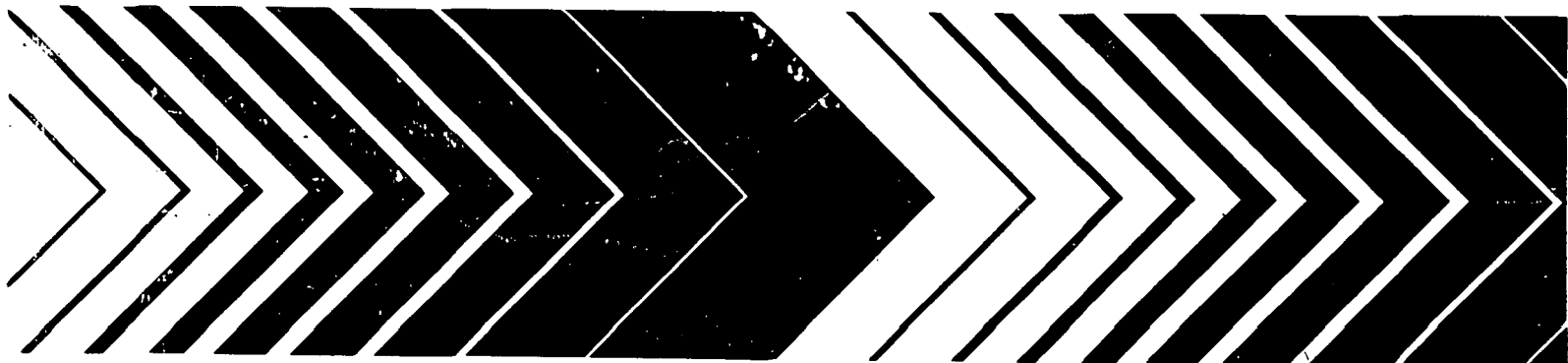&EPA **Technical Assessment of the Current Tentatively Identified Compound (TIC) Protocol**

Technical Assessment of the Current
Tentatively Identified Compound (TIC) Protocol

September 1997

J.R. Donnelly
Task Lead
Lockheed Martin Environmental Services

U.S. Environmental Protection Agency
National Exposure Research Laboratory-Las Vegas
Environmental Sciences Division
Environmental Chemistry Branch

G. Wayne Sovocool
Work Assignment Manager

Notice:

The U.S. Environmental Protection Agency (EPA), through its Office of Research and Development (ORD), partially funded and collaborated in the research described here. It is intended for internal EPA use only. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

Acknowledgments:

CONTENTS

TABLES

FIGURES

EXECUTIVE SUMMARY

The National Exposure Research Laboratory-Las Vegas (NERL-LV) conducted research on tentatively identified compounds (TICs) using Superfund samples and data submitted by the Contract Laboratory Program (CLP). This research effort is intended to provide valuable information for Superfund regarding TICs, which comprise approximately 90% of the analytes detected in Superfund samples. [The studies presented in this report on TICs used Superfund data submitted by the CLP.] These studies involved reviewing the CLP GC/MS hard-copy data and the raw data files to assess the effectiveness of current CLP protocols for TICs, and to provide information complementary to that reported by the CLP.

In this project, 99,513 TICs were reported in 792 Sample Delivery Groups (SDGs) studied. Of these, the CLP reported identifications with Chemical Abstracts Service (CAS) numbers for 16%. Not all of these identifications were correct, however. It was estimated from this study that perhaps 30% of these 16% were correct, and possibly another 10% were correct except that the TIC was an isomer of the compound whose CAS number was reported. Forty-one percent of the TICs were listed as being partially identified, and the remaining 43 percent were reported as "unknown". Examples of partial identifications include "unknown chlorinated aromatic," "unsaturated hydrocarbon," "unknown PAH" (polynuclear aromatic hydrocarbon), etc. In many cases, TICs were reported by the CLP as "unknown" despite having a high-probability mass spectral library match found by the data system. Data reviews conducted in this study emphasized analytical results for TICs in soils because this matrix type was found to be much more likely than water to contain compounds of interest.

An overview of the data indicated that the most commonly reported classes of TICs were saturated and unsaturated hydrocarbons. The next most prominent groups of compounds were PAHs and aromatic compounds, which were frequently substituted with aliphatic hydrocarbons. Higher molecular weight steroid compounds (i.e., cholesterol) as well as PCBs were also reported. Elemental sulfur was reported by the CLP laboratories in ca. 50% of soil samples. The Target Compound List (TCL) phthalates were found in ca. 80% of the soil samples, and non-TCL phthalates were found in ca. 10%. TIC mass spectra were found, including some with recognizable chlorine and/or bromine ion groups, for which there are no library spectra in the computerized mass spectral data bases. Such spectra must presently be manually interpreted. Additional compounds could be included in the NIST and Wiley mass spectral data bases to assist in identifying TICs for environmental monitoring efforts. These compounds include higher molecular weight PAHs, and industrial process solvents and chemicals. Additional pesticide metabolites and degradation products should be included; some are available in hardcopy form but not in commonly used software data bases.

The reporting trends from the laboratories were highly varied with respect to accuracy of reporting TICs. While some laboratories made honest efforts to identify the TICs, others simply labeled all TIC peaks as unknown and made no attempt to accurately identify the TICs. Several laboratories would only commit to identifying sulfur and reported all other TICs as unknown. Some laboratories used the name of the library match compound with the highest score on the library search, treating that as the identification regardless of whether that identification was reasonable based upon manual spectral interpretation.

Data system algorithms and procedures studied in this project included background subtraction, Biller-Biemann spectral isolation and library searching, and McLafferty Probability Based Matching (PBM). These software-based capabilities are implemented in the mass spectrometer vendor's data systems. Procedures for mass spectral resolution enhancement were also developed and tested, using the concept

proposed by Colby and in-house developed spreadsheet-based macros and other procedures. A comparison of the two library searching algorithms was performed on twenty CLP data files. Library searches were conducted with an HP DOS data system using PBM and with a Finnigan system using the Biller-Biemann algorithm. The results demonstrated that the two algorithms performed to the same level of quality and reliability for compounds in the molecular weight ranges associated with semivolatile TICs.

Commercial MassTransit™ software was tested for its ability to facilitate converting data files into formats usable by other data systems, including spreadsheets on personal computers. Interfacing the different MS data system formats through this software effectively accomplished a standardization of TIC data. This software converted GC/MS data from various contractors into ASCII text.

These data were subsequently imported into an EXCEL™ spreadsheet and manipulated using a macro written in-house to perform mass spectral resolution enhancement. The spreadsheet macro performed the sorting and statistical and mathematical procedures necessary to separate the TICs from interfering compounds. The extracted mass spectra were compared with reference spectra contained in the NIST database. Results of these comparisons showed that usually 80 to 90 percent of the ions contained in the reference spectra were successfully extracted using this method. This procedure improved mass spectral quality, and the data system's ability to perform successful library searches. The fit quality parameters showed systematic improvements after subjecting the data to resolution enhancement procedures.

In summary, this project investigated the effectiveness of current TIC reporting under the CLP protocols. Identification of TICs was found to be of variable quality across participating laboratories. The commonly used mass spectral data systems were found to provide essentially the same results. The libraries (Wiley and NIST) contain numerous entries that are not necessary for environmental studies and increase search time somewhat, and could be improved by adding additional compounds relevant to environmental monitoring. The available algorithms and data system procedures are satisfactory and provide virtually identical results. Mass spectral resolution enhancement procedures could materially help in identifying TICs by separating TIC spectra of interest from those of aliphatic hydrocarbons or other background signals. Several recommendations were made to improve the effectiveness of TIC reporting, although the specific mechanisms for implementation would require further study and testing:

(1)   require reporting the first library match of TICs on Form 1 if the library match meets a specified probability level, rather than "unknown" regardless of search results;

(2)   require reporting of TICs when the mass spectra, or library matches, indicate the presence of heteroatoms such as N, P, S, halogen, or heavy metals;

(3)   incorporate RRT data and frequency of occurrence of TICs into a data base, and add RRT criteria to aid in TIC identification;

(4)   extend the mass range to 600 Daltons (mass units, Da), with a tune emphasizing greater sensitivity for masses above 300 Da, for better detection and identification of TICs in this higher molecular weight range;

(5)   recommend or require that different GC temperature programs or column phases be used in a second analysis to improve separation of some TICs; and

(6)   because sulfur was found in approximately 50% of the soil sample data, improvements to the GPC procedure, a cleanup with copper, or some other improved procedure would be worthwhile;

(7)   identify additional compounds suitable for addition to the mass spectral data bases, because they have been found, or would be anticipated in known types of waste sites.

We understand that the first recommendation is currently being implemented by OERR. The remaining recommendations may not be suitable for a contract mechanism. instead, they may be better implemented by having an experienced laboratory specialize in the identification of TICs, with the freedom to use instruments and altered conditions for the solution of specific problems.

1. PURPOSE:

The National Exposure Research Laboratory-Las Vegas (NERL-LV) conducted research on tentatively identified compounds (TICs) using Superfund samples and data submitted by the Contract Laboratory Program (CLP). This research effort was intended to provide valuable information for Superfund regarding TICs. The initial effort included developing strategies and procedures necessary to assess and enhance the TIC information obtained through the CLP on samples submitted by the Regions. State-of-the-art approaches already developed through Office of Research and Development (ORD) research were applied to maximize the potential the identification of compounds in samples from Superfund sites. The results of these research-level studies on actual samples are designed to assist the Regions and the Project Officers in the following ways:

1)   measure the efficacy of current TIC identification and reporting procedures for compounds present in Superfund samples;

2)   identify improvements that may be needed for specific sample types, analyte classes, or programmatic reporting requirements;

3)   identify compounds having potential human or environmental risk;

4)   detect compounds, singly or in a series or group, that may be useful for determining the source of the pollution, or for tracking separate effluent streams or point sources.

2. TECHNICAL CONSIDERATIONS

The studies presented in this report on TICs used Superfund data submitted by the CLP. All regular analytical services (RAS) data available at NERL-LV i.e., those data known to be generated under the CLP protocols were studied. These studies involved reviewing the CLP GC/MS hard-copy data and the raw data files to assess the effectiveness of current CLP protocols[1] for TICs, and to provide information complementary to that reported by the CLP. The results included identifying TICs that may have value for environmental monitoring and remediation, confirming the presence of suspected compounds, and where necessary, correcting misidentifications. The study targeted both hazardous and potential marker (tracer) compounds for determining the source of contamination. Some compounds may not be considered toxic or suitable source marker compounds, but the techniques could be used equally well on other Superfund samples for components that could have these characteristics.

