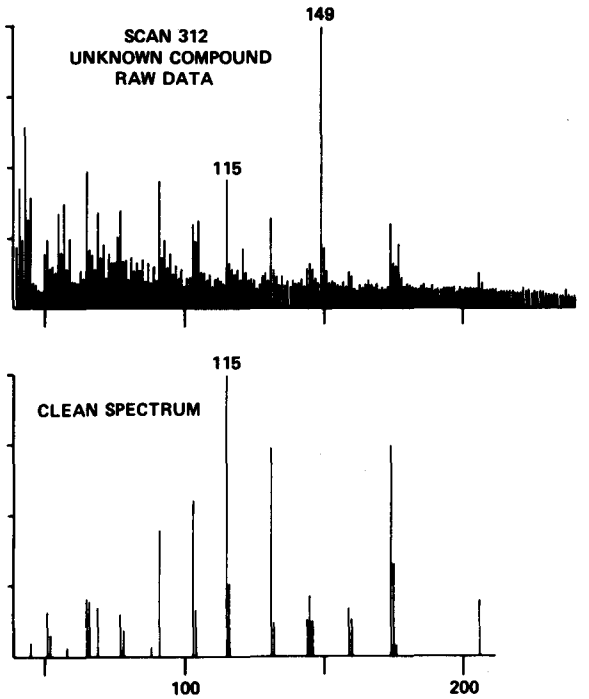

Figure 3-27. Result from CLEANUP for the second of the two neighboring components discussed in Fig. 3-26. The figure shows the raw (upper) and corrected (lower) spectra corresponding to scan 312. The extracted spectrum does not compare well with any authentic spectrum in our library but has been detected at similar concentrations in urines from both abnormal and control patients.

In Figs. 3-26 and 3-27 we illustrate CLEANUP's analysis of two consecutive scans of the methyl ester derivatives of urinary organic acids from a patient suffering from phenylketonuria (PKU), shortly after initiation of dietary control (a diet low in phenylalanine). The raw data for scans 311 (Fig. 3–26) and 312 (Fig. 3-27) are presented together with the respective "clean" spectra. For these low-concentration, closely eluting components (actual elution time difference is 1.7 spectral scan times), we obtain interpretable spectra in spite of the severe background and neighboring component distortions. The first component is identified tentatively as diethyl phthalate, whose authentic spectrum compares favorably with that illustrated in the lower half of Fig. 3-26. Such artifacts are routinely found in extracts of body fluids, apparently arising from plastic materials used in the collection of samples. The second component (Fig. 3-27) is an unknown compound whose identity has not been pursued further because the results of the HISLIB analysis (see Section d) indicated that this same component

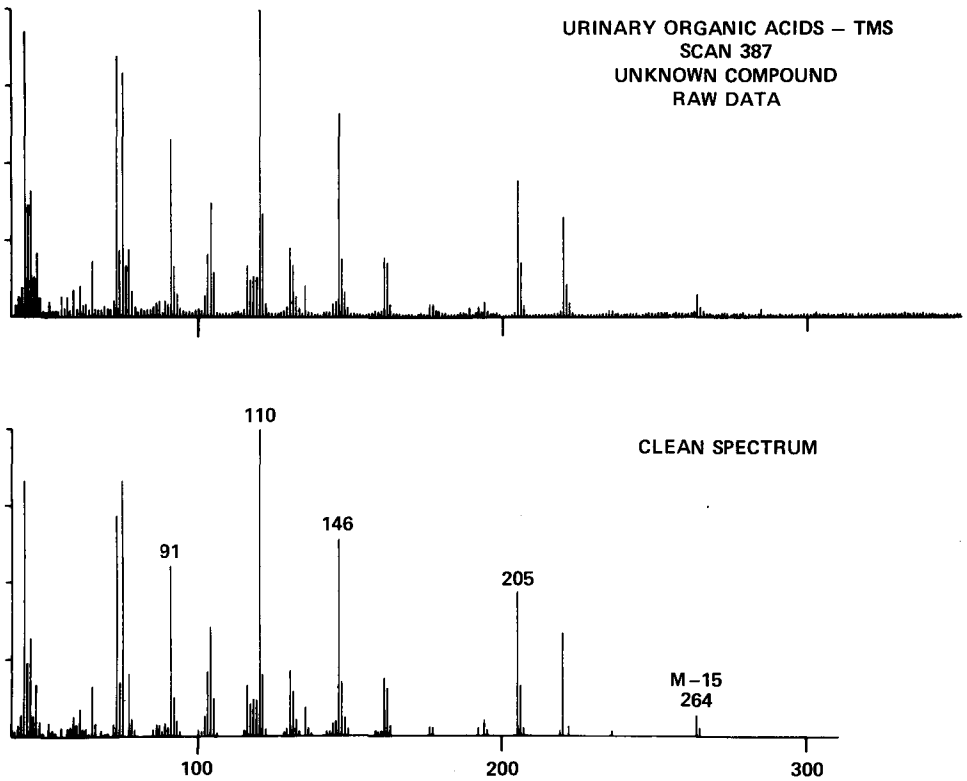

Figure 3-28. Example of the result of applying CLEANUP to the trimethylsilylated urinary organic acid fraction obtained from the same patient as in Fig. 3-26. The top spectrum represents the raw data and the bottom spectrum is the corrected spectrum for an abundant component at scan 387. (The further structural analysis of this material using the DENDRAL programs is discussed in Chapter 7.)

is also observed in low concentration in GC/MS profiles of organic acid urine fractions obtained from normal patients used as a comparison in this study.

The third example (Fig. 3-28) illustrates a result obtained from running CLEANUP on a GC/MS analysis of the trimethylsilylated urinary organic acid fraction isolated from the patient mentioned above. Fig. 3-28 shows the raw and "clean" spectra for a single component present in moderate concentration in the urine. The primary differences in the spectra are the absence of low-abundance background and column bleed peaks in the "clean" spectrum, together with slight variations in relative abundances of some peaks. In this example, the pattern of peaks important for structural analysis is clearly discernible in the raw data and it is important to note that CLEANUP is relatively transparent to spectra that are already of good quality. [The resulting spectrum did not match with high confidence any spectrum in our library of mass spectral data (19), however, and the DENDRAL programs were used to study the structure further. The spectrum of the methyl ester of this component provided information important to the solution of the structure, as described in Chapter 7.]

ii. DESCRIPTION OF METHOD

There are two key steps in the systematic resolution of GC/MS data. The first is to detect where in the GC profile each eluted component shows its maximum ion abundance and the second is to extract, from information in the regions about these maxima, representative spectra for each of the detected components. A basic assumption of our approach is that the mass spectra of two neighboring components can be distinguished; that is, there exist some masses (resolved singlets) for which ions occur in the mass spectrum of one component but not in the other, and vice versa. A schematic representation for two closely spaced materials is given in Fig. 3-29. By locating such "resolved" or "singlet" ion profiles at these masses (detected on the basis of the morphology of the mass fragmentograms), one can statistically

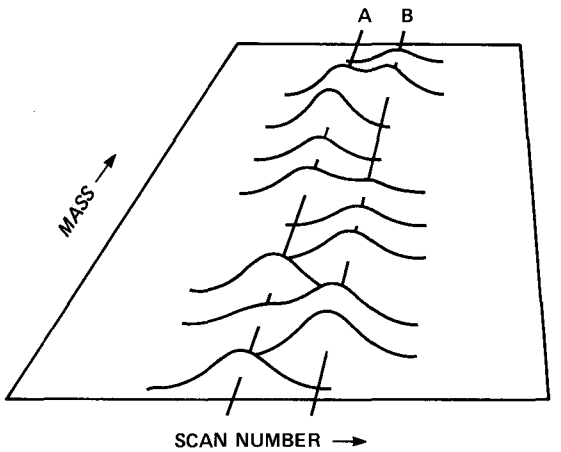

Figure 3-29. Schematic representation of a set of partial mass fragmentograms for two closely spaced eluting materials. Components A and B each have some ions with unique masses and others that are shared.

infer the positions of the components present and derive tabular models of the individual ion profile shapes and thereby the shapes of the GC peaks themselves. These models (one for each eluted component) can be used subsequently to separate the unresolved ion profiles for the mass fragmentograms of all other masses. The use of tabular ion profile models derived from the data accurately accommodates the a priori *unknown* peak shapes of particular components without solving for multiparameter, nonlinear model functions. Since the data are sampled often enough to satisfy the sampling theorem (20), these tabular models contain the information necessary to reconstruct a continuous ion profile envelope and can therefore be used as if they were continuous analytical models. For the typical GC peak and ion profile shapes encountered, the collection of 5 to 10 mass spectra over the duration of a singlet component peak represents a sufficient sampling frequency. In addition, the mass-by-mass analysis of the profile complexes facilitates the mass-dependent subtraction of background. (The large variation in background levels for different masses is a function of both the type of GC column used and the mixture being analyzed.)

