



ENVIRONMENTAL RESEARCH BRIEF

Assessment of Tentatively Identified Compounds in Superfund Samples

J. M. Long and J. M. McGuire*

Abstract

Stored mass spectral data for 27 semivolatile samples analyzed by 7 private laboratories under contract with the U.S. Environmental Protection Agency were reanalyzed at the Environmental Research Laboratory, Athens, GA (AERL). Results of the reanalysis were compared with the original contract laboratory results. In instances where specific compound identifications had been made by a contract laboratory, AERL identifications agreed 36% of the time, disagreed with the identification 11% of the time, disagreed on the presence of the GC peak 19% of the time, or concluded data were insufficient for identification 34% of the time.

Background

Public Law 96-510, entitled "The Comprehensive Environmental Response, Compensation and Liability Act of 1980" (commonly known as Superfund), authorized, among other things, testing and monitoring of waste sites. To accomplish this, the Superfund Contract Laboratory Program (CLP) was established and comprehensive methods were implemented for contract laboratories to use in analyzing and reporting target analytes (1, 2). In addition to the target analytes, the statement of work required each contractor, for each semivolatile sample analyzed by GC/MS, to conduct mass spectral library searches to determine the possible identity of as many as 20

semivolatile components not listed on EPA's Target Compound List (3). The selections were to be those compounds having the greatest concentrations. Substances that exhibited responses less than 10% of the nearest internal standard were not to be considered. A further requirement was that the 1985 or most recent release of the National Bureau of Standards spectral library was to be used to conduct the searches. Reporting requirements stipulated that compounds not meeting the complete identification requirements contained in the statement of work should be reported as "unknown," or "unknown hydrocarbons," or "unknown aromatic," etc. In other words, the tentative identification should be as specific as possible even for compounds identified as "unknown."

At the time the tentatively identified compounds (TIC) concept was included in the statement of work, it was thought that such tentative identifications might provide information that would be useful in modifying the target list. Since then, thousands of TIC identifications have been made and a need has appeared for assessing their reliability. This work is the first part of such an assessment. Multi-spectral identification or confirmation of selected TICs is in progress at this time and will be reported separately as the second step in the assessment process.

Approach

The TICs made by the contract laboratories were compared to those made by the AERL, using the same mass spectral data but with different analytical protocols,

*Environmental Research Laboratory, Athens, GA 30613



in order to determine the reliability of the contract laboratory identifications. The data for this research brief were generated from 7 contract laboratories on a total of 27 extracts and were processed by computer programs developed by AERL personnel (4) and others (5-11). The best mass spectrum for each GC peak was located and extracted from its background by the programs. The concentration was then estimated, and as many as ten possible identifications using Probability-Based Matching (PBM) were compiled for each resolved peak in the mass chromatogram. The library, containing 110,000 spectra, used in this work is a sub-set of the complete Wiley library and is larger and more extensive than the NBS library used by the contract laboratories.

The complete AERL computer identification programs rely heavily on historical relative retention data, which were not available for this study. Accordingly, AERL developed the following rules to facilitate selection of the best match.

1. The value of the PBM derived ratio $k/(k + \Delta k)$ should be greater than or equal to 0.50, and should be significantly higher than that for the next best hit.
2. The value of the PBM k should be greater than or equal to 40.
3. Priority should be given to k 's with "+" signs, which indicate the presence of an ion in the unknown spectrum at a mass corresponding to the molecular weight of the library match.
4. Priority should be given to matches that have relative retention time agreement (where available).
5. If two adjacent scans have the same value of $k/(k + \Delta k)$ and if these are the best HITS, the one with the higher concentration should be chosen.
6. A match should be chosen if the value of $k/(k + \Delta k)$ is greater than or equal to 0.95 and the contamination value is less than 2.
7. No match having 3 PBM flags or one with "anhydride" as part of the chemical name should be chosen.
8. The same identification should not be chosen more than once in the same run.

Using this screening procedure, a technician was able to, in many instances, eliminate all but two or three of the ten most probable identifications. From the two or three that passed the screen, the most probable identification then was chosen by the analyst based on his reexamination of the data.