In this project, 99,513 TICs were reported in 792 Sample Delivery Groups (SDGs) studied (Table 2.1). TICs were found to comprise approximately 90-95% of the analytes detected in the Superfund sample data studied. Data reviews conducted in this study emphasized analytical results on soil, because this matrix type was found to be much more likely than water to contain compounds of interest. Of these, the CLP reported identifications with Chemical Abstracts Service (CAS) numbers for 16%. Not all of these identifications were correct, however. It was estimated from this study that perhaps 30% of these 16% were correct, and possibly another 10% were correct except that the TIC was an isomer of the compound whose CAS number was reported. Forty-one percent of the TICs were listed as being partially identified, and the remaining 43 percent were reported as "unknown". As a result, 84% to 90% of the analytes in Superfund samples remain unidentified under the current CLP reporting procedures. Examples of partial identifications include "unknown chlorinated aromatic," "unsaturated hydrocarbon," "unknown PAH" (polynuclear aromatic

hydrocarbon), etc. In many cases, TICs were reported by the CLP as "unknown" despite having a high-probability mass spectral library match found by the data system. Long and McGuire studied the data sets on 27 samples, agreeing with the CLP TIC identifications 36% of the time. They recommended discouraging the use of "unknown" for reporting TIC identities.[2]

It should be noted that traditionally an absolute chemical structure determination is based upon synthesis by a rational method giving unique products or by X-ray crystallography. It is generally impractical to separate individual analytes from environmental samples to perform such structure determinations, because of separation difficulties, low concentrations of the analytes, and cost or sample throughput considerations. Secondary but powerful identification methods may involve selected, problem-specific combinations of techniques such as nuclear magnetic resonance, optical spectroscopy, and mass spectrometry.

While less rigorous, the CLP identifications of Target Compound List (TCL) analytes are usually reliable because they are derived from (a) mass spectra matched against spectra of chemical standards obtained on the same instrument, and (b) capillary column gas chromatographic (GC) relative retention times (RRTs) matched to those of the chemical standards obtained on the same instrument. In contrast, the "identifications" of TICs are of lower reliability because authentic standards were not used for reference spectra and retention times. For TICs, the library data base spectra may have been obtained under different experimental conditions (instrumentation, sample introduction), and a GC RRT data base to assist in TIC identifications is not specified in the CLP protocol. However, the EPA set up a TIC Work Group of involved and interested parties, and the Group has developed a GC RRT data base for many TICs. In many cases, partial identifications may be sufficient to decide if further investigation is needed. For example, mass spectral data consistent with the assignment of chemical features (halogens, aromatic rings, alkyl groups, etc.) may be sufficient for some purposes. Reinterpretations and identifications made in this project utilized authentic standards whenever possible. In other cases, the identifications could not be considered fully confirmed but are still not as "tentative" as those reported by the CLP.

Currently, CLP laboratories are required to provide TIC identifications with the data package submission. For the semivolatile fraction, laboratories are required to report (per sample) a maximum of 20 TIC identifications (30 TICs for CLP OLMO3.1), whose individual areas are over 10% of that of the nearest internal standard.[1] They submit the top three data system "hits" or tentative identifications in the data package, and are expected to interpret these results and report the most likely identification on Form I. The quality of the interpretations reported on this summary form is important because the raw data printouts may not be retained by many data users due to space limitations. The accuracy and completeness of these identifications vary widely among laboratories. Problems and inconsistencies result for the users of TIC data (see Section 5, below).

Table 2.1. Summary of TIC Data Study.

TIC Reporting by CLP Laboratories				Data Reviewed in this Study	
Total TICs	Ident. with CAS #	Partial Ident.	Unknown	# of SDGs Reviewed	# of Samples Reviewed
99,513	15,946	40,651	42,916	792	8,078
100%	16.0%	40.8%	43.1%		
12.3 = Average number of TICs reported per sample					
10.2 = Average number of samples per SDG					

3. DATA REVIEW PROCEDURES

The investigators examined existing hard copy data from the CLP to identify and select cases that were likely to contain environmentally significant TICs, based upon results reported under the CLP. Additionally, they evaluated analytical interferences or sample contamination that, while not environmentally significant, will hamper the identification effort of underlying significant TICs. An initial screening of likely data was made using the CLP data package Form I for TICs, resulting in the identification of a good data cross-section.

For this study, "significant" TICs included those that may increase risk, or help to characterize and differentiate wastes by source or effluent stream. Simple aliphatic hydrocarbons, for example, were not considered significant on an individual basis, although in aggregate, they might assist in an identification (e.g., a type of fuel) or in source fingerprinting. It was not the intent of this project to identify all TICs. No further study was made on work conducted by Viar & Co. for EPA.[3] Viar concluded that most TICs are believed to be relatively harmless or do not have any relevant toxicity data for risk assessment. The type of study that they performed necessarily utilizes a subset of incomplete toxicological data, and these general conclusions may not apply to specific situations where a TIC could be found that is "significant" for risk or source determinations. It is well to remember that unidentified compounds with established health and environmental concerns e.g., PCBs, were originally recognized while monitoring for other compounds, as organochlorine pesticides.

Investigators looked for TICs that appeared to contain atoms and functional groups which may be found in environmental contaminants of concern, including P, Cl, Br, F, N, S, PAHs, and organometallic compounds containing heavy metals. Emphasis was placed on seeking spectra with significant intensities of upper mass ions [above ca. 200 mass units, or Daltons (Da)], or masses below 200 Da that are relatively intense in the spectrum and potentially characteristic. Library hits were verified, when questionable, by reviewing the GC/MS raw data file tapes to obtain additional information from the original analyses, such as better quality spectra, better background subtractions, or non-reported spectra. During this process, mass spectral interpretations were augmented by the application of normally used data system algorithms.

3

4. SPECIFIC PROBLEMS IN DETECTING, IDENTIFYING, AND QUANTITATING TICS

In this project, the accuracy of TIC identification and quantitation by the CLP-specified procedures were assessed. This study also evaluated the frequency of problems seen with respect to chromatographic resolution such as the possibility that overlapping GC peaks prevent accurate identifications.

TICs whose areas are over 10% that of the nearest internal standard are approximately quantitated against that internal standard. The assumption is made that the response factors for the two compounds are the same. Several TICs that can potentially appear in sample extracts may be trace components from required quality control (QC) solutions e.g., low percentage impurities or degradation products in surrogates, internal standards, etc. The nature of the TIC definition (10% of internal standard) can mean that irrelevant compounds can become contractually significant in a clean sample. It is important to run associated blanks to determine if TICs are actually of environmental origin.

Chromatographic peak shape and resolution is generally good. However, organic acids do not generally chromatograph well on the GC columns commonly used for EPA analysis of semivolatile compounds. The chromatographic peaks tend to be skewed in a way that frequently causes poor integration by the data systems. These integration errors result in inaccurate quantitations.

Co-elution of large amounts of hydrocarbons with smaller amounts of TICs may occur. If hydrocarbons are present in sufficiently large quantities, the portion of the composite mass spectrum due to the lower-concentration TICs may be given minimal significance or neglected by the computerized matching algorithm.

Gel-permeation chromatography (GPC) cleanup of soil samples is required in the March, 1990 CLP statement of work (OLM01.0) and all subsequent statements of work. This cleanup removes large amounts of high molecular weight hydrocarbons and sulfur from soil sample extracts. However, soil samples containing high amounts of oil or sulfur have the potential of overloading the GPC column. These hydrocarbons interfere with the identification and quantitation of target compounds, and the identification of TICs. High levels of sulfur were also encountered in some water samples.

Instances of reporting and interpretation errors by CLP laboratory personnel were found during the data reviews. A few instances were found of errors such as manual integration of a peak that eluted during an electronic power glitch without adequate discussion of the problem and manual integration of electronic noise as a chromatographic peak. A more common, and significant error found in this study was reporting all TICs as "unknown" even if the data system library match was good. This last situation may be a result of less contractual emphasis on these compounds, where the laboratories save the time and cost that would be expended in TIC interpretations.

The following subsections discuss several scientific issues relating to the CLP method protocol. These issues may cause TIC data not to be reported or limit the mass spectral interpretation specialist in assessing the data.

4.1 Scan range limitations

The CLP SOW currently states that the scan range for the analysis of semivolatile compounds is to be 35 to 500 Da. This can cause the misidentification of compounds that have significant masses or clusters of masses above m/z 500. Certain TICs require special analyses using higher mass ranges. Structure

assignments may be made, or structural features inferred, from fragment ions of TICs having molecular weights over 500 Da and from neutral losses calculated by mass differences among fragment ions or between fragment ions and the molecular ion. Some compounds exhibit characteristic low mass (below 35 Da) fragments, such as C_2 at m/z 24 of highly unsaturated hydrocarbons, m/z 26 from CN, m/z 27 from HCN, CH_2NH_2 at m/z 30, H_2S at m/z 34, and fragments containing C, H, and N or O at m/z 29 and 30. Hence, benefits could be obtained from extending the scan range in either direction, for certain TICs. The CLP has not extended the mass range to avoid detector saturation by air peaks (m/z 28, 32), and in recognition that relatively few TICs of concern having molecular weights over 500 are chromatographed using the protocol method.

4.2 Library deficiencies

TIC mass spectra have been found, including some with recognizable chlorine and/or bromine clusters, for which there are no library spectra in the computerized mass spectral data bases. Such spectra must presently be manually interpreted.

Significant differences in spectral peak intensities have resulted from the use of various tunes, types of instruments, and inlet systems. It is normally not evident to the CLP laboratory or other library data base user what exact analytical conditions were used to generate each NIST reference spectrum. Where seemingly duplicate spectra are present, they may have been generated under different conditions, and one of them may provide a better match than another to the unknown.

Additional compounds could be included in the NIST mass spectral data base to assist in identifying TICs for environmental monitoring efforts. These compounds include higher molecular weight PAHs, and industrial process solvents and chemicals. The inclusion of additional natural products such as vegetation decomposition products would allow the investigator to eliminate such compounds quickly from further study. Additional pesticide metabolites and degradation products should be included; while some are available in the literature (hardcopy form), they are not available in commonly used software data bases.

4.3 Limitations of low resolution quadrupole GC/MS

Some GC/MS spectra do not adequately differentiate between and precisely define the number of chlorine and bromine atoms present in an analyte. This situation was noted in this study when low intensity responses are encountered, when the lowest intensity members of mass clusters are not observed above the noise level, and occasionally when large numbers of these atoms are present and the differences in relative ion intensities are small among the candidate Br_j, Cl_k, and Br_mCl_n clusters. Also, quadrupole instruments commonly employed by laboratories that work under the CLP may not have adequate sensitivity for the detection of some analytes and/or precise measurement of smaller peaks (such as carbon-13 isotopes). Efforts are made through use of logical neutral mass losses in the spectra and manual interpretation to resolve this problem and obtain additional information about TICs. Molecular ions of some compounds are weak or missing, a deficiency that has been discussed by Donald Scott of EPA-NERL-RTP.[4-6]

4.4 Data system capabilities

There are different data processing techniques among the major vendors of quadrupole GC/MS data systems used for environmental analyses. The two firms with perhaps the longest experience in environmental applications for their systems (Finnigan and Hewlett-Packard) provide relatively advanced software to perform necessary data manipulations such as background subtraction and peak enhancements. Data systems from these two companies are used to report approximately 80% of the data submitted to the US EPA under the CLP.

The two systems noted above list the detected masses of chromatographic peaks to two decimal points. This information is occasionally useful for determining whether a given mass spectral peak is due to a fragment containing negative mass defect atoms yielding masses less than the integer mass (chlorine, bromine, fluorine, etc.) as contrasted with hydrocarbon ion series having positive mass defects yielding masses greater than the integer mass. The appearance of mass clusters is more generally useful for the detection of bromines and chlorines in unknowns.