(1) *Determining Locations for Eluted Components*

The detection of eluted mixture components involves finding the location of each material in the GC/MS data, even if it does not have a corresponding ion current maximum in the overall total ion current trace. Ideally, for a given component, the mass fragmentograms for all its ion masses will show maxima at the same time. In practice, this holds for well-resolved materials. However, for mixtures only partially resolved by the GC, peak overlap and background contributions can cause profile maxima for neighboring components to show significant variation in their apparent positions. For this reason, only singlet ion profiles are used for component position determination.

Candidate singlet profiles may be distinguished from doublet or background peaks by the feature that they are relatively sharp. We measure profile "sharpness" by averaging the magnitude of the profile's logarithmic derivative. By this definition, "sharpness" is independent of amplitude for peaks of identical shape. A peak with a computed sharpness below a threshold appropriate to the experimental conditions is considered to be either an artifact of the gas chromatograph (background peak) or a multiplet and is not included in the detection process.

We compute two histograms of candidate singlet ion profile positions and select as component locations those places where both histograms exhibit significant maxima. The first histogram measures the number of singlet ion profiles that reach maxima in each time interval. The second histogram measures the total singlet ion abundance above background at these maxima. These two types of histogram contribute complementary information for judging eluted component locations. Profile locations are measured to one-third of a spectral scan time and appropriate shifts are included to account for the fact that higher masses are measured at different times in each scan than are lower masses. This statistical approach, looking for "clusters" of ion profile maxima in the histograms, does not depend upon a correct decision for each profile but rather on a preponderance of good decisions looking over all the data. It will generally fail to resolve components that are eluted within less than about 1 spectral scan time of each other.

(2) *Calculation of Corrected Spectral Amplitudes*

Once the locations of components in the GC effluent have been determined, we proceed to compute a "resolved" spectrum for each material. The background (contributed by GC column bleed, MS background, and possible tailing from nearby high-concentration materials) is distinguished from the component signal by the fact that

it varies much more slowly with time. Reasonable estimates can be made by assuming that for any particular mass fragmentogram the background amplitude is constant in the vicinity of a given component. This approach to background determination, using the actual fragmentogram characteristics around each eluted component, automatically tracks changes in the background levels observed during a run. This zero-order background approximation is subject to some error. A more accurate approximation would involve representing the background variations over a larger span of spectral scans. The assumption of a locally constant background estimate is justified, however, in that it produces results within the error limits from other data uncertainties.

To complete the estimation process we use a model profile to determine the contribution of each mass fragmentogram to the spectrum of the eluted component. Much work has been done on the analytic approximation of GC peak shapes (21, 22). Our experience has been that relatively simple models do not adequately approximate the range of shapes encountered, and more complex models require large amounts of computing to determine model parameters. To obtain the profile shape and definition necessary for multiplet resolution within reasonable computing resources, we use tabular singlet profile models taken directly from the data. Such models, defined at discrete sample points, can be evaluated at any required intermediate point by interpolation [since the sampling theorem (20) is satisfied] and automatically reflect any peak asymmetries that may be present. For a given eluted component, the model will be independent of mass, assuming that relative molecular fragmentation probabilities do not change with pressure within the mass spectrometer.

During the process of computing the detection histograms, a list is kept of the sharpest, unimodal ion profiles in the region under analysis. When a component is detected in a given region, a model profile is then immediately in hand that can be used in the ion abundance estimation and background removal process. The local minima just on either side of the model profile are used to estimate and remove the local background under the model. The selection of the sharpest profile as our model has worked well in producing models that are singlets and suffer least from interference by background and neighboring fragmentogram peaks.

Given singlet ion profile models for each eluted component, the individual mass fragmentograms can be corrected and the true mass spectral intensities for the components estimated. For the fragmentograms exhibiting peak maxima "near" the locations of the detected components, the ion profiles at the various masses are aligned on a common time origin to account for the time difference between collection of low- and high-mass data. Then, assuming a constant background, b, over the region of 5 to 10 scan intervals under consideration and letting P_t, Q_t, R_t, ... represent the interpolated component profile models at times t (after normalization to unit area), the amplitude of the actual fragmentogram profile Y_t at time t can be approximated by

$$Y_t = pP_t + qQ_t + rR_t + \cdots + b \qquad (1)$$

where $p, q, r, \ldots$ measure the eluted component amplitudes above background. Note that this model assumes a superposition principle based on the assumption of constant relative fragmentation probabilities and a linear encoding of ion current information. If ion current data are obtained from nonlinear electronic systems or read from film, the peak model itself would be amplitude-dependent and this linear analysis could not be applied until appropriate amplitude linearization corrections were made. From the model above we can derive an estimate for the component amplitudes $p, q, r, \ldots$ and the background amplitude b by standard least-squares procedures. It is worth noting that this method, using tabular peak models and eluted component locations obtained from the detection histogram analysis, reduces the calculation for each mass spectrum intensity to the solution of a set of linear equations. Specifically, this avoids iterative methods for determining the parameters of a theoretical peak model and for determining component time positions.

Fragmentograms are selected for this analysis on the basis of several criteria. Given the nominal eluted component positions from the detection histogram analysis, a fragmentogram is excluded (mass spectrum assigned zero intensity at that mass) if it has no significant local profile maxima or, for singlet components, if its maximum is displaced from the detected component position by more than two-thirds of a spectral scan time on either side.

In practice, we have not implemented this full procedure beyond the doublet case. Through the following approximations, reasonable results are

achievable within available minicomputer resources. Using the histogram method described earlier, neighboring components are handled with a "look-ahead" procedure. That is, information about a component that has just been detected is stored and the detection algorithm is applied to the data in the immediate neighborhood by extending the range over which the detection histograms are calculated. If by including this extended region an additional component is detected, we record the position of its mode, select a model profile for this second component using the sharpness criterion, and initiate a doublet resolver algorithm. At present, the extended histograms project six spectral scan widths beyond the position where the first eluted component of the multiplet was detected (limited by computer memory).

The doublet model represents an oversimplification of the case of more complex multiplets, but by applying it to successive pairs of eluted components (taking first-order account of peak tail contributions from any earlier component), it provides acceptable accuracy and peak resolution effectiveness. Amplitude results for masses that belong to the second component of the doublet are stored temporarily until this component is moved into the processing window, at which time they are incorporated into the analysis of the newly detected component.

iii. CLEANUP RESULTS AND LIMITATIONS

The program based on the algorithm outlined in the preceding sections has been tested on a wide variety of biological samples. It fits comfortably into a DEC PDP-11/45 computer (with 32K words of memory) and takes approximately 5 to 10 min to analyze a raw GC/MS data set of 600 mass spectra (scanned from masses 40 to 450). Much of this time is spent in reading the raw data from the disk and other input–output operations. Copies of the program, written in FORTRAN, are available from the authors. Currently, this program runs as part of an automated analysis system for the GC/MS analysis of urine and blood samples. The program typically reduces the raw GC/MS data set of approximately 600 spectra to a set of about 60 resolved spectra for detected eluted components that are then matched against a library of mass spectra of biological compounds (19). This whole process takes about 20 min and produces an analysis of the sample, with known compounds in the mixture identified and the remaining unknown set marked for further study by chemists or other programs such as HISLIB (see Section d) or DENDRAL programs (see Chapter 7).

In evaluating the performance of the program, a major issue is how well it is able to detect eluted components in the data. The detectability of unresolved components is clearly a function of their amplitude relative to neighboring components and background. One way to characterize this is to measure the ratio of the total ion abundance (sum of the mass spectrum amplitudes) in the resolved spectrum compared to that in the unprocessed spectrum, including background and overlap effects. In our experience, the program has difficulty detecting components with this ratio much below about 5%. Also, if two compounds are eluted within less than 1 spectral scan time of one another, the program will probably detect the presence of more than one component, but there is an increasing chance that it will fail to resolve truly representative spectra for the components. Such errors are dependent on the ion current ratio between adjacent components, the similarity of their mass spectra, and the reliability with which peak positions can be determined.

d. Quantitative Comparison of GC/MS Profiles (HISLIB)

i. INTRODUCTION

In routine applications of GC/MS to study complex mixtures of organic compounds, one frequently needs to compare quantitatively current results with those obtained earlier. Such comparisons are important, for example, in (a) validation of analytical procedures used to isolate components in complex mixtures; (b) development of historical libraries that might include complete summaries of past observations, compilations of controls, or other selected subsets of results; (c) computation of average mass spectra and relative retention indexes (RRIs) of known compounds to improve the quality of existing libraries of mass spectral data; or (d) rapid comparison of new data to previously compiled library(ies) to detect differences in kind and/or amounts of individual components. In this section, we describe a program, called HISLIB, designed to automate the task of quantitative comparison of GC/MS profiles obtained on complex mixtures of organic compounds.