Results

Table 1 represents an example of part of the information summarized in the report of TICs for an individual semivolatle extract by contract laboratories. Table 2, which closely resembles Table 1, contains the AERL identifications corresponding to those in Table 1. The data in Table 3 summarize the comparison of TIC reports, such

Table 1. Representative Identifications in TIC Report

CAS Number	Compound Name	Rt (minutes)	Est. Conc.	Q*
1. 140-76-1	Pyridine, 5-Ethenyl-2-Methyl	9.27	26000	J**
2.	Unknown Hydrocarbon	14.39	12000	J
3.	Unknown Hydrocarbon	15.97	44000	J
4.	Unknown	16.37	11000	J
5.	Unknown Hydrocarbon	16.90	25000	J
6.	Unknown	18.35	17000	J
7.	Unknown Hydrocarbon	18.85	51000	J
8.	Unknown	19.07	200000	J
9.	Unknown	19.74	46000	J
10.	Unknown Hydrocarbon	20.17	38000	J
11.	Unknown Hydrocarbon	20.25	37000	J
12.	Unknown	21.44	40000	J
13.	Unknown Hydrocarbon	21.54	34000	J
14.	Unknown	22.44	10000	J
15.	Unknown Hydrocarbon	22.62	24000	J
16.	Unknown Hydrocarbon	23.77	33000	J
17.	Unknown	25.04	25000	J
18.	Unknown	27.54	20000	J

*EPA Qualifier from statement of work

**J = estimated concentration or concentration is less than the quantitation limit

as the one in Table 1, for each sample, and the computer outputs resulting from AERL processing of the contract laboratory mass spectral data. For each TIC compound name entered in Table 3, the purity or probability of the spectral match is recorded, as are the identifications made by AERL personnel and the corresponding value of the PBM $k/(k + \Delta k)$.

Specifically, Table 3 summarizes the data obtained on the 27 semivolatle extracts by the 7 contract laboratories along with comparative data obtained by AERL by processing the mass spectral data generated by those contract laboratories. The information includes the number of TICs made, the number of those that are "specific," "generic," and "unknown" and the range in purity for compounds in the three groups, and a descriptor for the overall shape of the mass chromatogram. Purities are listed for laboratories using Finnigan instruments; probabilities are listed for those using Hewlett Packard instruments. Specific identifications for the purposes of this report are defined to be those employing specific compound names, e.g., n-hexadecane. Generic identifications are those employing chemical family names, e.g., unknown hydrocarbon. The remaining identifications are defined as unknown identifications. These employ only the descriptor, "unknown." For specific identifications, the comparative data obtained by AERL include a breakdown by four categories--agreement (A), disagreement (D), no-scan (NS), and misidentified (MIS). For generic and unknown identifications, the same categories, with the

Table 2. AERL Identifications Corresponding to TIC Report (Table 1)

Compound Name	Rt (minutes)	Concentration $\mu\text{g/kg}$
1. pyridine, 5-ethyl-2-methyl	9.27	8500
2. unknown hydrocarbon	14.39	11000
3. unknown hydrocarbon	15.97	13000
4. MIS	16.37	1000
5. Unknown hydrocarbon	16.90	4000
6. MIS	18.35	18000
7. Unknown hydrocarbon	18.85	14000
8. MIS	19.07	Saturated
9. MIS	19.74	20000
10. unknown hydrocarbon	20.17	7000
11. unknown hydrocarbon	20.25	8000
12. unknown hydrocarbon	21.44	18000
13. unknown hydrocarbon	21.54	9000
14. MIS	22.44	2500
15. unknown hydrocarbon	22.62	15000
16. unknown hydrocarbon	23.77	13000
17. MIS	25.04	9900
18. MIS	27.54	5900

exception of MIS, are used. The agreement and disagreement categories need no explanation. The NS category indicates that there was no scan in the AERL-processed data corresponding to the contract laboratory scan. These scans were absent due either to a known deficiency in AERL's peak recognition program or to the contract laboratory's reporting peaks that were not real. The MIS category refers to spectra that were not interpretable by AERL personnel based on GC/MS alone. The table also includes a range of values obtained for $k/(k + \Delta k)$ for specific, generic, and unknown identifications.