It was found that another data system, from a smaller vendor, had fewer software-based peak subtraction capabilities than the above systems. Although this system reports all masses as nominal masses, a data system command can be used to print the measured masses showing the mass defects. The mass defect is occasionally useful; a defect near 0.5 Da sometimes occurs, generally from presence of multiple bromine atoms in the analyte molecule. The data system may misassign the nominal mass when instrumental variability plus the mass defect combine to reach the 0.5 Da level. In this project, misassigned masses were found in analytes with negative mass defects, such as hexabromobenzene, and with positive mass defects, such as arylalkylamines. In the latter case, the actual mass defect was positive, and less than 0.2 Da.

4.5 Isomer identifications

Many times CLP sample reports list TICs with some non-isomer specific formation such as a "dibromobenzene" because chemical standards to determine relative retention times or retention indices were not available to determine which particular isomer was present. This situation may have significance for health concerns, in cases where different isomers have significantly different toxicities.

5. FINDINGS FROM CLP DATA REVIEWS

Samples were found with computerized library matches to sulfonamides, unusual ethers, steroids, vitamin-E, Cellosolves, thiophenes, acridines, PAHs, carboxylic acids, polychlorinated biphenyls (PCBs), pesticides, oxidized phosphines, natural products, and organometallics. Many of these matches were of poor quality. Some of the spectra in the data packages were difficult to match because of apparently poor spectral quality. The data examined in this study showed a wide variety of MS ions and fragmentation patterns characteristic of classes including long chain dioic (dicarboxylic) acids, from soap products, phthalates, aliphatic hydrocarbons and alcohols, and alkylated benzenes and naphthalenes.

An overview of the data indicated that the most commonly reported classes of TICs were saturated and unsaturated hydrocarbons. The next most prominent groups of compounds were PAHs and aromatic compounds which were frequently substituted with alkyl groups. Higher molecular weight steroid

compounds (i.e., cholesterol) and PCBs were also reported. Elemental sulfur was reported by the CLP laboratories in ca. 50% of soil samples. The TCL phthalates were found in ca. 80% of the soil samples, and non-TCL phthalates were found in ca. 10%. Of all the data reviewed, there was only one data case in which an organometallic compound (tetraphenyltin) was found.

Tentative identifications by the CLP varied from the uninformative "unknown" to the more informative, e.g., "trimethylnaphthalene isomer" or "unknown trimethylnaphthalene." Some TIC spectra were reported as unknown with the base peak m/z listed in the CLP report summary (Form I-F). The following are examples of reported items that were considered in this study to be reported as unknown:

Unknown
Laboratory artifact
Laboratory contaminant
Unknown (base peak m/z = 43)

The following are examples of reported items that were considered in this study to be partially identified:

Unknown aromatic
Trichlorobenzene isomer
Unknown hydrocarbon

The reporting trends from the laboratories were highly varied with respect to accuracy of reporting TICs. While some laboratories (ca. 2%) made honest efforts to identify the TICs, others (ca. 30%) simply labeled all TIC peaks as unknown and made no attempt to accurately identify the TICs. Some laboratories (ca. 2%) would only commit to identifying sulfur and reported all other TICs as unknown. Some laboratories (ca. 20%) used the name of the library match compound with the highest score on the library search, treating that as the identification regardless of whether that identification was reasonable based upon manual spectral interpretation. It was also noted that the method of reporting TICs varied among personnel within a given contract laboratory. There are several examples of all TICs being reported as unknown for half of the samples within a single SDG while efforts were made to identify or partially identify TICs for the remaining samples in the same SDG. These percentages indicate that the reporting of TICs is not emphasized by CLP, and the quality of TIC identifications appears to be independent of laboratory selection and payment for analyses.

In some cases, the TIC identity is known, but unreported. For example, two isomeric methylnaphthalenes exist; 2-methylnaphthalene is a Target Compound; 1-methylnaphthalene is a TIC. Their mass spectra are not distinguished by the data system search, but their GC RRTs are different, and the RRT for the TCL analyte was measured in the laboratory TCL standards analysis. Many labs report the TIC isomer as "unknown PAH" although they have enough information for a correct identification. Non-TCL phthalates were found as TICs in ca. 10% of the samples, but these compounds were not emphasized in this project. They may be present in the original samples or may be laboratory artifacts, but they are not usually considered hazardous.

The hard copy (paper) data submitted by the laboratories varied in content as well. The current SOW for organic analysis by the CLP requires that the laboratory submit the spectrum of the unknown followed by the spectra of the three best matches (if present). All laboratories do this but the format which is submitted often lacks information needed for further review or manual spectral interpretation. These

problems include spectra that have no or minimal labeling on the m/z axis, making it difficult or impossible to accurately determine the correct m/z value for a given ion. Names of spectral matches are often too long to present the entire name of the match on the graphical display. If the CAS number is not present (this occurs about one third of the time), the reviewer cannot determine the full name or exact structure of the compound.

6. DATA SYSTEM CAPABILITY STUDIES

GC/MS data systems used in this project included: HP RTE, HP DOS, HP UNIX, Finnigan INCOS™, VG DOS, and Extrel DEC. The HP and INCOS™ data systems were interconnected by a personal computer-based local area network (LAN) equipped with MassTransit™ Version 1.02a (Palisade Corp., 1993) and Excel Version 5.0 (Microsoft Corp., 1994) software.

Senior level chemists applied existing data system procedures to aid in the manual or automated identification of difficult or complex TICs from the existing CLP data. The procedures available included the following techniques: background subtraction, spectral enhancement using predefined data system algorithms combined with background subtraction, optimization of certain mass ranges, global and local normalization factors or tilting, changes in search speed, minimum and maximum molecular weight, minimum number of ions to search, comparison of retention times, and various types of spectral cleanup and background subtraction before library searching.

TICs that can cause analytical interferences were evaluated. These compounds varied in direct environmental significance, but they also can present difficulties in evaluating other closely eluting TICs. The feasibility of applying or designing computer algorithms that subtract these interfering TIC spectra was investigated, resulting in the development of an algorithm based on Colby's concept for spectral "deconvolution" or spectral enhancement.

We studied related, recent work on compound identification for applicability to the goals of this project. For example, Donald Scott (EPA-NERL-RTP) has developed a molecular weight estimator program;[4-6] Bruce Colby (Pacific Analytical Labs) has studied the concept of spectral deconvolution for overlapping peaks;[7-9] Stephen Stein (NIST) has studied the probabilities of correct identifications from mass spectral library searches,[9] and has tested search algorithms for compound identification.[10]

6.1 NIST and Wiley libraries

The Hewlett-Packard BigDB EI-MS library (130,000 spectra; 1986 version) was used for most library searching. This library includes the 49,000 entry NIST and 81,000 entry Wiley libraries, both also EI mass spectral libraries. The 75,000 entry NIST library (1992 version) was also used for searching. Numerous cases were noted where no library match was found by the data system, or the library matches were of such poor quality that the databases were ineffective in correctly identifying the TIC. Since the time of this study, these libraries have been expanded by the inclusion of new entries.

Three features of the data bases are noteworthy: 1) there are cases of multiple spectra present for the same CAS registry number (replicate spectra); 2) there are numerous instances where ions are present in the reference spectra at m/z values higher than the molecular ion mass cluster for the molecule. These higher mass ions indicate the presence of impurities or incorrect spectra; 3) quality indices (QI) are assigned to spectra in the data base. Factors 1) and 3) are qualitatively useful, and data interpreters can use them to augment manual mass spectral interpretations of unknown TICs against library spectra.

The Hewlett-Packard BigDB library has the advantage that it contains additional compounds not present in the NIST data base. However, searches using the HP software for compounds present in both libraries were found to be similar in effectiveness to searching the NIST library alone,. This situation indicates that spectral quality and accessibility via computerized searching are similar for the two libraries. The HP data system output did not specify whether selected spectra were from the Wiley or NIST data bases within the BigDB library. However, searching a larger library did take more data system time, and often gave more apparent matches that were caused by the presence of duplicate spectra of the same compound. Library searches using unique masses tend to be fast, whereas those using non-unique masses, such as those of aliphatic hydrocarbons, take longer due to the presence of many spectra in the data base having those masses in the spectra.

The distribution of compounds contained in the NIST 75K database as provided on the HP UNIX GC/MS data system is shown in Table 6.1.

It can be seen that about 4 % of the compounds present in the NIST library have molecular ions above the CLP-required scanning range (35-500 Da) and about 17% are duplicate spectra. For example, the 75K version of the NIST database includes nine spectra for decafluorotriphenylphosphine (DFTPP), four spectra for cholesterol, and five spectra for cholesterol trimethylether. Of the three spectra for hexachlorophene, one does not contain the molecular ion/isotopic molecular ion group (m/z 404 . . . 416), but rather a single low-intensity peak at m/z 407 (entry #74,279). The presence of multiple spectra for a compound results in more than one of the top three or five hits that are printed by the data system being the same compound, rather than different possibilities. This situation results in the data package containing fewer unique possibilities for the TIC identity, and the data user does not have potentially valuable candidates against which the TIC identification could be reinterpreted.

Some hydrocarbon spectra are incomplete, showing only the lower intensity but important ions above m/z 300 (e.g., hentriacontane and 3-methylhentriacontane) or m/z 150 (e.g., tritriacontane, 2-methyldotriacontane, 3-methyltritriacontane). Other hydrocarbon spectra are complete, showing the characteristic saturated hydrocarbon ions such as m/z 43, and 57. This presentation results in the lower-intensity responses for the high mass ions difficult to observe (e.g. tetratriacontane). Both the full spectrum and the partial higher mass spectrum are useful for certain purposes, the former for the molecular weight, and the latter for verification that the low mass saturated hydrocarbon pattern is present.

Because these mass spectral databases are intended to serve multiple purposes, they include compounds that are intractable (not suitable for introduction through the GC-peptides and other naturally occurring biological molecules of high molecular weight), and compounds that are not generally found at hazardous waste sites. The large databases increase the amount of time it takes to perform computerized library searches. A shorter subset of the data base without duplicate and aliphatic hydrocarbon spectra would reduce search times perhaps 15%. However, much of the time involved in using the larger library does not impact the cost of analysis, because searches are usually performed automatically. Labor costs are incurred for the time needed to interpret the results, prepare the summary form (Form I), and to make duplicate copies of the data package. The added costs should be balanced against the potential that an unexpected compound may be found in a sample, and could be identified if the data base were larger. As an example, a series of arylalkylamines was identified in a CLP-RAS case, partially because one member of the series was present in the 104,000-entry Wiley library. This library was searched in an effort to identify this series of TICs whose elemental compositions were known from high resolution mass spectrometric (HRMS) accurate mass measurements. Generally, when attempting to determine the chemical structure of unidentified compounds,

9

the largest possible library increases the probability of directly matching the compound or of recognizing substructural features from compounds having these in common.

Table 6.1. Molecular weight ranges of compounds in the NIST data base.

Molecular Weight Range	# of compounds	% of Database
15-34	19	0.03
35-99	1407	1.88
100-199	21274	28.43
200-299	19959	26.67
300-399	11187	14.95
400-450	3229	4.32
451-499	1840	2.46
500-599	1936	2.59
600-699	799	1.07
700-799	338	0.45
800-899	100	0.13
900-999	51	0.07
1000-1260	57	0.08
Duplicate spectra	12631	16.88
Total	74828	100.00

6.2 File formats

At present there are 36 file formats being used by various GC/MS data system vendors. The five most common formats applied to environmental analysis are the HP RTE, HP DOS, Finnigan INCOS™, VG DOS, and Extrel DEC formats. Due to differences in these formats, it has been difficult to compare quantitation and library search procedures between various software packages and hardware platforms. This comparison is difficult even with two data formats arising from the same GC/MS data system vendor.