We define a "GC/MS profile" for a GC/MS experiment as an assembly of data consisting of (a) the (unnormalized) spectrum of each component after component detection, background removal,

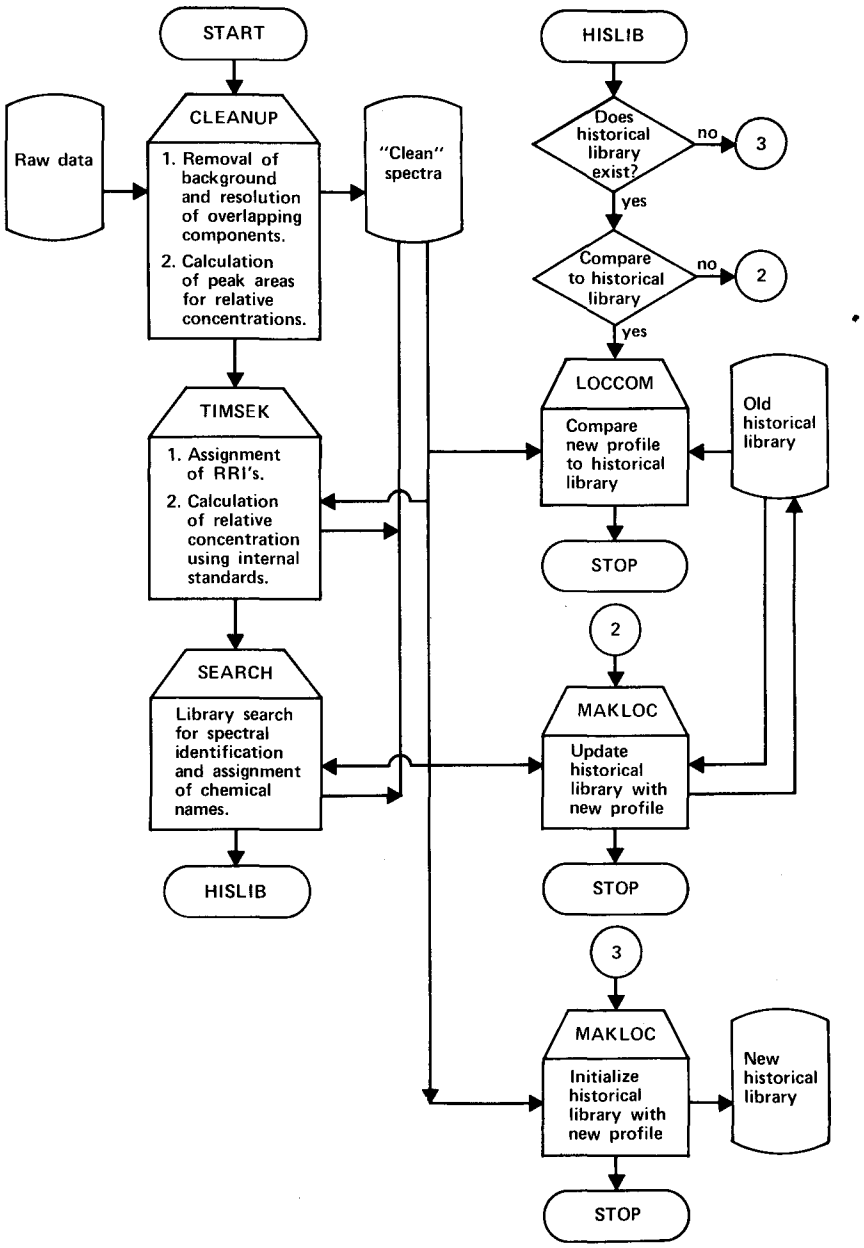

Figure 3-30. Major steps in processing a complete set of GC/MS data to establish and search an historical library.

and resolution of overlapping components; (b) the retention index of each component; (c) the relative concentration of each component; and (d) (optionally) a name for each component that may be a simple experiment code or a name associated with the component during routine library search.

This program is a logical synthesis and extension of several software tools (Fig. 3-30) and compares profiles of new mixtures with historical libraries of GC/MS data on related mixtures. Co-occurrence of components is established by matching both retention indexes and mass spectra, after processing by CLEANUP (see Section c), with the profiles in a historical library. Quantitation is done by comparing relative concentrations of components, calculated using internal standards. A detailed report of the methods involved and a comparison with other published work are given in the literature (16–18).

We use HISLIB after a number of preprocessing steps have been applied to experimental data to maximize the quality and specificity of the

extracted information. Because library matching, determination of RRIs, and, particularly, measurements of relative concentrations depend strongly on spectra free from background and overlapping components, we first process the GC/MS data with the CLEANUP program (15). Next we determine RRIs for each detected component to improve the specificity of library matches (23–25) and compute relative concentrations based on one or more internal standards. We then match each spectrum against an existing library of mass spectral data, in our case a library of compounds of biological interest (19). Finally, the resulting data are combined with previous results to update the historical library or are compared against an existing historical library. The flow of data through these steps is summarized in Fig. 3-30.

ii. PROCESSING STEPS FOR PROFILE COMPARISON

(1) *Automatic Determination of Relative Retention Indexes (RRIs)*

We use an extension of the method proposed by Nau and Biemann (23, 24) for determination of RRIs. The procedure is automatic and calculates reproducible RRIs under variable instrumental and experimental conditions arising from unavoidable changes in initial GC column temperature, carrier gas flow, or temperature programming rates. It requires only three internal hydrocarbon standards for the analysis of a GC/MS run.

As previously described (23, 24), each column is calibrated with a mixture of 18 hydrocarbons in the range n-C_{10} to n-C_{28}, thereby relating carbon numbers to mass spectrometer scan numbers. Subsequent GC/MS experimental runs using that column are processed using these calibration data as reference. Three of the hydrocarbons used in the calibration run are added to each experimental mixture. The CLEANUP program is run to isolate representative spectra and to assign scan numbers corresponding to elution times for each component. The TIMSEK program (Fig. 3-30) then locates the three added standards by matching their known spectra in windows about the expected elution scan numbers and fits the three observed hydrocarbon scan numbers to those corresponding in the calibration run. We assume that differences in conditions between a given experimental run and the calibration run can be accounted for by a linear transformation of the elution time scale,

$$S_{cal} = AS_{exp} + b \qquad (2)$$

where S_{cal} is a scan number in the elution time scale of the calibration file, S_{exp} a scan number in the elution time scale of the experimental run, and A and b the linear transformation coefficients. We determine A and b by minimizing the difference between the elution times of the three standards in the experimental and calibration runs using least-squares techniques.

Once A and b are determined, (2) is used to determine the effective scan number for eluted components in the experimental run as transformed to the calibration run time scale. These effective scan numbers are converted to RRIs by a linear interpolation or extrapolation using the nearest hydrocarbons measured in the calibration file (23, 24). (If the GC is operated isothermally, a logarithmic interpolation/extrapolation is used.)

This method differs from that of Nau and Biemann in that the least-squares fitting procedure takes explicit account of both linear offsets and expansion or contraction of the scan number/retention index curve rather than simply optimizing about the midpoint of the range (23, 24).

(2) *Determination of Relative Concentrations*

Relative concentrations are determined by TIMSEK (Fig. 3-30) based on any one or combination of the internal standards selected by the user prior to obtaining GC/MS data. Ideally, standards should be chosen that reflect the kinds of compounds one wishes to quantitate, the variety of analytical procedures used to isolate mixtures to be analyzed, the sensitivity of spectra to changing MS conditions, and other considerations that affect accurate and reproducible quantitation using any analytical procedure. We wish only to point out that care must go into the selection and use of such standards. TIMSEK uses a preestablished library of spectra of standards together with their RRIs. The standard(s) selected is searched for in the GC/MS data by looking for the closest spectrum match (5 below) within a narrow retention index window (±0.2 methylene unit). This is similar to the method of Sweeley et al. (25). Having found the internal standard(s), the relative concentration, ρ_i, of the ith component is calculated:

$$\rho_i = 100 \frac{\text{areal TIC of } i\text{th component}}{\text{areal TIC of internal standard}} \qquad (3)$$

The "areal" total ion current (TIC) measures the

area of the GC peak of the ith component, not simply its height. The area for each GC peak is derived from the raw mass spectral data using the peak model determined for each spectrum during CLEANUP (15). The intensity (ion abundance expressed as peak height) of each mass in the spectrum of the ith component is determined by fitting the data-adaptive peak model to the intensity profile for each mass (fragmentogram) about the position of elution of the component (15). Simpson's Rule is used to determine the area of the model peak. The areal total ion current for the ith component is then computed:

$$\text{areal TIC}_i = \frac{A_{i(\text{model})}}{h_{i(\text{model})}} \sum_m I_{im} \qquad (4)$$

where $A_{i(\text{model})}$ and $h_{i(\text{model})}$ are the area and height of the peak model for the ith component and I_{im} is the ion abundance (peak height) at mass m in the mass spectrum of the ith component after processing by the CLEANUP program.