For "specific" identifications in Table 3, the overall range of purity/probability values was: for contract laboratory A, 576-977; for B, 371-964; for C, 504-829; for D, 625-977; for E, 677-873; for F, 67-95; and for G, 52-93. The AERL $k/(k + \Delta k)$ range of values for identifications corresponding to and in agreement with those for contract laboratories was: A, 0.40-0.94; B, 0.67-1.00; C, 0.50-0.95; D, no data; E, 0.80-0.80; F, 0.66-1.00; and G, 0.55-0.91. Regression analysis indicated there is no linear correlation between either purity or probability values and $k/(k + \Delta k)$ values.

For "generic" identifications, the range of purity/probability values for contract laboratory A was 736-845; for B, 217-923; for C, 138-803; for D, no data; for E, 558-909; for F, 11-89; and for G, 15-81. The AERL $k/(k + \Delta k)$ range for identifications corresponding to and in agreement with those for contract laboratory A was 0.23-1.00; for B, 0.33-1.00; for C, 0.15-0.75; for D, no data; for E, 0.42-1.00; for F, 0.17-1.00; and for G, 0.50-0.91.

For "unknown" identifications, the range of purity/probability values for contract laboratory A was 290-800; for B, 197-768; for C, no data; for D, 683- 711; for E, 216-883; for F, 20-70; and for G, 11-38. The AERL $k/(k + \Delta k)$ range for identifications corresponding to and in agreement with those for contract laboratory A was 0.16-1.00; for B, 0.18-0.88; for C, no data; for D, 0.21- 0.72; for E, 0.29-0.82; for F, 0.19-0.86; and for G, 0.31-0.72.

There are two $k/(k + \Delta k)$ ranges for each identification category in each sample reported. The first range was obtained from those identifications for which there is agreement between the contract laboratory and AERL. The second range was obtained from those identifications for which there is disagreement.

It is evident from Table 3 that the values for AERL's $k/(k + \Delta k)$ are, for the most part, greater than or equal to 0.50 for the specific identifications. In a few instances, the compound identification having a value less than 0.50 was determined by the analyst to be reasonable and therefore was selected as the best HIT.

For the generic category, the range of $k/(k + \Delta k)$ values for each mass spectrum was obtained by selecting the lower and upper values from all the identifications comprising the generic group. The same is true for the unknown category.

The lower purity ranges for both generic and unknown identifications tend to be lower than those for the specific identifications. This is not unusual since only a poor correlation is expected between purity values and either generic or unknown identifications.

Table 4 summarizes the data contained in Table 3 and shows overall agreement and disagreement between the contract laboratory TICs and those determined by AERL. It is interesting to note that of the 478 contract laboratory TICs involved, 38% were specific identifications, 39% were generic identifications, and 23% were unknown identifications. AERL was in agreement with 36% of the specific identifications and in disagreement with 11% of them. In many instances, a disagreement on a specific identification would be considered an agreement on a generic basis. The designation of a GC peak as n-hexadecane by one of the contract laboratories and as "unknown hydrocarbon" by AERL is an example of this situation. This table indicates that AERL is in agreement by roughly the same percentage with four of the seven contract laboratories on specific identification and with five of seven contract laboratories on both generic and unknown identifications. The NS category comprised 19% and the MIS category 34% of the specific identifications, respectively. For the generic identifications, AERL was in agreement with 48% and in disagreement with 22%. For unknown identifications, AERL was in agreement 54% and in disagreement with 10%. The NS category comprised 15% of the generic and 22% of the unknown identifications.

Table 5 contains concentration estimates for each sample reported in Table 3. It appears that, for all three identifica-

Table 3. (Continued)

Contract Lab. (CL)	Sample #	EPA AERL Run #	TICs (CL)						TICs (AERL)													
			Specific Idents.		Generic Idents.		Purity Range		Specific Idents.		Generic Idents.		Purity Range									
			A	D	NS	MIS	A	D	NS	k+Δk	A	D	NS	k	k+Δk							
F	1	ER843	5	67-91	7	37-89	7	20-70	3	0	0	2	0.72-0.88	4	2	1	0.17-0.92	7	0	0	0.26-0.86	High
		44094															0.52-0.88					
F	1	ER844	5	70-95	7	11-81	7	25-59	3	0	1	1	0.85-1.00	5	1	1	0.18-0.96	5	0	1	0.25-0.71	Med
		44095															0.26-0.44					
G	1	ES062	13	52-92	1	15	--	--	1	5	7	0	0.55	0	0	1.	--	0	0	11	--	Low
		44071											0.40-0.81									
G	1	ES061	16	52-93	3	15-60	3	11-38	5	6	1	4	0.58-0.91	0	2	1	--	3	0	0	0.31-0.72	Med
		44070											0.33-0.96				0.27-0.47					
G	1	ES054	13	60-89	9	36-81	0	--	2	11	0	0	0.50-0.63	0	8	1	--	0	0	0	--	Med
		44068											0.29-1.00				0.34-0.81					