A universal GC/MS data format (netCDF) has been proposed, although the various instrument vendors have not fully agreed on its exact format. This data file standard is proposed by the International Association of Environmental Testing Laboratories as a universal file format for exchange of raw GC/MS data between different GC/MS equipment vendors. The common format will enable transfer of GC/MS data

files to different hardware platforms. This file format will allow processing of any one raw GC/MS data file using software from multiple vendors. To date, the netCDF format is standardized for GC data but not for GC/MS data. Converting netCDF to ASCII text with public domain software would allow for the potential of data fraud, by a lab manipulating the data because it is in the flexible, editable ASCII format. In addition to this concern, another limitation to the use of netCDF is that even if this format becomes available, additional software may be needed that would not be compatible with older systems.

The issue of file formats is significant for studies of mass spectral data system capabilities. Prior to this project, little work had been done to compare the various search algorithms for accuracy in identifying TICs found during the analysis of environmental samples. Most laboratories lack the ability to interconnect dissimilar data systems by means of a local area network (LAN) that could transfer data files and convert them from one format to another. Each data system requires that the data file be in the proprietary format used by that manufacturer. Therefore, we obtained information about available algorithms and compared the results of using different algorithms. To make such comparisons on identical data sets, we developed the capability to move files from one system to another through a DOS-based PC network, and subsequently interconvert file formats.

6.3 Background Subtraction

The use of software subtractions was investigated for TIC spectra in samples that were heavily contaminated with long-chain hydrocarbons. It was found that high concentrations of hydrocarbons would render such subtractions of minimal value. The presence of unsaturated and saturated hydrocarbons would further complicate the situation.

Software supplied with most systems for the analysis of environmental samples have simple background subtraction techniques for the removal of moderate interferences. The Finnigan system allows for the subtraction of a single scan, multiple scans on one side of the target peak, or background subtraction from both sides of the target peak if there are multiple interferences present. The HP system has similar capabilities. In this study, it was found that both of these systems were fully satisfactory.

6.4 Isotopic ratio and elemental composition program

A program written by Dr. Andrew H. Grange, currently National Research Council/NERL-ESD-LV Senior Research Associate, shows the correct peak cluster ratios for any combination of chlorine and bromine in the range from 0 to 10 bromines and 0 to 16 chlorine atoms. This program also calculates candidate elemental compositions of a molecular ion and isotopic molecular ions containing bromine and chlorine atoms. This program was used to support HRMS accurate mass determinations.[11,12]

7. ALGORITHMS AND PROCEDURES FOR TIC IDENTIFICATIONS

The process for determining TICs in mass spectral data involves two separate and distinct operations. First, a peak in the gas chromatogram must be isolated from other interfering compounds, and a representative mass spectrum must be obtained for the peak. Second, the mass spectrum corresponding to the peak must be searched against a mass spectral data base. There are two common methods currently being used to isolate and remove interferences in mass spectral data. These are application of manual background subtraction procedures and the use of various algorithms to isolate the spectra of interest.

7.1 Background Subtraction

The simplest method of mass spectral enhancement is background subtraction. This method was useful for the elimination of background responses but provided limited success when two or more components coeluted.

7.2 Biller-Biemann Spectral Isolation Algorithm

A more advanced background substration type of method for the isolation of spectra is the Biller-Biemann process, available on Finnigan and some other data systems. The data system searches the mass chromatogram and flags mass spectra where a set number of ions maximize within a set number of scans of each other. Typically the number of maximizing masses is three and the scan range is two scans. After this has been done, these scans are processed using an enhancement routine that discards all other ions that do not maximize within that retention time window. This algorithm has been used commercially on Finnigan™ data systems for about 20 years with good success.

7.3 Mass Spectral Enhancement

Fourier transform of data from intensity vs. time to intensity vs. frequency was tested. It was found that the required frequency resolution was not readily available. A two-dimensional package might prove useful, but the only commercially available packages were one-dimensional and were found to broaden the chromatographic peaks, decreasing the apparent resolution rather than increasing it as would be desired. A quadratic fit was found to be key (see below) in achieving the desired spectral enhancement and background rejection.

A method of mass spectral enhancement via background rejection was developed in this project, employing a concept proposed by Bruce Colby of Pacific Analytical Laboratories.[7-8] Colby suggested a resolution enhancement approach related to the Biller-Biemann process. It includes features that resemble a digital implementation of the widely used phase-locked amplifier in electronics.

This method of background rejection and spectral enhancement for the identification of unknowns is a potentially valuable substitute for conducting further analyses of the samples. The capabilities of modern personal computers makes it worthwhile to consider the use of mathematical methods for spectral enhancement and background subtraction to isolate the spectra of unknown compounds of interest from interferences. This is especially evident when there are large numbers of samples present from a site containing numerous interfering compounds.

In this study, we considerably expanded, developed, and implemented the resolution enhancement approach as a set of macros in a Microsoft EXCEL™ spreadsheet.[13] A description of the resolution enhancement approach follows below. Results of applying these algorithms to environmental samples are presented in Section 8.

The spectral enhancement approach was designed and implemented as a set of macros written in Visual Basic for Applications in Microsoft EXCEL™. Mass spectral data were accessed from a Hewlett-Packard data system and the National Institute of Standards and Technology (NIST) mass spectral data base. The data files were converted to ASCII text with MassTransit™ Version 1.02a (Palisade Corp., 1993) and imported into EXCEL™ Version 5.0 (Microsoft Corp., 1994). The intensity maxima (peaks) for each ion in

12

a mass chromatogram were determined and their raw retention times (scan numbers) were noted. The raw scan numbers were adjusted to fractional scan numbers for each m/z to match the precise time during a scan when that mass was measured. This adjustment was expressed as an offset term to the scan number, O_S, defined as follows:

$$O_S = (M_{Curr} - M_{Min})/(M_{Max} - M_{min}) \tag{1}$$

where

M_{curr} is the current mass
M_{min} is the minimum mass sampled, usually m/z = 35
M_{max} is the maximum mass sampled, usually m/z = 500,

The numerator represented how far into the scan the mass occurred and the denominator represented the full scan range. Hence, for linear scanning, O_S represented the simple fraction of the full scan completed at the given mass.

Near the apex of a mass chromatographic peak, the peak generally appeared to be reasonably parabolic, and could be represented by the following quadratic equation:

$$Y = aX^2 + bX + c \tag{2}$$

where

Y is the intensity
X is the retention time (scan numbers)
a, b, and c are constants to be determined.

The quadratic form was exactly fitted to each peak ion intensity and the intensities of the preceding and succeeding scans. By selecting these three intensities, the three coefficients could be exactly determined.

The retention time of the apex of the fitted curve was expected to be an accurate estimation of the true retention time of the constituent of interest, within constraints imposed by signal noise from scan to scan. This apex was readily found using the common technique of taking the first derivative of the quadratic function and setting it equal to zero, as follows:

$$Y' = 2aX + b = 0 \tag{3}$$

where

Y' is the first derivative.

A variety of methods, including packaged optimization routines in EXCEL™, were available for solving this algebraic problem. A fast, extremely simple axis-translation solution to the problem was selected that reflected the simplicity of the mass chromatographic data in a spreadsheet format. A peak intensity, I_O, presented in scan co-ordinates in the spreadsheet, was the origin for the fitted quadratic equation, and each of the adjoining intensities was set to coordinates of 1 and -1, respectively, with values of $I_{R(ight)}$ and $I_{L(eft)}$. These changes of variable provided a simple closed form for the offset of the apex, O_A, in scan coordinates:

$$O_A = -(I_R - I_L)/2(I_R + I_L - 2I_O) \tag{4}$$

13

The optimized retention time for the mass was obtained by adding O_S and O_A to the raw scan number of the peak. Once a peak was identified, the entire optimization procedure was compactly performed by a single line of code in the macro that contained only one multiplication and two division operations.

The quadratically optimized retention times, masses, and observed peak intensities from the mass chromatograms were stored in a list as they were calculated. After all of the peaks were optimized, the list was sorted with respect to retention time. The intensities of all masses falling within selected sequential retention time windows (summing intervals) over the entire chromatographic time range were summed and placed into a second list of retention times and intensities. The summing interval duration was set at the beginning of the experiment, typically between 0.1 and 0.33 scans. The selection of a summing interval shorter than one scan yielded a total ion current chromatogram with greatly enhanced resolution, showing distinct, baseline resolved peaks under the high levels of background signals. The mass spectra obtained for these mathematically resolved peaks were free of most background mass responses because few of the background masses maximized coincidentally with the masses of the peaks under investigation.

After the spectral enhancement procedure was complete, a peak of potential interest was selected, giving the background-rejected mass spectrum. The mass spectrum was exported with the appropriate header to a text file. This file was imported into a mass spectrometer data system through a program such as MassTransit™, and the mass spectrum was searched against a reference data base such as the NIST mass spectral library. Table 7.1 outlines the steps required to perform a spectral enhancement analysis of a mass chromatogram.

Table 7.1. The steps performed in the spectral enhancement operation.

Step	Action
1	Convert the raw data into text format.
2	Parse data using EXCEL™ version 5.0 into a two dimensional matrix placing the reported intensity for each ion present in a cell indexed to the mass and scan numbers.
3	Start a sliding window looking for all ions that are present in three successive scans, and identify each case where the ion maximizes in the center of the window.
4	Adjust the nominal scan time for time lag due to instrument scanning, O_S.
5	Adjust the peak retention to observed ion distribution using quadratic fit, O_A.
6	Place the time corrected data in a new matrix and sort with respect to the adjusted retention time.
7	Apply a filter to group ions based on adjusted retention times and generate a resolution enhanced chromatogram.
8	Extract ions within a given time range (peak) to produce a mass spectrum with enhanced background rejection

7.4 Biller-Biemann Library Search Algorithm

The Biller-Biemann type library search algorithm currently used by Finnigan and Fisons (formerly VG Masslab) uses a "sliding window" to determine the 16 most significant ions present in the unknown. This window is typically 20 Da wide. This width is meant to ensure that all ions contained in characteristic tightly-grouped mass clusters (such as from chlorine, bromine, or patterns exhibited by heavy metals) are retained as significant peaks to be searched against the reference spectra contained in the data base. The algorithm also uses peak intensity weighting. It multiplies the intensity of each peak by its mass number to give higher priority to low intensity peaks at higher masses. This weighting is especially important when the molecular ion is of low intensity. The algorithm then selects the 16 most intense peaks, and searches this reduced "chemically significant" mass spectrum against a condensed version of the specified library that has the 16 largest "chemically significant" peaks of every entry. After determining the 20 best candidate spectra, the algorithm then searches the full unknown spectrum against the full spectra of those 20 best candidates from the library. The ranking can be performed against quality of fit, reverse fit, or no fit, at the discretion of the user. During the expanded search, the algorithm again adjusts or weights the experimental ion intensities by multiplying them by the corresponding m/z value, increasing the significance of low intensity high mass ions.