If more than one standard is used, the basis for relative concentrations is the average of the areal total ion currents for the standards. The inclusion of multiple standards provides the opportunity for a better statistical basis for computing relative concentrations, since statistical fluctuations in measuring the areal TIC of one are reduced by averaging with the areal TICs of the others. Depending on the relative quantities and reproducibility of the various standards included, a weighted average may be appropriate to account for different relative a priori uncertainties in the TICs among them. In our case, these are comparable and a straightforward average is used. An improvement in quantitation standard reproducibility can be expected increasing approximately as the square root of the number of standards included.

(3) *Assembling an Historical Library of GC/MS Profiles*

An historical library is assembled by HISLIB by taking the GC/MS profile from an experiment and adding it to the library. If the library is initially empty, the profile becomes the library. If the library already contains at least one profile, the new profile is added as follows. Each spectrum in the new profile is compared to each spectrum in the library within a narrow retention index window (e.g., ±0.2 methylene, or ±20 RRI, unit for our work). A spectral match score is calculated:

$$\text{spectral score} = 1000 \frac{\left[\sum_m e_{m(\text{prof})} e_{m(\text{hist})}\right]^2}{\sum_m e^2_{m(\text{prof})} \sum_m e^2_{m(\text{hist})}} \qquad (5)$$

where spectra are reduced to the two most abundant ions every 14 u (26) and the spectral intensities are encoded before matching. The terms $e_{m(\text{prof})}$ and $e_{m(\text{hist})}$ represent the encoded intensities at mass m for the new profile and the historical library, respectively. They are quantized to have values 0, 1, 2, or 3, corresponding to the relative intensity ranges 0 to 4, 5 to 16, 17 to 64, and 65 to 100% of base peak, respectively.

The definition in (*5*) has several useful properties, based on Schwartz's inequality (27). The spectral match score calculated is independent of the order in which spectra are compared. If two ions of the same mass are present, a positive contribution to the score results. More abundant ions are weighted more heavily because of the squared term. The score is guaranteed to be between zero and 1000, 1000 representing a perfect match. Equation (*5*) is similar to the "degree of coincidence" score used by Jellum et al. (28), except that (*5*) uses encoded peak heights rather than just the number of peaks.

The spectral match score and the proximity of the retention indexes are combined through an heuristic evaluation function (*6a*) that yields the final score. This "final score" is the spectral match score weighted by a trapezoidal function (*6b*) that penalizes for disparate RRIs. The weight is unity if the difference in RRIs is less than 5 units and decreases linearly to a threshold weight as the absolute difference in RRIs becomes greater than 5 units up to the empirical cutoff of 20 RRI units:

$$\text{final score} = (\text{spectral score}) \cdot W(\delta\text{RRI}) \qquad (6a)$$

where $\delta\text{RRI} = (\text{RRI}_{\text{exp}} - \text{RRI}_{\text{lib}})$ and RRI_{exp} and RRI_{lib} are the relative retention indexes for the experiment and library components, respectively. The weighting function, $W(x)$, is defined by

$$\begin{aligned} W(x) &= 1 && |x| < 5 \text{ RRI units} \\ &= 1 - \frac{\text{maxscore} - \text{minscore}}{15\ \text{maxscore}} \times \frac{(|x| - 5)}{} && 5 \leq |x| < 20 \\ &= 0 && 20 \leq |x| \end{aligned} \qquad (6b)$$

where maxscore = 1000 and minscore = 400.

If this final score exceeds 400, the experiment compound is considered a potential match to the library compound. If there is more than one potential match between closely eluting experimental and library compounds, the ambiguity is resolved by a procedure (see below) that maximizes the overall correspondence between the pattern of experimental and library components. The minscore value of 400 was derived empirically by examining the distribution of scores obtained by matching spectra in chemically related subsets of our library (19) with the other spectra in that subset.

(4) *Assignment of New Spectra to the Historical Library*

The last step in correlating a new profile with an historical library (Fig. 3-30) involves selecting between alternative matchings of experimental and historical library spectra with similarly high final scores. This occurs frequently among isomeric compounds with similar spectra and retention indexes, and accidentally as, for example, among compounds whose spectra are similar owing to domination of the fragmentation pattern by ions from a functionality added during derivatization.

We have implemented a pattern-matching procedure to resolve such ambiguities. Briefly, the procedure attempts to maximize the consistency between a new experiment and the library, assuming that they are derived from similar mixtures. In a region containing ambiguities, a matrix representing every possible correspondence between experiment and library spectra is analyzed using an algorithm that can trace and rank all self-consistent "paths" through the matrix (29). Such paths include those which create new entries in the historical library, that is, paths with some spectra in the new profile not being matched to any existing spectrum in the historical library. Consistency constraints on the assignments include: (a) the scoring threshold must be exceeded for a match to be considered, (b) RRI order must be preserved, and (c) a spectrum in either set can be assigned to at most one spectrum in the counterpart set. Finally, the "best" assignment is that which has the highest total score, where the total score is the summation of scores (*6a*) for each candidate pairing of spectra between the two sets (the score is not incremented for a spectrum found to be only in one set). This procedure is driven strongly toward maximum overlap between the two sets of spectra. This is justified when the matching threshold is high enough to reject dissimilar spectra and the GC/MS profiles are from related mixtures.

Once specific assignments have been made, spectra from the new profile are added to the library. New entries are created for components that scored less than "minscore" against library entries, or that were assigned as new entries by the above pattern-matching algorithm. When a pairing with an existing library entry is made, the new spectrum is averaged with the library spectrum for that entry, effectively weighting each contributing spectrum by its total ion current. At the same time, the new relative concentration and retention index are averaged with the previous values. Note that an important advantage of this approach is that components need not be identified by name, only by occurrence in terms of RRI and mass spectrum (30).

(5) *Comparing New Profiles to the Historical Library*

Once a suitable historical library has been prepared, subsequent profiles can be compared to it to detect similarities and differences (Fig. 3-30). In practice, we use the same program used to assemble the library to perform the comparisons, changing only a flag that prevents using the new data to update the library and that causes a summary output to be produced indicating the results of comparison. Individual users may select different formats for such a summary. The one used in our laboratory was chosen to focus the attention of the user on components observed in significantly different relative concentration and on new components present in the profile regardless of relative concentration. Reference 16 contains an example of results for such comparisons.

iii. APPLICATIONS AND LIMITATION

The facilities provided by HISLIB suggest many types of applications. Examples include checking on the reproducibility of variables involved in instrumentation and analytical procedures used to study complex mixtures and detailed intercomparisons of complex mixtures such as those encountered in diagnostic medicine, where enhancements of GC/MS techniques are desirable (31). For example, we have presented examples of the use of HISLIB to evaluate two isolation procedures of organic acids from human urine (16). Since that time, a third, simpler method has been developed and HISLIB was utilized to compare the new

method with the previous two (6). The program quickly provided results on the precision of replicate extractions for each method and on components isolated more effectively by one procedure than the others.

Because the historical library can be updated at will, it is easy to maintain a long-term history of analyses of a particular type of mixture. Maintenance of several such libraries for different types of mixtures is a simple task. In fact, different historical libraries can be compared with one another, opening the possibility for comparison of results among laboratories engaged in similar research.

HISLIB averages spectra of the same compound. Thus statistical variations in ion abundances are reduced as additional examples are encountered. The resulting average spectrum is frequently of much higher quality than single spectra in existing libraries, and mechanisms have been implemented for adding averaged spectra to or replacing spectra in our primary library. This provides a procedure for gradual improvement of spectral libraries with time. In addition, RRIs are included with the spectra, improving the specificity with which subsequent spectra are matched to the primary library.

The method of comparing new profiles to an existing historical library quickly focuses attention on known materials present in abnormal quantities and on new components. The latter become subjects for more sophisticated structure elucidation procedures (32) that can now use the (high-quality) mass spectral data directly to assist in solving the structures of unknowns (33).

There are several limitations to the current implementation of the HISLIB procedure that should be mentioned. We have not yet thoroughly investigated variations in relative concentrations with instrument operating parameters. The performance of any mass spectrometer may change as a function of time. Any change in performance that affects the ionization of the internal standard(s) relative to other mixture components will affect results of quantitation. This can be avoided in part by using several different standards in each run.

The spectrum-averaging scheme makes no decisions about including ions of low abundance—all are included. Ions that occur infrequently are diminished in importance as additional spectra are averaged, but they are not rejected because we have not yet developed adequate heuristics for removing such ions.

e. GC/HRMS Data System Design

i. INTRODUCTION

Low-resolution MS is a very powerful tool for the analysis of biological materials, particularly when coupled with effective computer support to minimize tedious and error-prone manual data processing. However, detailed identification of mixture components depends upon satisfactory library matching of spectra or upon inferring the structure from spectral clues and other information that may be available from a knowledge of the origin and preparation procedures used on the sample. When novel components are encountered, such identification procedures may not succeed and one must seek other sources of information to determine the structure of the eluted component. One very useful refinement is the use of HRMS, which can give precise information about the elemental compositions of spectral fragments (34). These can greatly assist the chemist's process of structural inference or can serve as an input to other computer programs designed to infer biomolecular structures (see Chapter 7). One of the virtues of HRMS is that it can also be used in conjunction with GC so that little additional sample preparation is necessary. One might consider using GC/HRMS routinely in place of GC/LRMS, but the sensitivity advantages, greater instrumental simplicity, and lower cost of GC/LRMS quickly relegate GC/HRMS to more special studies of structural unknowns.