*5 replicate identifications excluded
 **3 replicate identifications excluded

Legend

- A Number of agreements between contract laboratory and AERL on identifications
- D Number of disagreements between contract laboratory and AERL on identifications
- NS No scan in AERL processed data corresponding to that for contract laboratory
- MIS Spectrum of compound not interpretable by AERL
- Purity Term used by Finnigan to indicate goodness of match between unknown and library spectra
- Prob. (probability) Term used by Hewlett Packard to indicate goodness of match between unknown and library spectra
- k/(k + Δ k) Term used by AERL to indicate goodness of match between unknown and library spectra
- Baseline Drift Term used by AERL to indicate overall shape of mass chromatogram

Table 4. Condensed TIC Statistics on Agreement/Disagreement of Contract Laboratory/AERL

Contract Lab. (CL)	TICs (CL)				TICs (AERL)									
	No.				Specific Identifications				Generic Identifications			Unknown Identifications		
		% Specific	% Generic	% Unknown	% A	% D	% NS	% MIS	% A	% D	% NS	% A	% D	% NS
A	136	35	7	58	44	8	27	21	100	0	0	65	20	15
B	111	11	69	20	50	17	0	33	61	21	18	77	14	9
C	55	25	75	0	57	0	0	43	56	34	10	0	0	0
D	20	90	0	10	0	0	33	67	0	0	0	50	0	50
E	30	17	63	20	20	0	40	40	53	0	47	67	33	0
F	58	26	43	31	60	0	13	27	68	24	8	94	0	6
G	68	62	19	19	19	52	19	10	0	77	23	23	0	77
Total	478													
Mean		38	39	23	36	11	19	34	48	22	15	54	10	22
Std. Dev.		28	31	18	23	19	16	18	36	28	16	32	13	30

Legend

- % Specific Percentage of contract laboratory identifications that employ specific chemical names
- % Generic Percentage of contract laboratory identification that employ chemical family names
- % Unknown Percentage of contract laboratory identifications that employ only the descriptor "unknown"
- % A Percentage of contract laboratory identifications agreed upon by AERL
- % D Percentage of contract laboratory identifications disagreed with by AERL
- % NS Percentage of contract laboratory identifications for which there were no corresponding scans in AERL data
- % MIS Percentage of contract laboratory identifications not interpretable by AERL

tion categories, the agreement between contract laboratories and AERL is within a factor of three.

Conclusions

Overall, the agreement between AERL and the contract laboratory identifications for both specific and generic identifications appears to be fair and also roughly equivalent for five of the seven contract laboratories. Generic and unknown identifications comprising 62% of the total is indicative, to some extent, that perhaps fewer samples should have been analyzed in order to obtain more thorough interpretations of the data generated. Future work statements should be written in a manner to strongly discourage the use of "unknown" identifications. Such identifications should be used only as a last resort. It was observed in at least one instance that the same specific compound identification appeared more than once in a single TIC report. This suggests that this particular report did not receive a great deal of review. Finally, in several instances, it appeared that relative retention time data were ignored in assigning compound identities.

Acknowledgments

Paul Kimsey's help in applying the AERL rules to screen the computer outputs is gratefully acknowledged, as is the advice of Dr. Susan Richardson and Al Thruston, Jr. concerning compound identifications.