During the main search, the fit (FIT), reverse fit (RFIT), and purity (PUR) are determined for the complete experimental spectrum against each of the candidate spectra. The FIT ranking rates the degree that the library spectrum is present in the unknown. The RFIT ranking rates the degree to which the unknown spectrum is contained in the library spectrum. The PUR ranking rates the resemblance of the unknown to the library entry. The three parameters FIT, RFIT, and PUR are scaled using values which have a range of 0 to 1000. Clean spectra usually produce matches that have numerically high FIT and PUR values. Spectral search results with high FIT but lower PUR indicate the presence of coeluting compounds or interferences that have not been adequately subtracted out. It is left to the data system operator to decide whether the library search results will be ranked by FIT, RFIT, or PUR.

7.5 Probability-Based Matching Algorithm

A second library search algorithm, used primarily by Hewlett-Packard and Extrel, involves determining the uniqueness of the ions present in the unknown spectrum. This algorithm was developed by Dr. Fred W. McLafferty at Cornell University and is called probability-based matching (PBM). Initially all ions present in an unknown spectrum are assigned uniqueness values based on the number of times each mass occurs in the NIST data base. The unknown is then compared against all entries in the NIST data base. According to the HP 59872 RTE MS Data System Manual, this algorithm is based on the fact that the probability particular ions will occur follows a log normal distribution and that the probability of finding higher mass ions decreases by a factor of two each increment of 130 mass units.

The PBM algorithm uses only a reverse search to determine the ranking of NIST candidate spectra against the unknown. A reverse search means that each library spectrum is compared against the experimental spectrum to determine if the library spectrum is contained in the experimental spectrum. The PBM procedure is significantly different from the Biller-Biemann technique used by Finnigan and Fisons; however, the output is similar to that produced by the Finnigan method. The various parameters that are reported and are useful include the following: Prob--the probability that the NIST data base spectrum matches the unknown spectrum; K--the confidence factor, from 15 to 250, with a high number indicating

great similarity between the unknown and the library entry; dK-the difference between a perfect match and the confidence factor K, with a low dk value generally indicating a good match.

The McLafferty PBM algorithm is currently in use by Hewlett-Packard and Extrel for the identification of unknown compounds. This algorithm uses a filter to reduce the number of ions present in the unknown spectra to a subset of between 15 and 26 chemically significant ions by eliminating three ions (m/z 18, 28 and 32) and then eliminating fragments that represent illogical neutral losses (i.e., loss of 9 Da) from the spectrum of the unknown. The intensities of the remaining ions, including the molecular ion, are weighted by their masses, as occurs in the Biller-Biemann procedure. The mass peaks are assigned values based on the probability of their occurrence (uniqueness values) in the mass spectral data base as a whole. The filtered spectrum is used to search a condensed subset of the NIST and/or Wiley data base to choose candidate spectra that will then be matched against the full mass spectrum of the unknown. The final comparison uses all peaks present in the original spectrum against the full spectra of the candidate matches present in the reference database.

The PBM algorithm offers the advantage of fast search speed, especially for compounds that have high-intensity molecular ions above 300 Da. The tilting function accommodates differences in instrument tuning and variations caused by differing instrument types (magnetic versus quadrupole). This algorithm performs less well for compounds having low molecular weights (e.g., hydrocarbons) and compounds that have very few ions (e.g., acetone), as discussed below.

7.6 Performance Comparison of PBM and Biller-Biemann Algorithms

A comparison of the two library searching algorithms was performed on 20 CLP data files. Library searches were conducted with an HP DOS data system using PBM and with a Finnigan system using the Biller-Biemann algorithm. The results demonstrated that the two algorithms performed to the same level of quality and reliability for compounds in the molecular weight ranges associated with semivolatile TICs.

The PBM algorithm performed less well for low molecular weight compounds with few ions because PBM uses 16 to 25 ions for pre-searching and matching, and because the uniqueness values associated with low mass fragments are usually small.

7.7 Normalization and Tilting Algorithms

The library search algorithms include the optional feature of global and local normalization factors (Finnigan) or tilting (Hewlett-Packard) for matching the library spectral relative intensities against the unknown spectrum. This feature adjusts the spectrum of the unknown to matches in the library database.

Global normalization multiplies all ion intensities in the unknown spectrum by a global normalization factor to make the average intensity of the peaks similar to those of the library entry. In a second step (local normalization), individual peak intensities in the unknown spectrum are normalized to the corresponding peaks in the library spectrum. This system does not alter peak intensities by more than a factor of two, and peaks that are not found in both spectra are not normalized. This procedure is intended to correct for mass spectral differences that may occur when the same chemical compound is analyzed under different experimental conditions.

16

In the Hewlett-Packard tilting procedure, a comparison is made between the library spectrum and the unknown. The coefficients for a quadratic equation are determined to provide normalization factors for the best fit of the peak intensities in the library spectrum to those of the unknown. The library entry is then rescaled using these factors, and library search results are reported using the rescaling coefficients that gave the best results. Unlike the Finnigan data system that adjusts the unknown mass spectrum to fit it to the library entries, HP adjusts the library entries to get them closer to the experimental mass spectrum of the unknown. The two procedures provided similar results in this project.

8. ADVANCED DATA HANDLING STUDIES

One goal of this work was to investigate the option of identifying TICs that were reported as unknown using computer software and related studies of the data, as opposed to conducting further analyses of the samples. For highly contaminated samples, it has traditionally been necessary to re-extract the sample using different procedures, to perform alternate clean-up methods on the extracts, or to use highly specialized instrumentation to remove interferences and separate the TICs of interest from mass spectral contaminants. These procedures are time consuming and costly. In addition, it is only possible if samples or extracts are available.

The capabilities of modern personal computers makes it worthwhile to consider the use of deconvolution algorithms, background subtraction techniques, and other mathematical methods to isolate the spectra of unknown compounds of interest from interferences. This is especially evident when there are large numbers of samples present from a site containing numerous interfering compounds.

After a given data file is translated into a suitable format, spectral isolation or enhancement can be performed in a matter of minutes as opposed to the hours of time necessary for re-extraction/dilution, sample cleanup, and concentration. If an effective method of spectral deconvolution were generally available in an appropriate software format, the time necessary for the identification of many TICs could be reduced from several labor hours to about 20 minutes. This advantage is particularly useful when limited amounts of the initial sample are available.

In this study, advanced data system and aftermarket software-based procedures were applied to archived data to provide a more complete assessment of current CLP TIC reporting status. We purchased the commercial MassTransit™ software and tested its ability to facilitate converting data files into formats usable by other data systems, including spreadsheets on personal computers. Interfacing the different MS data system formats through this software effectively accomplished a standardization of TIC data.

8.1 MassTransit™

MassTransit™ is a commercial software product developed by Palisade Corporation. It accepts data files in 36 different GC/MS data file formats (Table 8.1) and produces output files in any of the seven formats listed below in Table 8.2.

Table 8.1. Data file input formats, including commercial brands, supported by MassTransit™

Anelva AGS-7000	JEOL Mario
Anelva DOS	JEOL DA50000-DA7000
Balzers Quadstar 420	JEOL CAMP
Balzers Quadstar 421	Kratos DS90
EPA	Kratos MACH3
Finnigan INCOS	MASPEC
Finnigan ITS40	Nermag SIDAR
Finnigan ITS80	netCDF
Finnigan MAT SS300	Netzsch
Fisons/VG MassLab	Palisade
Fisons/VG Lab-Base/Trio	Perkin-Elmer Qmass 910
Fisons/Thermolab	Shimadzu PAC200
Fisons/VG 11/250	Shimadzu QP-5000
Fisons/VG JCAMP	Shrader System
Hitachi	Teknivent Vector/1
HP Chemstation	Teknivent Vector/2
HP RTE	Text
JEOL Complement	Varian Saturn

Of these input formats, INCOS™ (Finnigan), and two Hewlett-Packard formats were available for testing through the DOS-based PC network. The network connection was needed to convert the data into a personal computer DOS-based format. The Fisons data system was not connected to the network, so this format was not tested in this study. Note that the Finnigan ITS40™ and ITS80™, and the Varian Saturn™ formats are ion trap data systems.

Table 8.2. Output data file formats, including proprietory commercial, supported by MassTransit™

HP Chemstation™
netCDF
EPA
Text
Palisade
Fisons/VG JCAMP
Teknivent Vector/2

For these studies, the HP Chemstation™ and text output formats were used. Concerning the other MassTransit™ output formats, netCDF is not finalized; EPA format is for 9 track tape storage and is not a true data system format for data handling. The Text format allows reading but is not used by data systems. However, we used the Text format output files to transfer ASCII data to EXCEL™ spreadsheets to perform spectral enhancement analyses. Palisade is a proprietary MassTransit™ format. The Fisons and Teknivent formats were not available for data processing in this study.

We tested the capabilities of MassTransit™ on a PC-based local area network (LAN), taking raw data files from a Finnigan INCOS™ data system and converting them to HP Chemstation™ DOS-based data system file formats. A raw Finnigan data file from a Finnigan INCOS™ data system was converted to EPA data file format using the Finnigan EPA utility program. The file was then downloaded onto a PC using the trivial file transfer protocol (TFTP) and converted to the HP Chemstation™ data format using the MassTransit™ software. Next, the file was transferred over the LAN using file transfer protocol (FTP) to the HP UNIX data system, where library searching was performed with the PBM algorithm.

8.2 Molecular weight estimation

Lockheed requested and received information and the program for molecular weight estimation from Donald Scott (EPA-NERL-RTP).[4-6] Our preliminary review of this program indicated that it was most useful for low molecular weight organics (e.g., gas sample analysis). In some cases, the molecular ion was estimated

only to within several Daltons. This program in its present state of development was less useful for higher molecular TICs, where the correct mass of the molecular ion was very important for identification.

8.3 Application of Colby's concept

The original reports on mass spectral enhancement and background signal rejection techniques demonstrated their potential on solutions of known reference standards.[7-8] In the present study, subject data files on actual environmental sample analyses were translated into a format suitable for a personal computer spreadsheet. Spectral isolation or enhancement was performed in a matter of minutes, as opposed to the hours of time necessary for re-extraction, sample cleanup, and concentration.

The capabilities of this method were used to resolve a complex chromatogram that appeared to have several indistinct and broadly eluting components into a highly resolved elution pattern containing potentially significant components, separated from the background contamination. In Figure 8.1, pollutants having unique mass spectral features were separated from broad aliphatic hydrocarbon background signals that eluted across about 20 scans. The procedures were used also to increase the quality of mass spectra selected as a result of inspections of these chromatograms. Improvements in mass spectral quality were measured in terms of the increase in quality of mass spectral fit parameters reported by the data system performing the library search.