Data systems for photoplate and electronically scanned HRMS systems have been developed for many years (1) requiring various levels of operator intervention to assure accurate data reduction. The combination of GC with HRMS places special emphasis on reliable automation of the data system and analysis procedures because of the very large amount of information collected and the complexity of data reduction (35). The following sections give summaries of several aspects of our GC/HRMS data system that we have found to be important for reliable, automated operation. These include optimum data sampling rates, data-adaptive peak detection and multiplet resolution, pattern-driven reference peak location and scan modeling, rapid elemental composition matching, and effective operator feedback for monitoring of instrument performance and selective scan reduction.

ii. CRITERIA FOR OPTIMUM DATA SAMPLING RATE

One of the most frequently encountered limitations of GC/HRMS in biomedical applications is its sensitivity. For electronically scanned instruments, the determining factors for sensitivity include the amount of material that can be introduced into the mass spectrometer via the gas chromatograph, the ionization and focusing efficiency of the ion source, the resolution at which the instrument is operating, and the time that can be spent integrating the ion current at the sensor for each sample point in a mass spectrum. Assuming optimized gas chromatograph and source operation, the elution rate of the GC defines the time scale (scan rate) over which spectra must be collected to measure a particular mixture component. In turn, the ion integration time is determined by the instrument resolution, scan rate, and the required electronic sampling rate of the ion detector output. To maximize the ion integration time (and sensitivity), the sampling rate should be as low as possible. However, it must be high enough to encode the mass peak shape information needed for centroid calculation and mass multiplet resolution. A typical operation mode of many HRMS data systems is based on sampling rates that guarantee at least 20 samples across each mass peak (35, 36). Our analysis and experience shows that this criterion is too high by more than a factor of two and would result in a significant loss of sensitivity in applications involving GC/HRMS, in addition to placing a very large data rate burden on the data system computer. Based on the results summarized below, we have been operating our high-resolution system with a sampling rate, r, given by

$$r > \frac{11.5R}{t_{dec}} \qquad (7)$$

where R is the spectrometer resolution (5% peak width) and t_{dec} the exponential scan time per decade. This sampling rate is equivalent to requiring 5 to 10 sample points across a mass peak and assures, for trapezoidal or Gaussian-like peak shapes, that the peak envelope may be accurately reconstructed from suitable data interpolations.

A key issue in achieving these results is the type of ion detection system used. On many systems the output analog signal from the ion multiplier is lowpass-filtered and sampled periodically with a sample-and-hold analog-to-digital converter. The bandwidth of the lowpass filter should be chosen appropriate to the sample rate. It must be high enough so as not to reduce instrument resolution and low enough that ions arriving since the last sample measurement contribute as much weight as possible to the next one. When a simple RC filter is used (typical of many commercial systems), these criteria are hard to reconcile and typically the bandwidth is adjusted for acceptable resolution, which means that ions arriving early in the sample interval will have a lower weight than those near the end. As the sampling rate changes, this bandwidth must also change to preserve the "optimum" trade-off. We believe this issue accounts for the higher sampling rate requirements published earlier (36), where an analysis compared mass errors in an experiment run obtained by reducing the data using every sample point collected at 10K samples/s and by reducing the same data using every fourth sample point. There was no apparent adjustment of the effective filter bandwidth for the lower-sample-rate case to take advantage of the improved ion statistics available. This, combined with the limited amount of data analyzed, accounts for the conservative result published.

In contrast, our system uses an "ideal integrator" that linearly integrates the signal contribution for each ion detected between samples and, after its output is measured for a given sample point, is reset to zero. Thus each incident ion contributes maximally to the digitized signal, no matter when it arrives between samples (no filter rolloff), and each sample is independent of the others. Since the detector integrates uniformly between sample points and is reset to zero at the start of each interval, there is no blurring of adjacent samples, and the effective bandwidth of the "ideal integrator" automatically adjusts for different sampling rates.

One can intuitively sense that an "optimum" sampling rate exists for particular instrument performance parameters from the signal behavior at the extremes of low and high sampling rates. For very low rates, peaks will be encoded into a single integrator cycle or sample point. The location of the actual peak is subject to a large error, since the peak may have occurred anywhere within the integrator time and the same output would have resulted. Also, such single sample peaks would be very difficult to separate from other background noise and many false detections would be made. On the other hand, for very high rates, a peak will

have many samples taken in its course. Each sample will be the result of integrating the ion current for only a relatively short time and hence will have an amplitude subject to higher ion "shot" noise errors the higher the sampling rate. For a given peak ion current and at high-enough sampling rates, there may be a significant likelihood that too few ions arrive for some samples so that they will dip below the background noise and the peak envelope will be fragmented into two or more separate peaks. Between these two extremes lies an optimum sample rate for encoding the peak location and shape information.

Analytically, if the ion detector output is measured with an ideal integrator, the centroid errors introduced by sample spacing and by ion statistics are defined by

$$E_c = \frac{\int t f(t-t_0)\,dt}{\int f(t-t_0)\,dt} - \frac{\sum m\,\delta t(\mu_m + N_m)}{\sum (\mu_m + N_m)} \qquad (8)$$

where $f(t-t_0)$ = continuous peak envelope (centered at t_0) measuring the expected ion current distribution

δt = sample interval

$\mu_m = \int_{(m-1/2)\delta t}^{(m+1/2)\delta t} f(t-t_0)\,dt$ = expected number of ions arriving during interval m

N_m = deviation between the actual number of ions collected during interval m and the expected number

We define the area A to be the total expected number of ions in the peak,

$$A = \int f(t-t_0)\,dt = \sum \mu_m$$

and assume that the ion fluctuations arise from Poisson arrival statistics so that

$$\langle N_m \rangle = 0$$

$$\langle N_m N_n \rangle = \begin{cases} 0 & m \neq n \\ \mu_m & m = n \end{cases}$$

where the sample independence implied by the ideal integrator has been used. Equation (*8*) may be expanded and the expected error and variance in the centroid calculations estimated by averaging over peak locations and ion arrival fluctuations. Keeping terms to second order in the ion noise, one can deduce that

$$\langle E_c \rangle = 0$$

and

$$s_c^2 = \langle E_c^2 \rangle = s_{\text{samp}}^2 + s_{\text{ion}}^2 \qquad (9)$$

where s_c^2 is the variance in centroid calculations consisting of two independent components. The first component, s_{samp}^2, is that due to discrete sampling and is given by

$$s_{\text{samp}}^2 = \frac{(2\pi)^2}{A^2} \int dw |\tilde{\varphi}(w)\tilde{f}(w)|^2 \qquad (10)$$

where $\tilde{\varphi}(w)$ is the Fourier transform of the sawtooth function,

$$\varphi(t) = t - m\,\delta t \qquad (m-\tfrac{1}{2})\,\delta t < t < (n+\tfrac{1}{2})\,\delta t$$
$$m = \cdots, -1, 0, 1, \cdots$$

and $\tilde{f}(w)$ is the Fourier transform of the expected peak envelope. The second component, s_{ion}^2, is that due to ion arrival statistics and is given by

$$s_{\text{ion}}^2 = \frac{s_{\text{peak}}^2}{A} \qquad (11)$$

where s_{peak}^2 is the calculated width of the ion peak.

These expressions can be evaluated analytically for particular peak shapes and are in good agreement with Monte Carlo simulations of centroid error statistics. The dependence of centroid error on sample rate for Gaussian peaks sampled with an "ideal integrator" and containing various numbers of ions is shown in Fig. 3-31. The key result

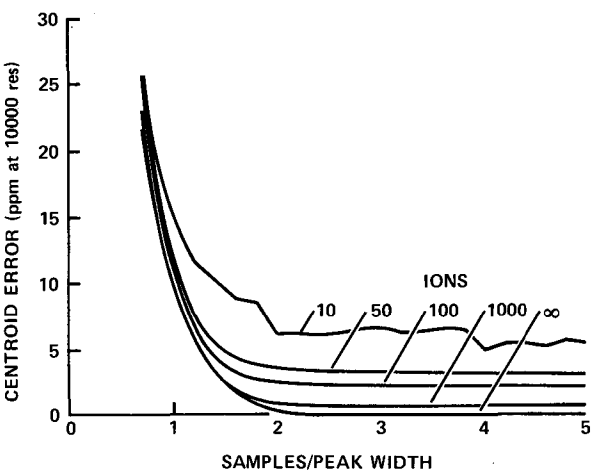

Figure 3-31. Results of a Monte Carlo simulation of the relationship between computed centroid accuracy and sample rate for Gaussian peaks containing various numbers of ions. The peaks are sampled with an "ideal integrator" and sample values are subject to Poisson-distributed fluctuations.

evident in these plots is that there is little improvement in *centroid* accuracy to be gained by sampling the ion current output more frequently than about two or three samples per peak! These plots are not to be interpreted to mean that the errors included here are the only ones affecting mass accuracy. There are other significant error sources (e.g., electronic noise and exponential scan model defects) that further compromise mass accuracy, so one may not achieve these theoretical accuracy limits in practice. In fact, the other error sources make it difficult to evaluate the effects of sampling and ion statistics alone using actual spectrometer data. The main effect to be observed in a proper statistical analysis of actual data is that there is no improvement in mass measurement accuracy by increasing the sampling rate above the limits given above.

