United States
Environmental Protection
Agency

Environmental Research
Laboratory
Athens GA 30613

Research and Development

EPA-600/S4-83-029 Sept. 1983

EPA

Project Summary

A Computer Survey of GC/MS Data Acquired in EPA's Priority Pollutant Screening Analysis: System and Results

W. M. Shackelford, D.M. Cline, F. O. Burchfield, L. Faas, G. Kurth, and A. D. Sauter

The screening analysis phase of the best available treatment (BAT) review of wastewater treatment techniques by EPA was initiated to assess 21 industrial categories for the 129 "priority pollutants." Implicit in the purpose of the screening analysis for these pollutants was the notion that the raw gas chromatography/mass spectrometry data would be saved for later evaluation for compounds not on the priority pollutant list. To this end, a system of computer programs was built that automatically extracted the pure spectra of components in a GC/MS run; matched the spectra against a reference library, and dealt appropriately with matched and unmatched spectra. Matched components were entered into a database for statistical studies to determine their priority for further study. Unmatched spectra were compared to each other to find recurring unknowns so that priorities for *ab initio* identifications could be set. Component software was obtained from Stanford University (CLEANUP) and Cornell University (PBM); some software was written at the Athens Environmental Research Laboratory.

The automated survey techniques appeared to work well on most of the GC/MS data. The system was efficient and cost effective, for tentative identification of the major components in the samples.

This Project Summary was developed by EPA's Environmental Research Laboratory, Athens, GA, to announce key findings of the research project that is fully documented in a separate report of the same title (see Project Report ordering information at back).

Introduction

In June 1976, the U.S. Environmental Protection Agency (EPA), as a result of court action by several environmental groups, was directed by a Consent Decree from the U.S. District Court in the District of Columbia to assess the wastewater of 21 industrial categories for 65 chemical substances and to prescribe the best available treatment (BAT) for the effluent. To begin the task, a scheme for analysis of the wastewaters for the 65 substances had to be designed.

Although some of the 65 substances were unique chemical compounds, many included whole classes of compounds (*e.g.* polynuclear aromatic hydrocarbons). Realizing that such classes of compounds could contain literally hundreds of individual members, EPA included for analysis only those members that had been previously identified a significant number of times, were produced in quantity by industry, and were available as analytical standards. The now familiar 129-compound priority pollutant list was the result of this work.

Even though the list of 129 specific substances made the analysis task manageable, the plaintiffs in the court action were concerned that some members of the chemical classes not on the 129-compound list would be missed in the

analysis procedure. Because it was generally agreed that computerized gas chromatography/mass sepctrometry (GC/MS) would be the analysis tool of choice, the advantage of saving all raw GC/MS data for later processing to look for compounds other than the priority pollutants became obvious. The state-of-the-art GC/MS instrumentation includes a computer system; thus, the data would be saved in computer-readable format for later study. Magnetic tape, the cheapest mass storage medium, was chosen for recording all GC/MS data from sample analysis.

Initial analysis of each sample at the laboratories operating under EPA contract was to be directed only toward compounds among the 129 priority pollutants. Although EPA might have contracted for a general survey of all compounds in each sample, a number of limiting factors precluded this approach:

- Cost of a general survey was estimated at $2000 per sample versus $700 per sample for the limited analysis.
- Time was extremely important. Although the data acquisition times for general survey and specific analysis are the same, data-evaluation times could be 5 to 10 times longer for survey analysis (if only computer matching were required for identification). Decreasing the number of samples per unit time by a factor of 5 to 10 would have played havoc with the court-ordered deadlines.
- Management of the large volumes of unconfirmed data would have required a massive secondary effort to confirm and collate the results of the survey analysis.

By requiring that all data from each GC/MS acquisition be sent to a central location for survey processing, EPA assured proper management of non-priority pollutant data and at the same time obtained timely response on the priority pollutants directly from contractor laboratories at a reasonable cost. Because all parties involved in the Consent Decree had agreed that the non-priority pollutant data were of less immediate need, no part of the spirit of the Consent Decree was sacrificed; yet provision was made for assessing all the data for compounds other than the 129.