We converted GC/MS data from various contractors into ASCII text format using the MassTransit™ software as discussed above. These data files were subsequently imported into an EXCEL™ spreadsheet and manipulated using a macro to perform the sorting and the statistical and mathematical procedures necessary to separate the selected peaks from interfering compounds. The resulting total ion current mass chromatograms were evaluated to determine whether the duration of summing intervals significantly affected the results. Figures 8.1a through 8.1d show the effects of varying the summing intervals from 0.05 scan to 0.5 scan. Summing intervals in the range of 0.05 to 0.1 scan did not provide sufficient integration of individual mass responses for reliable detection of small peaks, such as the one in the range of 7.0 to 8.0 scans (denoted #1 in Figure 8.1). Using a scanning interval of 0.33 scans, this peak was readily observable. At the wider summing intervals such as 0.5 scan, closely eluting peaks such as those occurring in the scan range of 131 to 134 (denoted #2 in Figure 8.1) were not resolved adequately. These peaks were readily seen to be distinct with a summing interval of 0.1 or 0.2 scans. Based on these observations, the optimum summing interval seemed to be between 0.2 and 0.33 scans. Under these conditions the correct elution width of a pure compound needed to contain all of the compound's ions can range up to approximately 1.5 scans, as demonstrated by the enhanced mass chromatograms of indeno(1,2,3-cd)pyrene and dibenzo(a,h)anthracene shown in Figure 8.2. In another case, using a 0.2 scan summing interval, the EXCEL™ macro successfully baseline-resolved 10 known components that eluted within a period of 20 scans.
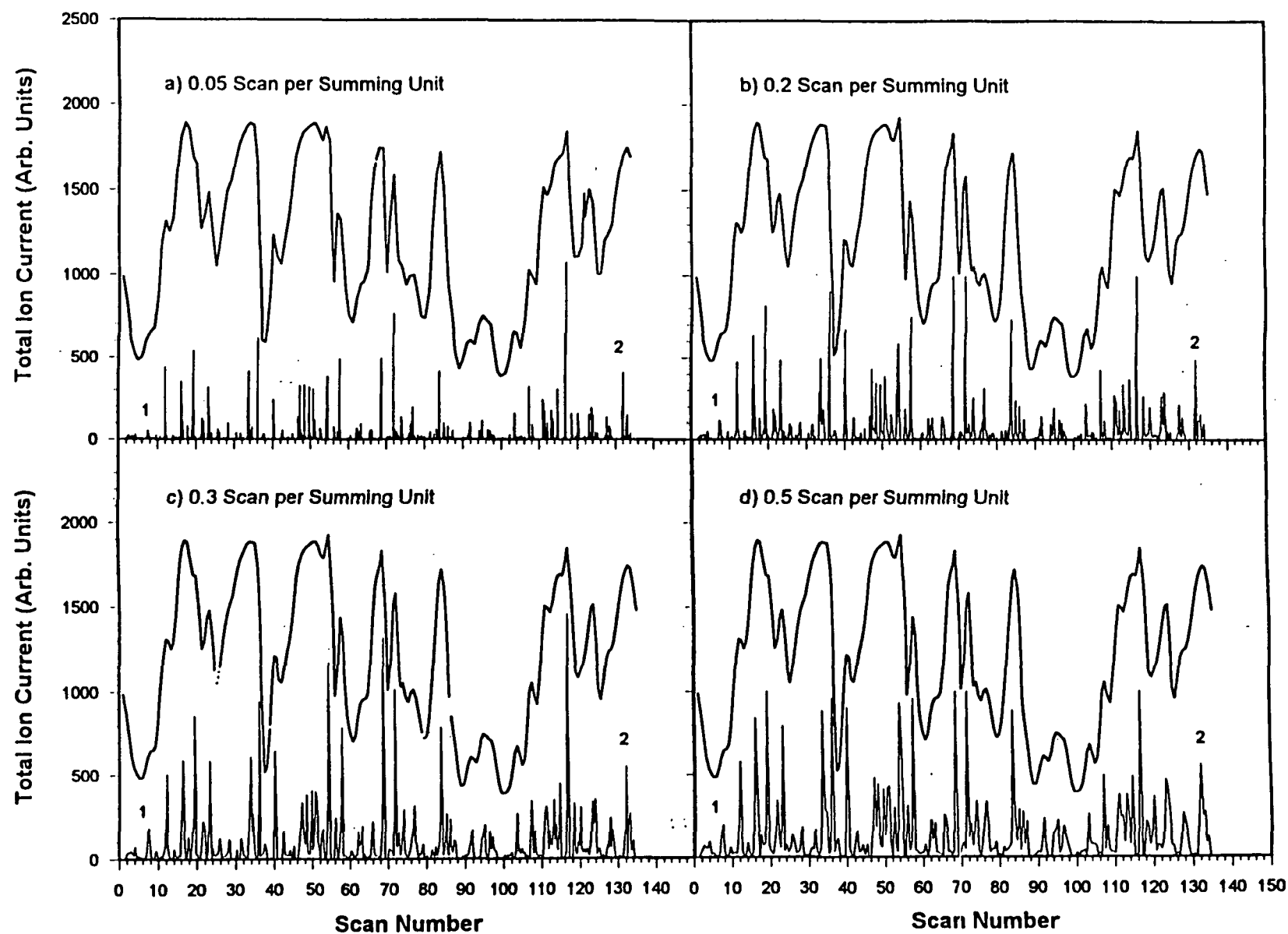
Figure 8.1. Comparison of native and enhanced total ion current chromatograms demonstrating effects of different summing intervals.
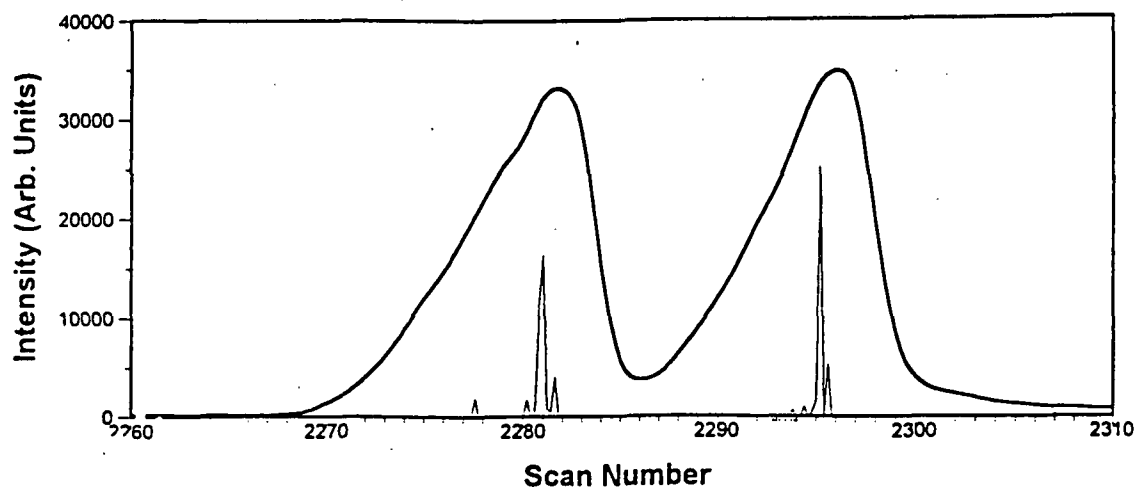
Figure 8.2. Enhanced total ion chromatogram of indeno(1,2,3-cd)pyrene and dibenzo(a,h) anthracene showing natural peak widths of late-eluting compounds.

Initial tests of the spectral enhancement algorithm were performed using standards containing known coeluting compounds to check the technique for accuracy in separating the spectra of coeluting analytes. The extracted mass spectra were compared with reference spectra contained in the NIST data base. Results of these comparisons showed that usually 80 to 90 percent of the ions in the experimental spectrum that were common to ions in the reference spectra were successfully extracted using this method. The ions that did not extract well were those with a low ion current and a relatively high noise level.

A dramatic demonstration of the method's ability to extract improved mass spectra from a chromatogram with high background signal is shown in Figure 8.3. This example, based on the mass spectrum of naphthalene, showed the value of the spectral enhancement technique for library searches. The enhanced total ion chromatographic peak profile shown in Figure 8.3.a predicted that the mass spectrum was contained in the scan range 17.2 to 18.0. The mass spectrum in Figure 8.3.b was obtained from a mass spectral data system by using the standard background subtraction technique of averaging the three spectra at the peak max and subtracting the average mass spectrum from the two adjoining minima. The data system was unable to identify this mass spectrum. The enhanced mass spectrum shown in Figure 8.3.c (scan range 17.2 to 18.0 scans) was correctly identified as naphthalene by the data system, which also produced the reference mass spectrum in Figure 8.3.d. The failure of the standard procedure appeared to be heavily dependent on the absence of the masses at m/z 127 and 129. The artificial enhancement of the peaks at m/z 50, 51, 61 to 64, and 74 to 78 did not hinder the search routine, because it strongly weighted the apparent molecular ion in making identifications.

Other examples of improved mass spectral quality were provided by the spectral matching quality indicators generated by the mass spectral data system for background subtracted mass spectra directly extracted by the data system against those that were enhanced and extracted using the spectral enhancement, background rejection technique described in this report. The comparisons shown in Table 8.3 utilized the data for the mass chromatogram in Figure 8.1. Both of the examined fitting quality parameters showed small but systematic improvements after subjecting the data to the spectral enhancement procedures. Such

22

apparently small improvements can, however, provide significant improvements in mass spectral library search results, as has been demonstrated on solutions of standards.[8]

Table 8.3. Comparison of native and enhanced mass spectral quality indicators.

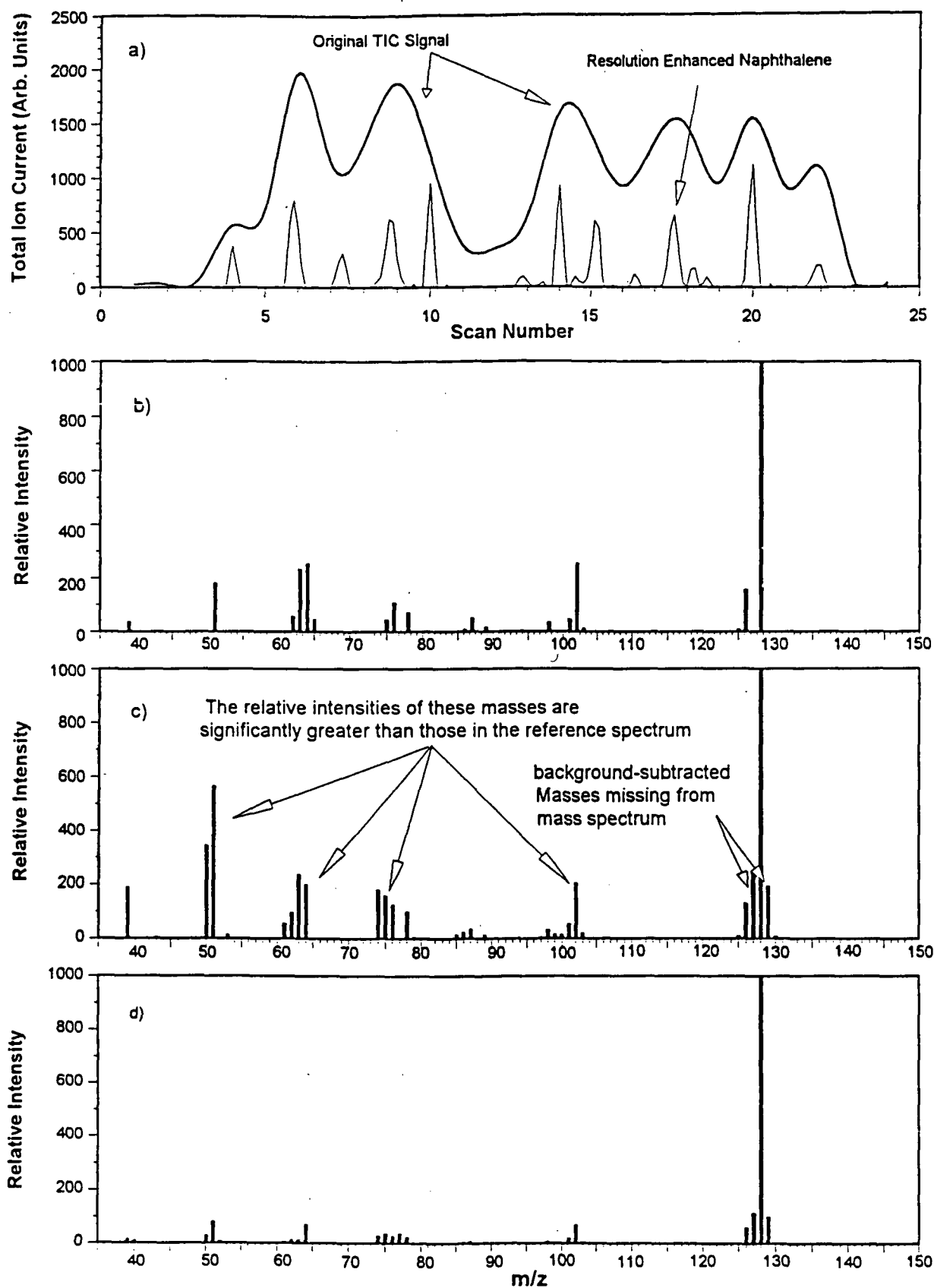| Compound	Scan # (Fig. 8.1.c)	Native Qual. Fact.	Enhanced Qual. Fact.	Native Cross Corr.	Resolution Enhanced Cross Corr.
2,5-Dimethyl-benzo[b]thiophene	12	89	91	9606	9927
Tetradecane	16	89	93	8462	9621
1-Ethylnaphthalene	19	96	97	9904	9954