In some circumstances, there is another data analysis requirement, however, that warrants sampling faster than two or three samples per peak, even with the resulting loss in sensitivity. When unresolved mass multiplets are encountered in samples, either between the sample ions themselves or between sample and reference compound ions, a higher instrument resolution must be used or the peak envelope must be more accurately encoded to allow more sophisticated computer analysis. The sampling rate limit derived above, based solely on centroid errors, clearly cannot give a very good approximation to the detailed peak envelope. The sampling rate required for this latter case is given by the Nyquist sampling criterion (20) based on the complexity of the envelope shape. The Nyquist criterion states that the sampling rate must be at least twice the highest-frequency component in the Fourier transform of the envelope. Given samples collected with at least this rate, an interpolation procedure exists to compute any intermediate point on the peak envelope. For a Gaussian peak shape, there is clearly no highest-frequency component. But, in the presence of noise, it does not make sense to encode frequencies above where the noise transform begins to dominate the peak envelope transform, that is, where the envelope errors from noise dominate interpolation errors from undersampling the data. In fact, from the statistical properties described earlier in the discussion of centroid error, it is clear the noise amplitude increases with increasing sampling rate. Since we must encode the envelopes of peaks with differing numbers of ions, we have adopted a compromise for the highest peaks, consistent with these noise limits. We sample at a frequency of twice that where the peak transform falls to 5% of its maximum. This rate corresponds to a sample spacing of about five samples across the peak envelope. Examples of the trade-off between sample rate and envelope reconstructability are given in Fig. 3-32. Figure 3-32a shows a Gaussian peak

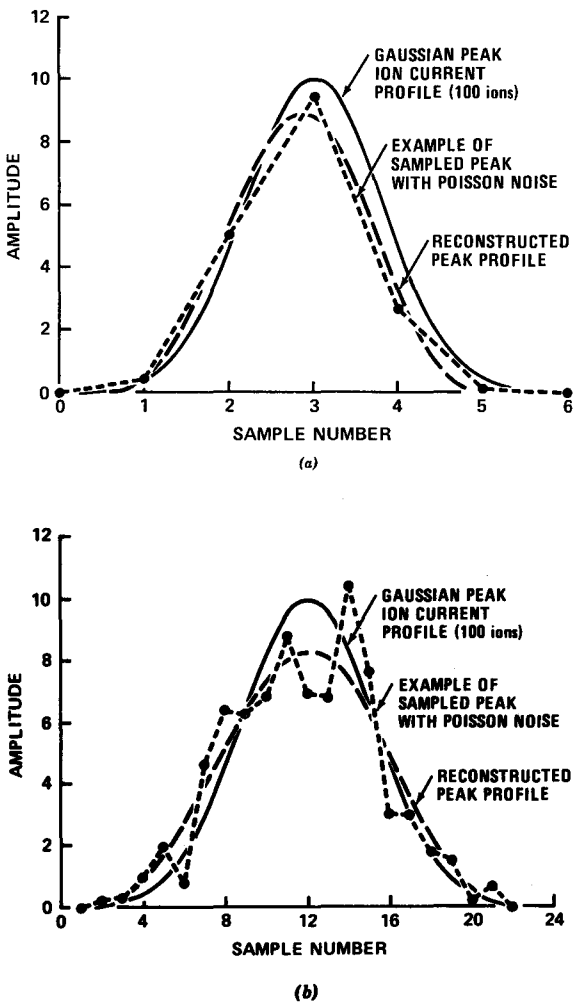

Figure 3-32. Examples of the effects of sampling rate on the reconstructability of peak envelopes. Each graph shows the expected ion current profile (a Gaussian with total area of 100 ions), an example of a set of sampled data points (measured with an "ideal integrator") with Poisson ion fluctuations, and an optimally reconstructed (interpolated) envelope. (*a*) Ion current profile sampled at the limit given by (7). (*b*) Ion current profile sampled at four times the limit given by (7). Note that the higher sampling rate (*b*) does not offer an advantage, since each sample point integrates fewer ions and hence is subject to greater amplitude uncertainties. These uncertainties offset the benefits of the higher sampling rate.

containing 100 ions, sampled according to the criterion above, and then reconstructed with an optimum filter (20) to estimate the original nominal peak. Figure 3-32b shows the same Gaussian peak, sampled at a rate 4 times the foregoing limit and also optimally reconstructed. Note that the noise in the envelope samples of this latter case is considerably higher than in the former and that no improvement results in the estimation of the original peak. The fact that the centroid of the peak in Fig. 3-32a is shifted to the left by about 3 ppm is a statistical accident of the particular peak example shown and is not related to the lower sampling rate (3 ppm is consistent with the expected centroid error for 100 ion peaks as shown in Fig. 3-31).

iii. PEAK DETECTION AND MULTIPLET RESOLUTION

Another important feature of an automated high-resolution data system is the ability to perform correctly under varying instrument operating conditions. On a day-to-day basis, many parameters of instrument performance are subject to change affecting, for example, signal gain, system noise level, resolution as a function of mass, mass peak shapes, and background level. Before sample measurements are made, a mass spectrometer is typically tuned up and calibrated by running reference compound spectra. However, during the extended time required for GC/HRMS operation, some of these parameters may change. Our data system has been designed to construct dynamic models of instrument parameters for use in setting thresholds for mass peak detection and measuring singlet peak characteristics to detect and resolve multiplets. These models are derived directly from the data as part of the real-time data reduction processing.

(1) Data-Adaptive Peak Detection Threshold

In processing instrument scans we take advantage of the fact that most of the scan time is spent between mass peaks. This gives an opportunity to set the best threshold for detecting mass peaks so as not to miss minor peaks or to overburden the system with noise spikes. A short time after the scan is started, while filling up the first data buffers to be recorded, we compute a signal amplitude histogram based on 400 sample points. This histogram measures the frequency of occurrence of amplitude values versus amplitude. Since the amplitude of background noise is random, this histogram will have a distribution predominantly determined by the statistics of stray ion events, dark current, and other electronic noise. The mode of the distribution measures the background offset and the standard deviation the noise amplitude. Occasionally, a peak may be included in the histogram collection. This will show up as a broadening of the noise histogram on the side above the mode, since the peak rides nominally on the average background level. Thus we can estimate the noise level, even in the presence of mass peaks, by measuring the histogram width below the mode. A data-adaptive detection threshold can then be set equal to the histogram mode plus an appropriate factor times the noise amplitude. This approach adapts well to background shifts and eliminates concerns about background changes with ion multiplier or other electronic adjustments.

(2) Data-Adaptive Multiplet Resolution

Even at common HRMS operating resolutions of 10,000, mass multiplets are encountered. In batch mode, given enough sample, one can resort to "ultra-high-resolution" operation to resolve such multiplets. In GC/HRMS this is not possible. Because of critical sensitivity constraints in GC/HRMS mode, it is important to minimize instrument resolution requirements so as to maximize sensitivity. For this reason, we reduce the resolving power of the mass spectrometer to 1 part in 5000 to 8000 to achieve better sensitivity, even though this increases the probability of unresolved mass peaks. In this mode, the data system must be able to cope with such unresolved mass peaks in order to reliably reduce the scans.

During instrument calibration with the reference material, we analyze the size, shape, and spacing of mass peaks to report to the operator information about instrument performance. This includes scan speed and duration, instrument sensitivity, resolution as a function of mass, and the quality of the fit between the actual instrument scan and the exponential scan model used in the data system to compute accurate sample masses. These data are used iteratively to tune the instrument. During a GC/HRMS run, however, peak shapes may change because of instrument tuning drifts or because of varying degrees of ion source saturation with effluent concentration. Since in order to resolve peak multiplets one needs to know the characteristics of a singlet peak, we must extract a model for a singlet peak from the real-time data. This model must also track the

variations of singlet peak shape with mass that are present in most instruments.