The analysis laboratories were required to supply each sample extract along with the GC/MS data as a second provision for possible later analysis of the sample. Thus, should some compound be tentatively identified in the GC/MS data, it could be confirmed by reanalysis of the corresponding extract. Also, recurring components, not identifiable from their mass spectra, possibly could be identified using another analysis technique on the saved extract.

The screening analysis phase of BAT review was expected to require the qualitative/semi-quantitative analysis of about 4000 samples. Each sample analysis involved GC/MS data acquisition for at least five fractions: a volatile organics analysis (VOA), a VOA blank, an extractable base/neutral (B/N), an extractable acid (ACI), and a direct aqueous injection (DAI). Other blank, standard, and pesticide confirmation runs also were needed. All calculations for the task were based on 20,000 GC/MS runs (4000 samples x 5 fractions). Considering that each GC/MS run was expected to contain some 500 to 1000 individual spectra, the magnitude of the task of evaluating these data is evident.

Implicit in this data evaluation task was the development of a computer system that might evaluate the data in a manner comparable to a human using computer-aided spectrum extraction and spectra matching to tentatively identify all sample components. An additional goal was the identification of those spectra which did not match any spectrum in the reference library yet were seen in multiple GC/MS runs. Thus, a library of compounds tentatively identified in each industrial category, as well as a library of recurring but unidentified spectra, were to be generated for use in effluent regulation. Also, the data in both libraries were to be studied in a subsequent project, which will reanalyze the saved extracts. Tentative identifications made in that project could be confirmed by comparison with standards, and recurring but unidentified spectra could be examined for *ab initio* determination of compound identity.

System Description

The PDP 11/70-based GC/MS Data Survey System consisted of computer hardware and software programmed to accomplish the following functions:

1. Inventory all incoming magnetic tapes and sample extracts.
2. Copy the data on each magnetic tape to a second tape in an internal use format and plot the reconstructed gas chromatogram.
3. Retrieve data as necessary from tapes in batch mode.
4. Extract the spectra of components in each GC/MS run from the background spectra in the run.
5. Match the extracted spectra with a library of reference spectra.
6. Check if matched spectra have been seen before under the same circumstances.
7. Check spectra that are not matched against their fellow unmatched spectra.
8. Generate reports on the numbers of matched spectra by industry, fraction type, analytical laboratory, GC/MS run conditions, etc.
9. Provide graphics capability necessary to view the data from any run.
10. Search any run for specific compounds.

Inventory System

To inventory and track the 20,000 GC/MS data runs and the estimated 12,000 extracts (a B/N, ACI, and pesticide fraction for each of 4000 samples), a database management system was implemented. This system was the INFORM management program, a well-known tool for database management. In INFORM were kept the GC/MS data run descriptors that allowed physical location of each run and corresponding extract and all available information about the sample.

As each magnetic tape or extract was received at the Athens Laboratory, it was manually entered into the INFORM system. Important parameters entered for each data run were the tape on which it was found, the EPA sample number, an Athens Laboratory run number, the fraction type, and various GC/MS parameters. The corresponding data for the extract included all of this information and the precise location of the extract in a freezer.

During the inventory process, data received from contractor laboratory tapes were copied onto an Athens Laboratory tape in a format that was both more space-efficient and damage resistant. Thus, the original tape and a backup copy were saved. The backup copy, which had only the Athens Laboratory number for identification of each run, was used for all data processing needs. Confidentiality of the data was maintained through the use of the backup copy so that descriptive data were not associated with the GC/MS data. Software that had access to both descriptive and GC/MS data was password protected.

At the time of tape conversion from the contractor's format to the Athens Laboratory format, the data of each run were scanned and a reconstructed gas chromatogram (RGC) plotted. The RGCs were then bound in volumes to serve as references at the time runs were submitted for analysis. Inspection of an RGC by a chemist might result in the discarding of the cor-

responding run because of obvious flaws such as absence of peaks or premature end of data.