References

1. Fisk, J.F., A.M. Haebeler, and S.P. Kovell, *Spectra*, Volume 10, Number 4, 22 (1986).
2. Friedman, D., *ibid*, 40.
3. USEPA Contract Laboratory Program, Statement of Work for Organic Analysis, Multi-Media Multi-Concentration, 10/86. Rev: 1/87, 2/87, 7/87, 8/87.
4. Shackelford, W.M., D.M. Cline, L. Burchfield, L. Faas, G. Kurth, and A.D. Sauter, "Computer Survey of Gas Chromatography/Mass Spectrometry Data Acquired in the U.S. Environmental Protection Agency Screening Analysis: System and Results," pp. 527-554 in "Advances in the Identification and Analysis of Organic Pollutants in Water," ed. L.H. Keith, Ann Arbor Science, Ann Arbor, MI (1981).
5. Smith, D.H., M. Achenback, W.J. Yeager, P.J. Anderson, L. Fitch, and T.C. Rindfleisch, *Anal. Chem.*, 49, 1623 (1977).
6. Dromey, R.G., J. Stefik, T.C. Rindfleisch, and A.M. Dufield, *Anal. Chem.*, 48, 1362 (1976).
7. Pesyna, G.M., R. Venkataraghavan, H.R. Dayringer, and F.W. McLafferty, *Anal. Chem.*, 48, 1362 (1976).
8. Atwater, B.L.(F.), D.B. Stauffer, F.W. McLafferty, and D.W. Peterson, *Anal. Chem.*, 57, 899 (1985).
9. Stauffer, D.B., F.W. McLafferty, R.D. Ellis, and D.W. Peterson, *Anal. Chem.*, 57, 1056 (1985).
10. McLafferty, F.W. and D.B. Stauffer, *J. Chem. Inf. Comp. Sci.* 25, 245 (1985).
11. Shackelford, W.M., D.M. Cline, L. Faas, and G. Kurth, *Anal. Chem. Acta.*, 146, 15 (1983).

Table 5. Comparison of TIC Concentrations

Contract Lab. (CL)	EPA Sample # AERL Run #	Conc. (CL)			Conc. (AERL)		
		Specific Idents.	Generic Idents.	Unknown Idents.	Specific Idents.	Generic Idents.	Unknown Idents.
A	FG489 44121	2700	--	2800	2300	2000	2200 (µg/kg)
A	FG496 44140	900	--	9500	700	280	6100 (µg/L)
A	FG490 44128	230000	--	340000	7300	--	4400 (µg/kg)
A	FG488 44127	26000	33000	46000	8500	11500	6400 (µg/kg)
A	FG494 44138	230	--	500	200	225	290 (µg/L)
A	FG495 44139	900	--	6400	1200	--	6000 (µg/L)
A	FG493 44137	650	--	2700	400	--	2100 (µg/L)
A	FF397 44136	120	--	200	400	--	300 (µg/L)
B	AK077 44082	2400	3600	1100	1300	3000	300 (µg/L)
B	DH939 44080	39000	17000	16000	7700	6000	1600 (µg/kg)
B	ER728 44081	100	50	80	80	60	90 (µg/L)
B	YD028 44076	--	108000	74000	--	7000	500 (µg/kg)
B	YD037 44078	57000	4900	--	3200	16000	-- (µg/kg)
B	YD035 44073	2500	2200	1700	1000	600	25 (µg/kg)
C	DH441 44165	7600	11000	--	3100	1400	-- (µg/kg)
C	DH438 44170	1000	2200	--	700	--	-- (µg/kg)
C	DH444 44166	13000	3800	--	2100	400	-- (µg/kg)
C	DH449 44171	500	900	--	200	100	-- (µg/kg)
D	CQ538 44072	900	--	500	--	160	230 (µg/kg)
E	CR385 44075	12000	18000	12000	13000	22000	4000 (µg/kg)
F	GE325 44092	470	--	89	200	--	200 (µg/L)
F	ER837 44093	59000	25000	12000	32000	30000	14000 (µg/kg)
F	ER843 44094	6500	2100	2100	3200	2800	1600 (µg/kg)
F	ER844 44095	46000	30000	33000	16000	35000	20000 (µg/kg)
G	ES062 44071	900	7500	250	1100	--	-- (µg/L)
G	ES061 44070	1700	3900	900	600	1000	200 (µg/kg)
G	ES054 44068	700	1300	--	800	200	-- (µg/kg)

Note: Mention of trade names or commercial products does not constitute endorsement or recommendation for use by the U.S. Environmental Protection Agency.

United States
Environmental Protection
Agency

Center for Environmental Research
Information
Cincinnati OH 45268

Official Business
Penalty for Private Use \$300

EPA/600/M-89/030

•

•