General approaches to multiplet resolution based on optimum detection filters (20) are too costly for routine application in real time. Rather, we have developed a scheme that is very economical for the most common case of doublets and that can be run in real time, deferring more complex but fortunately rare multiplet cases for later analysis. In our approach, we characterize each peak by its area, centroid, and its second and third moments.* A singlet model is initialized by analyzing the first 10 significant peaks of a scan, assuming that these will be predominantly singlets. We throw out deviant peaks in this set by a "majority rules" logic. Successive peaks are classified as singlets or multiplets by a comparison of their second moments with the observed second moments in the singlet model using the following discriminant function:

$$\text{If } s^2 > s_0^2 + f\,\mathrm{SD}(s_0^2), \text{ then}$$

$$\text{else} \quad \begin{matrix}\text{the peak is a multiplet} \\ \text{the peak is a singlet}\end{matrix} \qquad (12)$$

where s^2 and s_0^2 are the second moments of the sample peak and singlet model, respectively, SD is the standard deviation of the second moments of the peaks comprising the singlet model, and f is a factor relating to the confidence that singlets will not be considered as doublets (we typically use $f = 2$). If the peak is not a singlet, it is analyzed as a doublet with the areas and locations of the two components given by

$$A_{\text{small}} = \frac{A}{2}\left(1 - \sqrt{1 - \frac{4}{4+G}}\right)$$

$$A_{\text{large}} = \frac{A}{2}\left(1 + \sqrt{1 - \frac{4}{4+G}}\right) \qquad (13)$$

$$X_{\text{left}} = X_0 - \sqrt{\frac{A_{\text{right}}}{A_{\text{left}}}(s^2 - s_0^2)}$$

$$X_{\text{right}} = X_0 + \sqrt{\frac{A_{\text{left}}}{A_{\text{right}}}(s^2 - s_0^2)}$$

where "small" and "large" refer to the larger and smaller of the multiplet components, respectively, and "left" and "right" refer to the time ordering of the peaks in the scan. Also, X_0 is the centroid of the multiplet profile and

$$G = \frac{(K^3 - K_0^3)^2}{(s^2 - s_0^2)^3}$$

where K^3 and K_0^3 = third moments of the multiplet profile and singlet model, respectively

and the left-hand peak is the smaller (A_{small}) if $K^3 < K_0^3$.

If the superposition of singlet models at the resolved locations and with the computed sizes in fact closely approximates the multiplet envelope, the peak complex is labeled a doublet. If not, the peak complex is a higher-order multiplet that has to be resolved by more complex means.

The singlet model must be constantly updated, since peak envelope shapes and instrument resolution change as a function of mass. As a new peak is classified as a singlet, if it is of significant size, it is added to the model and the oldest peak in the model removed. In this way the model adapts dynamically to the data along the mass scale as peaks are processed in real time during a scan.

These decision criteria and doublet resolution calculations are subject to errors that depend on the number of ions in the peak, the relative sizes of the doublet components, and their separation. For equal peaks containing 500 ions each and separated by 1 peak standard deviation (20 ppm at a resolving power of 10,000), the algorithm recognizes 98% of the doublet peaks encountered. For such doublets, resolved locations are subject to an error (std dev) of 5 ppm and the areas to an error of 40%. At a separation of 2 peak standard deviations, all 500 ion doublets are recognized, resolved centroids are accurate to 2.2 ppm, and areas to 11%. This method cannot resolve doublets that are not detectably wider than the singlet peak model. Such doublets can, however, be detected and resolved based on accurate mass measurements and inferred elemental compositions of ions (37).

iv. Reference Spectrum Detection and Scan Modeling

In GC/HRMS analysis of complex mixtures, one cannot guarantee that the reference compound will always have the most negative mass defects or that its peak amplitudes will be readily distinguishable from those of the sample. Thus a highly reliable means is needed to distinguish reference from sample peaks for use in accurate mass

* The nth moment of a peak is given by $\int (x - x_0)^n p(x)\,dx$, where $p(x)$ is the peak envelope (normalized to unit area) and x_0 is the peak centroid.

assignments. This scheme must track instrument scan drifts and relative changes in reference to sample amplitudes so that all spectra can be reduced without time-consuming manual intervention. Our approach has two phases; initial detection of a group of reference peaks followed by the incremental detection of neighboring reference peaks.

(1) *Initial Reference Detection*

To find an initial set of reference peaks in a sample spectrum, we match a pattern of reference peaks from the instrument calibration run (reference only) against a collection of "reasonable" candidate patterns in the sample spectrum. If $\{M_i, T_i\}$ is the set of mass–time pairs for pattern members in the reference run, there may be several reasonable candidates $(M_i, \{t_j\}_i)$ for each pattern member in the sample run. We use a pattern of 10 calibration peaks and consider up to three candidates for each pattern member in the sample spectrum. Such candidates are chosen on the basis of their proximity within windows to the locations of pattern members estimated from the reference run, taking into account the time shift observed for the candidate reference compound base peak, and a comparison of peak areas expected based on the calibration run.

The plausible reference patterns in the sample spectrum (one of the candidates chosen corresponding to each reference pattern member) are compared against the reference run pattern after adjusting the observed times of the sample pattern members to fit those of the reference run most closely. This adjustment uses a simple exponential scan model

$$M(t) = \exp\left[\frac{-(t-t_0)}{\tau}\right]$$

and accounts for possible changes in instrument scan parameters between the time of calibration and the sample run. To first order, if there were a change in scan parameters, the times of the calibration peaks would be shifted by

$$dt_i = dt_0 - \ln(M_i)\, d\tau \qquad (14)$$

Scan parameter adjustments are selected for each of the sample patterns so as to minimize the mean-squared difference in times between corresponding members in the sample and reference run patterns. We also admit the possibility that for some pattern members no reference peak can be found in the sample spectrum (e.g., if the reference compound is sufficiently suppressed relative to the sample). When a member is eliminated, a "mismatch" penalty is included in computing the mean-squared difference in times that is equal to the mean error in a fit of the exponential scan curve to the reference pattern. This prevents the best pattern from being that with all peaks thrown out.

The set of peaks in the sample spectrum corresponding to the reference pattern is then selected to be the candidate pattern that corresponds most closely to the pattern in the calibration run, that is, has the smallest mean squared adjusted time location difference between its members and those of the calibration pattern. This type of reference detection process works quite reliably in the presence of complex sample spectra and adapts to changes in the instrument scan function.

(2) *Incremental Reference Detection*

Once an initial set of reference peaks is identified in the sample spectrum, sample masses spanned by these peaks are interpolated using a scan model given by

$$M(t) = A + \exp\left[\frac{-(t-t_0)}{a+bt+ct^2}\right] \qquad (15)$$

which we have found most closely approximates the scan peculiarities of our Varian MAT 711 and AEI MS-9 instruments while, in addition, being easily computable. The mass offset (A in *15*) at infinite time is important since the asymptotic field of the magnet does not necessarily correspond to zero mass at the detector.

We also use this function to extrapolate to find the best candidate for the next reference mass. The system projects to find the expected time of the next reference peak and searches for candidates in a window around that location. If none is found, it is assumed the reference peak is missing and a projection is made for the next one. If one is found, it is assumed to be the reference peak and the sample peaks spanned by it are interpolated. If more than one candidate is found in the window, we pick the two closest to the expected location. Each of these is used to project for the following reference peak. If only one succeeds to find a unique peak, it is chosen as the appropriate alternative. If both succeed, the pair with locations best corresponding to those in the calibration run (after applying shifts for scan differences as described above) is chosen.

This scheme has proven quite reliable in finding reference peaks among sample spectra conflicting

with the reference spectrum. It also has the necessary data adaptivity to track instrument drifts over GC/HRMS experiment time.

v. ELEMENTAL COMPOSITION MATCHING

One of the principal results of HRMS is the tabulation of elemental compositions consistent with the mass assignments for spectral peaks. In GC/HRMS such composition matching is done many times and a highly efficient computer algorithm is essential. Our system uses a fast table-lookup matching algorithm described by Lederberg (38). By tabulating mass defects for a range of hydrogen, nitrogen, and oxygen combinations that have nominal masses equal to multiples of ^{12}C, a simple lookup in the table, using the observed defect as an index, will give a candidate composition. Other heteroatoms and ^{12}C-equivalent mass components are evaluated by sequential subtraction and further table lookup. The ability to translate from a mass defect to a composition by means of a lookup in a compact table is what makes this algorithm very fast.

vi. EXAMPLES AND LIMITATIONS

We have used GC/HRMS as an adjunct to the more routine GC/LRMS analyses of biological mixtures. When eluted components are encountered that cannot be identified by simple library search, we have as an option using GC/HRMS to give more precise clues to the structure, assuming that enough of the component is present. Examples of the use of this kind of information as a part of procedures for the determination of unknown structures are given in Chapter 7.

Using the procedures outlined above, the vast majority of scans taken during a GC/HRMS run can be reduced automatically, with *no* intervention by the operator. These techniques fail when the spectrum of the reference compound is overly suppressed during the elution of an abundant GC effluent. Such scans are not reducible to accurate masses and compositions.