When data were to be processed, a chemist identified the runs that passed visual inspection for processing. Software then retrieved the designated runs from the magnetic tape and prepared each run in turn for processing by the analytical system. The inventory system was reapplied when the run had been processed and descriptors contained in INFORM were necessary for reporting.

Analytical System

The analytical system consisted of four main parts: the internal standard locator, PEAK; the peak or spectrum extractor, CLEANUP; the spectrum matching system, PBM; and the result collator. Chemists had opportunities at various points during the process to make decisions that could end processing or affect further processing of any given component spectrum. Ideally, data analysis proceeded with minimal operator intervention. Only when the analytical system was presented with decisions that it was unqualified to make did the chemist intervene.

The program PEAK was developed to assure identification of internal standard location in each run. Because all subsequent processing of the data required knowledge of the internal standard in the run, it was imperative that software be available that would unambiguously define the location and area of the internal standard peak in each data run.

CLEANUP is a system of programs developed at Stanford University that finds and extracts the spectra of components in a GC/MS data run. Successive 16-scan windows are searched for ion peaks that have 2 ascending points, a maximum, and 2 descending points. When an ion peak is found, successive ions from mass 40 to 400 are checked to see whether any maximize within a distance of ± 1 scan number of the first found peak. When 8 or more such masses maximize simultaneously, a component peak is said to be detected. In this case, all the masses maximizing at this point are collected, their areas are normalized to the largest mass of the group, and they are passed along to the next phase of the analysis as a mass spectrum.

CLEANUP involves a number of checks to insure that such artifacts as column bleed, noise spikes, and background are not chosen as sample components. Criteria are input at the start of processing to insure that only ion peaks of a defined sharpness will be considered. This procedure will normally eliminate peaks caused by column bleed, which usually shows up in the form of broad peaks. Noise spikes, which are generally of only one- or two-scan duration, are guarded against by requiring a minimum of four scans in the ion peaks. Instrumental background noise caused by pump oil or other contaminants normally does not peak during a run; therefore, it does not interfere with the CLEANUP process.

The spectra extracted by CLEANUP were passed to PBM, a library matching program developed at Cornell University under an EPA grant. PBM, or probability based matching, employs a reverse search technique to compare a reference library of condensed spectra to a similarly condensed unknown spectrum.

Reporting

Reporting is accomplished in two ways. The first system is a series of hard copy outputs that describe the flow of data through the total system and the results generated from the data. The contents of the historical library can be printed out either in totality or as a listing of unique entries. The data can be sorted by parameters such as CAS number, RRT, GC column, analysis laboratory, industrial category, relative concentration, etc.

A second method for reporting was a graphics system that allowed the chemist to recall data and plot it in various ways. For instance, the raw data for a spectrum, the cleaned up spectrum, and the reference spectrum can all be plotted on the same screen simultaneously. The extracted ion current profile (EICP) for any ion can be plotted between any scan limits. Multiple EICP plots can be displayed on the screen. The graphics system is used by chemists to evaluate ambiguous results from the computer analysis.

Extraction Results

The extraction of information-containing spectra from the mass of spectra in a GC/MS run is the key to a successful automated system for GC/MS data analysis. Figure 1 shows an RGC of a group of 11 phenols. The scans 213 and 215 can be seen to be on opposing sides of an apparent single component peak. Manual subtraction of a baseline spectrum (e.g. 208 or 220) from spectrum 214 results in a spectrum that is not recognizable as any of the components injected. The use of CLEANUP to find spectra, however, reveals that the peak is actually the sum of two components. Figure 2 shows the resultant spectra of 213 and 215. Also depicted are spectra from the reference library that