Figure 8.3. a) Native and enhanced total ion chromatograms of a naphthalene-containing mixture; b) native mass spectrum of naphthalene in mixture retrieved by data system; c) enhanced mass spectrum of naphthalene in mixture; d) NIST reference spectrum from those chromatograms.

24

A modification of the spectral enhancement procedure by calculating the quadratically predicted intensity at each peak was tested. The difference between any given enhanced peak relative intensity and the relative intensity derived from raw data in the native mass chromatogram was less than 2%. Therefore, this modification to the approach was not studied further.

The stability of the quadratic fitting procedure was tested on 13 peaks with apex offset values (O_A) in the range of ±0.45 scans. Mass spectrometer noise was simulated by randomly adding noise values in the range ±5, 7.5, or 10 % of the experimental value to each of the three experimental values used to calculate the quadratic approximation. These treatments simulated total noise vs. signal levels of 10, 15, or 20 percent peak-to-peak (p-p), respectively. The standard deviation was plotted against the O_A value (Figure 8.4).

The data showed that for scan-to-scan noise levels up to 20 percent, p-p, the quadratically predicted O_A generally varied less than ±0.2 scans at the 95 percent confidence level (see Figure 8.4). This situation indicated that the underlying mass spectral signal maximized within 0.2 scan of the calculated value if the scan-to-scan noise was less than 20 percent, p-p. Although scan-to-scan noise information was not readily available for commonly used environmental mass spectrometers, it appeared to be unlikely that scan-to-scan noise levels would approach 20 percent in properly operating units. This value of 0.2 scans for an approximate O_A uncertainty was about the same as the optimum summing interval. Therefore, a quadratically optimized peak would fall within one summing interval of its correct position, an acceptable situation because mass spectral peaks were typically 4 to 6 summing intervals wide.

A simple comparison of the results obtained by performing the O_A optimization and by omitting this step is shown in Figure 8.5. The results obtained by correcting for the scan offset, O_S, but not for O_A are shown in Figure 8.5.b. These data were characterized by a series of peaks of unit scan width, as should be observed. The procedure simply identified those ions maximizing in that mass spectrometric scan. The appearance of doublets in most of the peaks was an artifact of the summing interval, 0.1 scans, and the distribution of ion intensities in common environmental mass spectra. The same enhanced mass chromatogram, but incorporating the quadratic optimization term, O_A, is shown in Figure 8.5.a. The primary effect of including inter-scan effects was a sharp reduction in the number of observable "peaks." This was particularly apparent in the region of scan numbers 1935 and 1939. The only major observable addition to the chromatogram was the emergence of a potential very narrow peak at scan number 1928.7. However, the appearance of two peaks in this region resulted from the short 0.1 scan summing intervals used; the two "peaks" merged into one when a 0.2 scan summing interval was used. This occurrence emphasized the importance of selecting the proper sized summing interval. Examination of mass peak sequences showed that neutral loss sequences were interspersed across both peaks so that they should be considered one peak, reinforcing the previous conclusion about selecting the proper summing interval length. Based on preceding results, the primary utility of the simple O_S-only retention time optimization procedure appeared to be in studying whether there were any systematic, mass-dependent principles such as low mass discrimination involved in the inter-scan corrections of the quadratic optimization procedure.

The capabilities of the spectral enhancement concept were ultimately limited by the characteristics of the EXCEL™ 5.0 spreadsheet. The spreadsheet permitted addressing any section of a mass chromatogram containing up to 16,000 records. The mass chromatogram was imported, the optimized mass retention times determined, and the resulting synthetic mass chromatogram was obtained in approximately 5 to 10 minutes. The resulting summary data such as total ion current mass chromatograms and mass spectra were called synthetic because they were created from selected subsets of the total data for the run of interest rather than
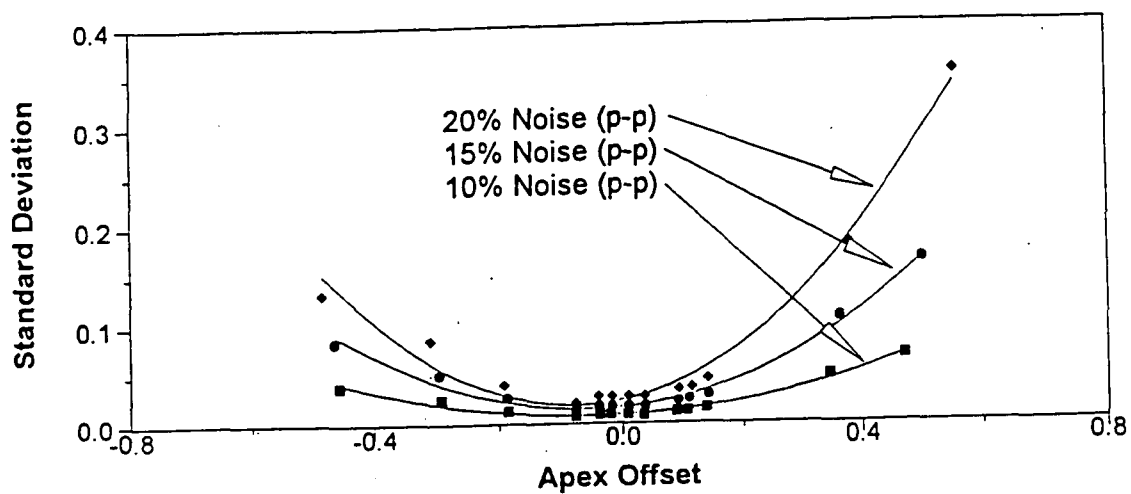
Figure 8.4. Plot of the standard deviations of simulated O_A under varying noise conditions versus the estimated O_A value.
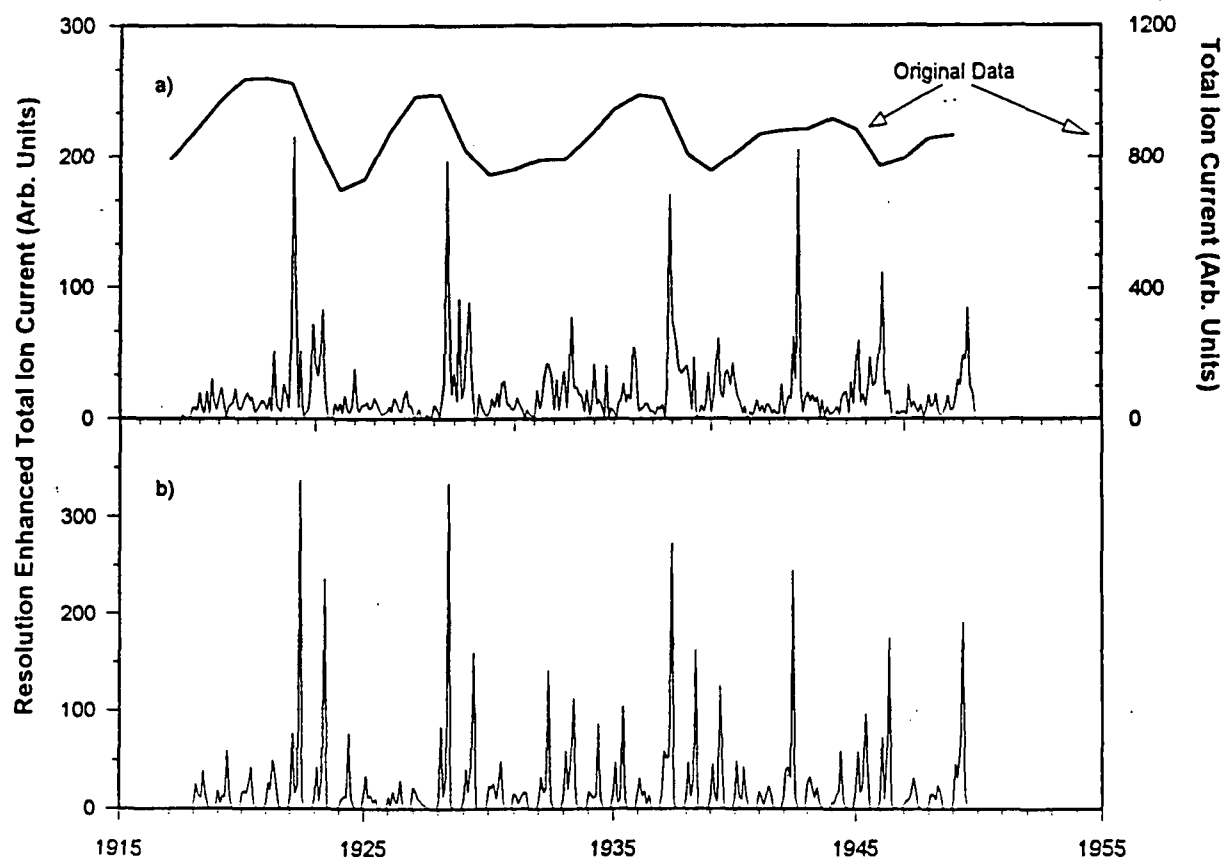

Figure 8.5. Comparison of a) spectral enhancement results using the quadratic fit, and b) spectral enhancement results without calculating the quadratic fit.

the complete native data set. The synthetic mass spectrum for each resulting peak of interest was readily obtained in hard copy or electronic format that could be analyzed by a mass spectral data system.

9. QUALITY ASSURANCE PROCEDURES

To conduct quality assurance reviews of TIC data, at a minimum the following information for TIC spectral matches should be included in the data package:

1) labeling of the mass axis or major fragments
2) the full name of the spectral match (if possible)
3) the CAS number associated with each spectral match
4) the molecular weight of the match
5) the composition of the spectral match
6) the ranking (score) of the spectral match and
7) a table of ion intensities vs. m/z values.