GC/HRMS is limited in its application by instrument sensitivity. In order to get reasonable elemental composition specificity on our instrument, we need to take mass spectra at resolving powers below 10,000 and usually at 5000. In order to use the CLEANUP procedure (see Section c) at our GC elution rates, sampling considerations demand that spectra be scanned as rapidly as is possible on our instrument (i.e., at about 2 s/decade). In practice, we often scan at lower rates to balance sensitivity, scan duration, and GC peak width. We use 8 to 10 sec/decade for packed GC column studies of natural marine products or minor constituents in urinary extracts. Our instrument is currently not equipped for capillary column GC/HRMS analyses, which can do a better job of separating closely eluting materials. Acquiring interpretable spectra with the much narrower peaks from capillary GC places even more emphasis on faster mass spectrometer scan rates and is a difficult problem that we have not been able to attack.

These constraints place rather stringent limits on the achievable sensitivity. Absolute sensitivity figures, of course, depend on the overall efficiency of the GC/separator/MS system and on the fragmentation pattern of the spectrum being analyzed. For free sterol mixtures, a common but very difficult application in our laboratory (7–10), we are able to measure spectra with a 100 : 1 dynamic range of ion abundances for sample amounts as low as 1 μg with our current instrument.

f. Conclusions and Summary

It seems clear that the computer will play an increasingly important role as a problem-solving tool for molecular structure elucidation in general and for GC/MS applications in particular. As more and more routine use is made of these analytic tools, the only tractable way of managing the volume of data produced and correlating the derived information with other experimental results is more fully automated data systems. Because the human being will be unable to systematically review intermediate results, it is imperative in such systems that reliability and accountability be built into the programs comprising them. Such responsibility requires an increasing level of introspection on the part of the programs as they perform their tasks. Assessments of the quality of the results the programs produce must be based on models of the instruments and data they are supporting. Such qualification of results can then be used to most effectively deploy human resources to problems of ambiguity, particular difficulty, or anomaly relating to past experience. The earlier sections describe several developmental steps in this direction that have been of significant benefit to our own applications and that we export to other laboratories within our available resources. We feel that as new commercial systems are engineered, they must pay increasing attention to the quality and depth of software support provided to analytic instrumentation.

ACKNOWLEDGMENTS

The development of the GC/MS systems described here has benefited from the work of many people and has been guided by the needs of an equally large number of people using the systems for biomedical applications. We gratefully acknowledge the major contributions and patience of the following individuals, who have contributed to various aspects of project guidance, refinement of the computer programs and instrument hardware, and the critique of results from their operational use in our laboratory: Profs. J. Lederberg and C. Djerassi; Drs. Alan Duffield, W. Pereira, and G. Dromey; and M. Stefik, N. Veizades, R. Tucker, and G. Jirak.

This work was supported by grants from the National Institutes of Health (RR-612 and GM-20832) and from the National Aeronautics and Space Administration (NGR-05-020-632). Computing support for interactive program development was provided by the SUMEX resource, funded by the Biotechnology Resources Program of the National Institutes of Health (grant RR-785).

REFERENCES

1. Waller, G. R., Ed., *Biochemical Applications of Mass Spectrometry*, Wiley-Interscience, New York, 1972.

2. Cram, S. P., and Risby, T. H., *Anal. Chem.* **50,** 213R (1978).

3. Burlingame, A. L., Shackleton, C. H. L., Howe, I., and Chizhov, O. S., *Anal. Chem.* **50,** 213R (1978).

4. Smith, D. H., and Carhart, R. E., in Gross, M. L., Ed., *High Performance Mass Spectrometry: Chemical Applications*, American Chemical Society, Washington, D.C., 1978, p. 325.

5. Fitch, W. L., Anderson, P. J., and Smith, D. H., *J. Chromatog.* **162,** 249 (1979).

6. Anderson, P. J., Fitch, W. L., and Halpern, B., *J. Chromatog.* **146,** 481 (1978).

7. Delseth, C., Carlson, R. M. K., Djerassi, C., Erdman, T. R., and Scheuer, P. J., *Helv. Chim. Acta* **61,** 1470 (1978).

8. Ayanoglu, E., Djerassi, C., Erdman, T. R., and Scheuer, P. J., *Steroids* **31,** 815 (1978).

9. Carlson, R. M. K., Popov, S., Massey, I., Delseth, C., Ayanoglu, E., Varkony, T. H., and Djerassi, C., *Bioorg. Chem.* **7,** 453 (1978).

10. Theobald, N., Shoolery, J. N., Djerassi, C., Erdman, T., and Scheuer, P. J., *J. Am. Chem. Soc.* **100,** 5574 (1978).

11. Fitch, W. L., Everhart, E. T., and Smith, D. H., *Anal. Chem.* **50,** 2122 (1978).

12. Fitch, W. L., and Smith, D. H., *Environ. Sci. Technol.* **13,** 341 (1979).

13. Pereira, W. E., Summons, R. E., Rindfleisch, T. C., Duffield, A. M., Zeitman, B., and Lawless, J. G., *Geochim. Cosmochim. Acta* **39,** 163 (1975).

14. Davis, R. M., *Science* **195,** 1096 (1977).

15. Dromey, R. G., Stefik, M. J., Rindfleisch, T. C., and Duffield, A. M., *Anal. Chem.* **48,** 1368 (1976).

16. Smith, D. H., Yeager, W. J., Anderson, P. J., Fitch, W. L., Rindfleisch, T. C., and Achenbach, M., *Anal. Chem.* **49,** 1623 (1977).

17. Smith, D. H., Yeager, W. J., and Rindfleisch, T. C., *Anal. Chem.* **50,** 1585 (1978).

18. Sweeley, C. C., Gates, S. C., and Holland, J. F., *Anal. Chem.* **50,** 1585 (1978).

19. Markey, S. P., Urban, W. G., and Levine, S. P., *Mass Spectra of Compounds of Biological Interest*, USAEC Office of Information Services documents TID 26553 P1, P2, and P3.

20. Hamming, R. W., *Digital Filters*, Prentice-Hall, Englewood Cliffs, N.J., 1968.

21. Grushka, E., Myers, M. N., and Giddings, J. C., *Anal. Chem.* **42,** 21 (1970).

22. Scott, C. D., Chilcote, D. C., and Pitt, W. W., *Clin. Chem.* **16,** 637 (1970).

23. Nau, H., and Biemann, K., *Anal. Lett.* **6,** 1071 (1973).

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

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

26. Hertz, H. S., Hites, R. A., and Biemann, K., *Anal. Chem.* **43,** 681 (1971).

27. See, for example, Kolmogorov, A. N., and Fomin, S. V., *Elements of the Theory of Functions and Functional Analysis, Vol.* 1: *Metric and Normed Spaces*, Graylock Press, Rochester, N.Y., 1957.

28. Jellum, E., Helland, P., Eldjarn, L., Markwardt, U., and Marhofer, J., *J. Chromatog.* **112,** 573 (1975).

29. Dijkstra, E. W., *Numer. Math.* **1,** 269 (1959).

30. Blaisdell, B. E., *Anal. Chem.* **49,** 180 (1977).

31. Stokke, O., *Biomed. Mass Spectrom.* **3,** 97 (1976).

32. Carhart, R. E., Smith, D. H., Brown, H., and Djerassi, C., *J. Am. Chem. Soc.* **97,** 5755 (1975).

33. Smith, D. H., and Carhart, R. E., in Gross, M. L., Ed., *High Performance Mass Spectrometry: Chemical Applications*, American Chemical Society, Washington, D.C., 1978, p. 325.

34. Beynon, J. H., *Mass Spectrometry and Its Application to Organic Chemistry*, Elsevier, Amsterdam, 1960.

35. Kimble, B. J., in Gross, M. L., Ed., *High Performance Mass Spectrometry: Chemical Applications,*

American Chemical Society, Washington, D.C., 1978, p. 120.

36. Halliday, J. S., *Adv. Mass Spectrom.* **4,** 239 (1968).

37. Burlingame, A. L., Smith, D. H., Mevren, T. O., and Olsen, R. W., in Orr, C. H., and Norris, J. A., Eds., *Computers in Analytical Chemistry,* Vol. 4 of *Progress in Analytical Chemistry,* Plenum, New York, 1970, p. 17.

38. Lederberg, J., *J. Chem. Educ.* **49,** 613 (1972).

9. The Upjohn Company

Lubomir Baczynskyj, Physical and Analytical Chemistry Research, The Upjohn Company, Kalamazoo, Michigan

a. Introduction

background subtract). This system functioned well within its original design limitations.

In 1974 we acquired an additional mass spectrometer, the Varian MAT CH-7, and a year later we replaced the Atlas MAT CH-4 with the Varian MAT CH-5 DF mass spectrometer. At the same time the disk storage capability of the IBM 1800 computer was expanded by the addition of two 2311 disk drives, which added 7 million words of disk space to the system. This allowed all raw and processed data to be stored on disk. The additional instrumentation, as well as the extra disk storage capabilities and the need for a more flexible and powerful processing of the data, prompted us to rebuild and rewrite the MS system.

An extension of the system became possible when the IBM 1800 computer was linked via a