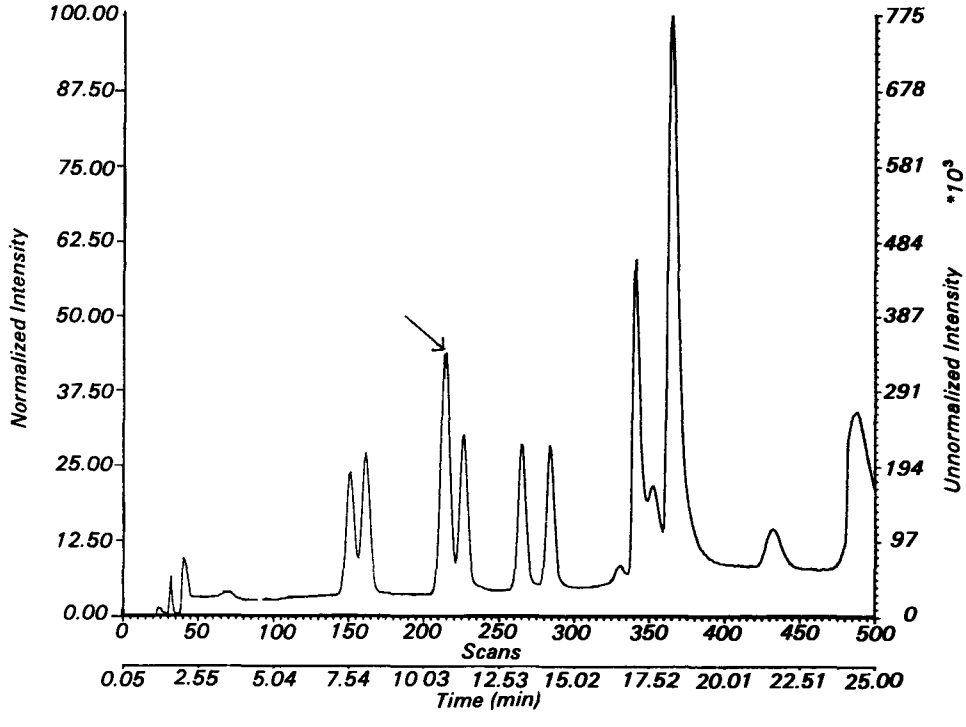

Figure 1. RGC of 11 component phenol standards. Arrow indicates apparent single component peak that is actually the sum of two components.

establish the identity of the two components. Although this example represents an ideal case in which standards were used with no interferences, it does serve to illustrate the ability of CLEANUP to separate components eluting within two scans of each other.

The data presented thus far indicate that for the systems studied, automated techniques are at least the equal of manual techniques for pointing out components in the run and identifying them by spectrum matching with a reference library. Spectrum extraction and identification are not always so clear cut. As shown in Table 1, despite the fact that more peaks are found with the automated method, the ratio of identifications-to-peaks has decreased. In fact, as the number of components in a run increases, identification becomes more and more difficult, even though the automated system apparently is able to deliver a spectrum for each component.

Spectrum-Matching Results

The spectrum-matching portion of the data analysis system has undergone the least modification. PBM has been evaluated in the literature and has been in use at the EPA Athens Laboratory for several years.

Selection of a database of reference spectra for use with PBM involved no small problem. Three databases were available: the Wiley collection, the National Bureau of Standards collection and the EPA master collection. The Wiley library contains 30,476 spectra of 30,476 compounds; the NBS library 25,025 spectra of 25,025 compounds; and the EPA master list ~40,000 spectra of ~32,000 compounds. The EPA list is the master list of spectra from which the NBS library was taken.

Because the GC/MS data used in this study came from a great variety of sources it was thought that "duplicate" spectra, *i.e.* multiple spectra of the same compound that differ slightly due to run conditions, in the database would be of some help in the matching process.

Table 2 compares the matching ability of two databases on the same spectra. As can be seen, Database II (the EPA master database) enjoys a distinct advantage over Database I (the NBS library) for the cases mentioned in the table. Comparison of the matches suggested by the two databases with the manual identification shows the superior ability of Database II. Data generated using the Wiley library showed similar shortcomings to the NBS library.