We tested conversions through MassTransit™ software between the HP formats and INCOS™, finding that all ion intensities and retention times remained the same. Conversions from and back to, HP Chemstation™ showed no changes in the data. Conversion from one HP format to the other, and then workup in the new format gave results identical to workup in the original format. Thus, it was concluded that MassTransit™ did not modify the data, and that the different data handling software packages available from HP treated the data in equally authentic ways.

Quality assurance procedures for analytical determinations included acquiring accurate mass measurements in triplicate, careful calibration of the mass range with PFK, and obtaining SIR-based accurate mass measurements.[11,12] Additionally, peak profiles were not considered valid unless points on each side of the maximum were observed in addition to the maximum. Some TICs were not amenable to CLP-type GC/MS analysis due to their low concentrations or to the absence of their mass spectra in the library data base. Other techniques such as HRMS, LC/MS, and GC/FT-IR provided valuable data in some cases.[12] Instances were found where the CLP MS data system rounded off masses incorrectly, where the low-intensity isotopic members (including the one of lowest mass) of halogen ion groups were not seen, and where the scan range of 35 to 500 Da used by CLP prevented the detection of the molecular ion.

10. CONCLUSIONS

In this project, 99,513 TICs were reported in 792 Sample Delivery Groups (SDGs) studied (Table 2.1). Of these, the CLP reported identifications with Chemical Abstracts Service (CAS) numbers on 16%. Not all of these identifications were correct. It was estimated from this study that perhaps 30% of the 16% portion were correct, and possibly another 10% were correct except that the TIC was an isomer of the compound whose CAS number was reported. Forty-one percent of the TICs were listed as being partially identified, and the remaining 43 percent were reported as unknown. Target Compound List (TCL) analytes were 5 to 10% of the total number of analytes in the Superfund sample data studied, with the remainder being TICs. If only about 16% of the TICs are reported with CAS number identification, and only about 30% to 40% of those identifications are correct, then perhaps only 5% to 6% of the TICs are correctly identified. The result is that approximately 84-90% of the analytes remain unidentified under the current CLP requirements.

An overview of the data indicated that the most commonly reported classes of TICs were saturated (branched and straight chain alkanes) and unsaturated (alkenes, dienes, etc.) hydrocarbons. The next most prominent groups of compounds were PAHs and aromatic compounds that were frequently substituted with aliphatic hydrocarbons. Higher molecular weight steroid compounds (i.e., cholesterol) and PCBs were also reported. Elemental sulfur was reported by the CLP laboratories in ca. 50% of soil samples. The TCL phthalates were found in ca. 80% of the soil samples, and non-TCL phthalates were found in ca. 10%.

TIC mass spectra were found, including some with recognizable chlorine and/or bromine clusters, for which there are no library spectra in the computerized mass spectral data bases. Such spectra must presently be manually interpreted. Instances were found where the CLP MS data system rounded off masses incorrectly, where the low-intensity isotopic members (including the one of lowest mass) of the halogen ion groups were not seen, and where the scan range of 35 to 500 Da used by CLP prevented the detection of the molecular ion.

Additional compounds could be included in the NIST and Wiley mass spectral data bases to assist in identifying TICs for environmental monitoring efforts. These compounds include higher molecular weight PAHs, and industrial process solvents and chemicals. Additional pesticide metabolites and degradation products should be included; some are available in hardcopy form but not in commonly used software data bases.

The reporting trends from the laboratories were highly varied with respect to accuracy of reporting TICs. While some (ca. 2%) laboratories made honest efforts to identify the TICs, others (ca. 30%) simply labeled all TIC peaks as unknown and made no attempt to accurately identify the TICs. Some laboratories (ca. 2%) would only commit to identifying sulfur and reported all other TICs as unknown. Some laboratories (ca. 20%) used the name of the library match compound with the highest score on the library search, treating that as the identification regardless of whether that identification was reasonable based upon manual spectral interpretation.

Data system algorithms and procedures studied in this project included background subtraction, Biller-Biemann spectral isolation and library searching, and McLafferty Probability Based Matching (PBM). These software-based capabilities are implemented in the mass spectrometer vendor's data systems. Procedures for mass spectral resolution enhancement were also developed and tested, using the concept proposed by Colby and in-house developed spreadsheet-based macros and other procedures. A comparison of the two library searching algorithms was performed on 20 CLP data files. Library searches were conducted with an HP DOS data system using PBM and with a Finnigan system using the Biller-Biemann algorithm. The results demonstrated that the two algorithms performed to the same level of quality and reliability for compounds in the molecular weight ranges associated with semivolatile TICs.

Commercial MassTransit™ software was tested for its ability to facilitate converting data files into formats usable by other data systems, including spreadsheets on personal computers. Interfacing the different MS data system formats through this software effectively accomplished a standardization of TIC data. This software converted GC/MS data from various contractors into ASCII text.

These data were subsequently imported into an EXCEL™ spreadsheet and manipulated using a macro written in-house to perform mass spectral resolution enhancement. The spreadsheet macro performed the sorting and statistical and mathematical procedures necessary to separate the TICs from interfering compounds. The extracted mass spectra were compared to reference spectra contained in the NIST database. Results of these comparisons showed that usually 80 to 90 percent of the ions contained in the reference

spectra were successfully extracted using this method. This procedure improved mass spectral quality and the data system's ability to perform successful library searches. The fit quality parameters showed systematic improvements after subjecting the data to resolution enhancement procedures. This approach was found to be effective in extracting mass spectra of individual compounds from background signals, including those of hydrocarbon mixtures having broad elution profiles on the chromatographic column utilized for the GC/MS acquisitions. The approach could have significant value for EPA and CLP as a rapid, cost effective alternative to special extract clean-up and reanalysis schemes for removal of chemical interferences that render the mass spectra of selected peaks/unknowns difficult or impossible to interpret.

Many analytes of potential interest are not amenable to GC/MS analysis. Other techniques such as LC/MS can be used to detect these compounds. LC/MS conditions can be selected for analytes that are thermally unstable, have low-volatility or high molecular weight, or are highly water-soluble. HRMS can be used to determine accurate masses and elemental compositions.[11,12] Some of the above suggestions may be difficult or costly to implement. An alternative would be for the CLP laboratory to "flag" samples for the Region or other requestor to consider sending to an Agency "expert" or research laboratory for further study of the TICs.

Seven possible improvements for TIC reporting and identification by CLP laboratories were identified:

(1)   require reporting the first library match of TICs on Form 1 if the library match meets a specified probability level, rather than "unknown" regardless of search results;

(2)   require reporting of TICs when the mass spectra, or library matches, indicate the presence of heteroatoms such as N, P, S, halogen, or heavy metals;

(3)   incorporate RRT data and frequency of occurrence of TICs into a data base, and add RRT criteria to aid in TIC identification;

(4)   extend the mass range to 600 Da, with a tune emphasizing greater sensitivity for masses above 300 Da, for better detection and identification of TICs in this higher molecular weight range;

(5)   recommend or require that different GC temperature programs or column phases be used in a second analysis to improve separation of some TICs;

(6)   because sulfur was found in approximately 50% of the soil sample data, improvements to the GPC procedure, a cleanup with copper, or some other improved procedure would be worthwhile;

(7)   identify additional compounds suitable for addition to the mass spectral data bases, because they have been found, or would be anticipated in known types of waste sites.

The first recommendation is currently being implemented for CLP TICs. The remaining six ((2)-(7)) recommendations are offered as tools for continued strengthening of the CLP, and may, as recommended by one of our external reviewers, best be implemented by using a centralized laboratory. This could be a federal laboratory performing research in compound identification and capable of supplementing the above recommendations with special MS techniques (e.g., CI-MS, HRMS,and MS/MS) and other non-MS techniques with adequate sensitivity (e.g,. FT-IR).

In summary, this project investigated the effectiveness of current TIC reporting under the CLP protocols. Identification of TICs was found to be of variable quality across participating laboratories. The commonly used mass spectral data systems were found to provide essentially the same results. The libraries (Wiley and NIST) could be improved by adding additional compounds relevant to environmental monitoring. The available algorithms and data system procedures are satisfactory and provide virtually identical results. Mass spectral enhancement procedures could materially help in identifying TICs by separating TIC spectra of interest from those of aliphatic hydrocarbons or other background signals.

11. REFERENCES

1. U.S. Environmental Protection Agency, Contract Laboratory Program, Statement of Work for Organic Analysis, Multi-media, Multi-concentration. Document Number OLM01.0, 1990. Including Revisions, OLM01.1--OLM03.1, Dec 1990--Aug 1994. U.S. Environmental Protection Agency, Cincinnati, OH.

2. J. M. Long and J. M. McGuire, "Assessment of Tentatively Identified Compounds in Superfund Samples," U. S. Environmental Protection Agency, Environmental Research Brief EPA/600/M-89/030; June, 1990.

3. Viar and Company, "Evaluation of the Toxicity of 798 Commonly Occurring Semivolatile Tentatively Identified Compounds," Interim Report and Final Report on SMO Performance Event 259, October 31, 1991 and December 31, 1991.

4. Donald R. Scott, "Rapid and Accurate Method for Estimating Molecular Weights of Organic Compounds from Low Resolution Mass Spectra," *Chemometrics and Intelligent Laboratory Systems*, 16, 193-202 (1992).

5. Donald R. Scott, A. Levitsky, and S. E. Stein, "Large Scale Evaluation of a Pattern Recognition/Expert System for Mass Spectral Molecular Weight Estimation," *Analytica Chimica Acta*, 278, 137-147 (1993).

6. Donald R. Scott, "Empirical Pattern Recognition/Expert System for Molecular Weight Estimation of Low Resolution Mass Spectra," *Analytica Chimica Acta*, 285, 209-222 (1994).

7. Bruce N. Colby, "Spectral Deconvolution for Overlapping GC/MS Components," *Journal of the American Society for Mass Spectrometry*, 3, 558-562 (1992).

8. Colby, B.N.; D'Arcy, P.H. Reliable Compound Identification at Low Levels in Complex Environmental Samples. *Proceedings of the 41st ASMS Conference of Mass Spectrometry and Allied Topics;* San Francisco, CA, 1993; p. 813.

9. Stephen E. Stein, "Estimating Probabilities of Correct Identification from Results of Mass Spectral Library Searches," *Journal of the American Society for Mass Spectrometry*, 5, 316-323 (1994).

10. Stephen E. Stein, "Optimization and Testing of Mass Spectral Library Search Algorithms for Compound Identification," *Journal of the American Society for Mass Spectrometry*, 5, 859-866 (1994).

11. Andrew H. Grange, Joseph R. Donnelly, William C. Brumley, Stephen Billets, and G. Wayne Sovocool, "Mass Measurements by an Accurate and Sensitive Selected-Ion-Recording Technique," *Analytical Chemistry*, 66, 4416-4421 (1994).

12. A.H. Grange, J.R. Donnelly, W.C. Brumley, and G.W. Sovocool, "Determination of an Elemental Composition from Mass Peak Profiles of the Molecular Ion (M) and the M+1 and M+2 Ions," *Analytical Chemistry*, 68, 553 (1996).

13.   N.R. Herron, J.R. Donnelly, and G.W. Sovocool. "Software-based Mass Spectral Enhancement to Remove Interferences from Spectra of Unknowns," *Journal of the American Society for Mass Spectrometry*, 7, 598 (1996).