In cases where the identical matched spectrum occurred in all the libraries (as was generally the case), no problems were

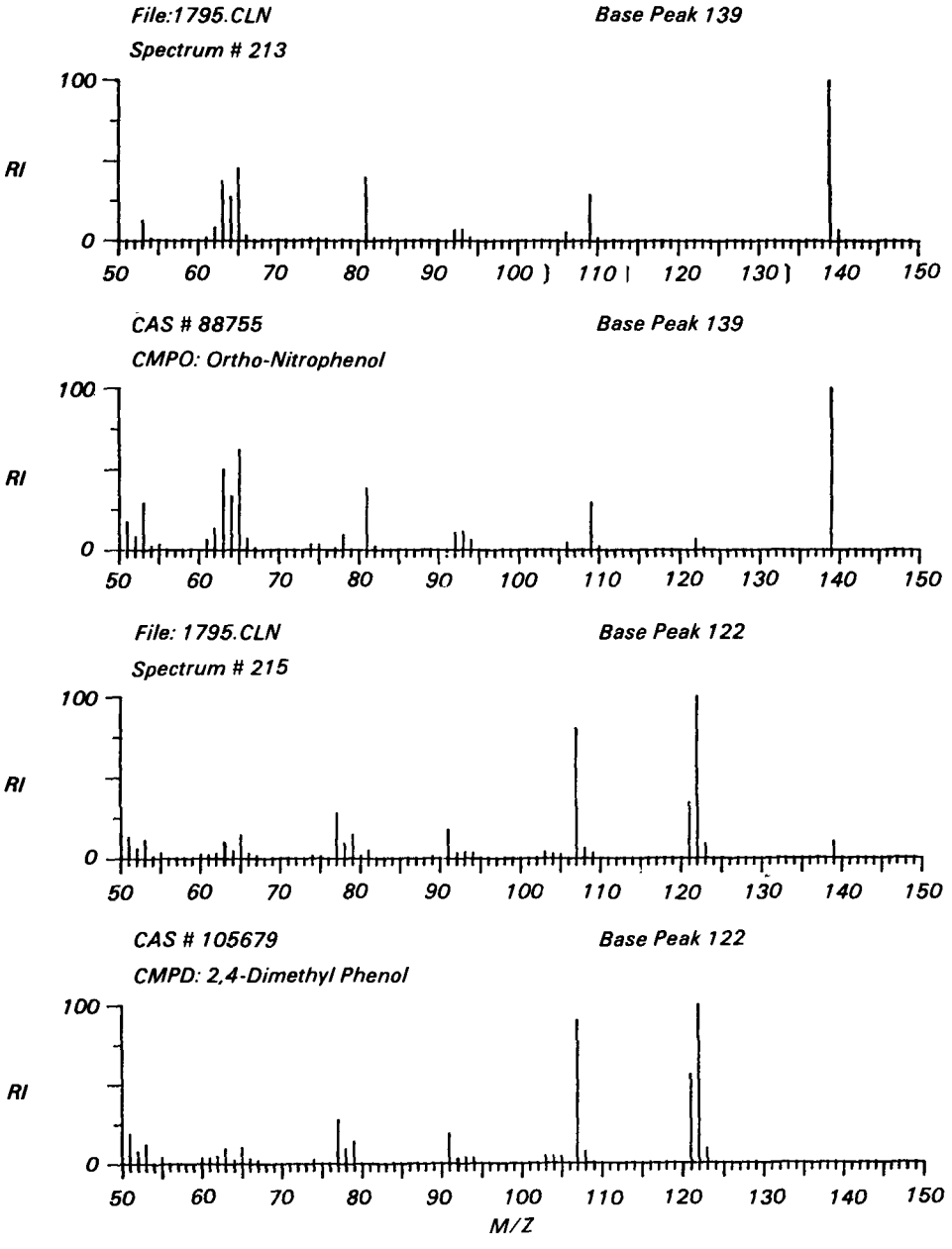

Figure 2. *Resultant spectra from scans 213 and 215 of figure 1 compared to matching library spectra.*

Table 1. *Comparison of Automated vs. Manual Peak Extraction (Values in Parentheses are Additional Components Identified by the Automated Method)*

Manual			CLEANUP-PBM		
Peaks	ID's	% ID	Peaks	ID's	% ID
30	18	0.60	46	18(6)	0.52
18	7	0.39	44	7(4)	0.25
31	11	0.35	41	11(2)	0.32
26	10	0.38	43	10(2)	0.28

Table 2. Comparison of PBM Matching Ability for the NBS Library (Database I) and the EPA Master Database (Database II)

Database I (K value, missing masses)	Database II (K value, missing masses)	Manual ID
Toluene (75+)	Toluene (75+)	Toluene
7-oxabicyclo 2,2,1 heptane (49+)	2-cyclohexene-1-ol (76+)	2-cyclohexene-1-ol
Phthalide (56, -2)	Methyl benzoate (69+)	Methyl benzoate
Hexacosanoic acid (102, -3)	Octadecanoic acid (105+)	Octadecanoic acid

(+ means that the molecular ion was matched within the proper intensity tolerance).

observed. In some cases, when spectrum extraction or run conditions slightly affected the spectrum, a match occurred only if there were duplicate spectra. Thus, the EPA master library was chosen for use in this work.

The reverse search approach of PBM is also useful for analyzing environmental samples with an automated system. Figure 3 is an example of a mixed spectrum obtained by CLEANUP from a VOA standard run. Although the two compounds (*cis*-1,3- dichloropropene and 1,1,2-trichloroethane) are not resolved, PBM was able to match both compounds in this spectrum. This ability of a reverse search to recognize components of mixed spectra is clearly an advantage.

Collator Results

The combination of MS data and retention indices provides a powerful tool in the automated extraction and identification of components in GC/MS data. Several other automated systems described in the literature rely on retention data as well as MS data for computer-aided identification of a known set of compounds. Although the match quality parameters vary from 45 to 100, the RRT variation is only 0.01 to 0.03. RRTs are particularly important in the identification of compounds such as alkanes or alcohols, which have little or no molecular ion and exhibit highly similar spectra.

Summary

Automated survey techniques for processing GC/MS data appear to work well on most of the data encountered in this study. The sensitivity of the CLEANUP-PBM package is not as great as that of a reverse search for specific ions, but it is adequate for the tentative identification of the major components in a sample. CLEANUP-PBM is cost effective when compared to the procedure in which an operator finds peaks, subtracts background, then matches the spectrum manually or by using computer search.

Use of a historical library for cataloguing data collected over long periods of time aids identification immensely by adding

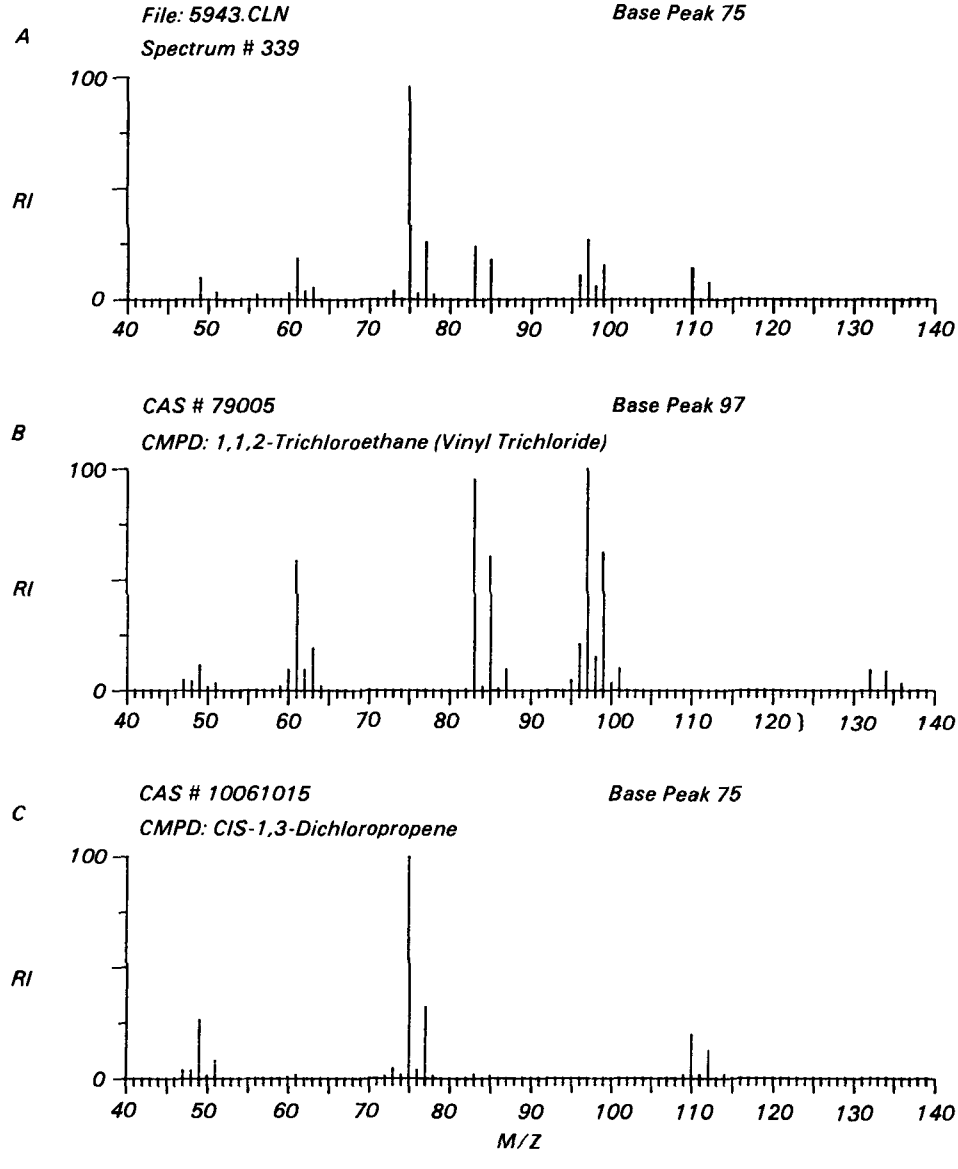

Figure 3. *A. Mixed spectrum extracted by CLEANUP from a VOA standard run. B and C. Library spectra are indicated as matched by PBM.*

Table 3. Comparison for RRT vs. K Variation for Selected Matched Compounds

Compound	Range of RRT	Range of K
Dioctylphthalate	*0.03*	*45 - 100*
Phthalide	*0.01*	*57 - 77*
Toluic acid	*0.03*	*48 - 85*

the dimension of GC retention data. Confidence in a tentative identification is heightened if corroborating GC retention data are available. This combination of spectral and retention data can effectively catalogue and highlight recurring unidentified substances for future study.

Study continues in two areas of the CLEANUP-PBM package. First, the proper compensation of background by CLEANUP is of concern because errors in intensity calculations reduce the chance that PBM will find a match for the spectrum. Studies underway indicate that background compensation similar to that used in PEAK would be effective in CLEANUP. Implementing such a routine is under study. Second, various parameters must be set for CLEANUP, and these are an area of concern. Parameters that pertain to peak shape and minimum area are inflexible during a given data run. Thus, changing chromatographic conditions that affect peak size and shape may cause CLEANUP to miss pertinent data. Automatic setting of these parameters during data processing is under study.

*The EPA authors, **W. M. Shackelford** and **D. M. Cline**, are with Environmental Research Laboratory, Athens, GA 30613. F. O. Burchfield, L. Faas, G. Kurth, and A. D. Sauter are with Computer Science Corporation, Falls Church, VA 22046.*
***W. M. Shackelford** is the EPA Project Officer (see below).*
The complete report, entitled "A Computer Survey of GC/MS Data Acquired in EPA's Priority Pollutant Screening Analysis: System and Results," (Order No. PB 83-220 111; Cost: $17.50, subject to change) will be available only from:
National Technical Information Service
5285 Port Royal Road
Springfield, VA 22161
Telephone: 703-487-4650
The EPA Project Officer can be contacted at:
Environmental Research Laboratory
U.S. Environmental Protection Agency
Athens, GA 30613