Analytica Chimica Acta, 210 (1988) 51–62
Elsevier Science Publishers B.V., Amsterdam — Printed in The Netherlands

EXPERT SYSTEM FOR INTERPRETATION OF THE INFRARED SPECTRA OF ENVIRONMENTAL MIXTURES

LI-SHI YING and STEVEN P. LEVINE*
School of Public Health, The University of Michigan, Ann Arbor, MI 48109 (U.S.A.)
STERLING A. TOMELLINI
Department of Chemistry, University of New Hampshire, Durham, NH 03824 (U.S.A.)
STEPHEN R. LOWRY
Nicolet Instrument Corporation, 5225 Verona Road, Madison, WI 53711 (U.S.A.)
(Received 1st September 1987)

SUMMARY

A program for the identification of the principal components of mixtures through interpretation of the infrared mixture spectrum (IntIRpret) was developed. This program, which was developed as a preliminary screening tool for unknown organic mixtures, has five main subroutines: the interferogram processing and peak-selection subroutine (PUSHSUB), the automated knowledge-acquisition subroutine (AUTOGEN), the system optimization subroutine (STO), the interpretation subroutine (PAIRS), and final processing subroutine to subtract spectral similarity (PAIRSPLUS). Principal advantages of this system compared to earlier systems are speed, flexibility and accuracy.

In order to satisfy the requirements of hazardous waste analysis [1–6], a program for automated waste mixture identification (PAWMI) through the interpretation of the infrared (IR) spectrum of the waste mixture was developed [7, 8] and tested on samples from hazardous waste drums [9]. Two limitations of PAWMI were that once a training set, consisting of a library of reference spectra, was defined, the rules for the inference engine (PAIRS) [10–16] had to be generated manually. The second limitation was that the PAWMI software for compound identification only uses information on peak location.

An approach to the automated generation of functional group interpretation rules for PAIRS was previously developed [16]. This system defines an "occurrence" value, which was used to weight information on peak position for the generation of expectation values for the presence of certain functional groups.

0003-2670/88/$03.50

Efforts by other investigators have included the fuzzy data set [17], as well as hierarchical tree [18] and table-driven [19] programs developed by Munk and co-workers, and the pattern recognition approach of Frankel [20]. Most of these systems were primarily aimed at identifying functional groups in compounds, as was the original PAIRS program. A related work aimed primarily at identifying compounds in mixtures was that of Lowry and Huppler [21], which used a search system based on Boolean logic. Many of these approaches owe their origins to earlier efforts by Jurs, Isenhour and co-workers [22–24]. Recent publications have included improvements in the PAIRS and hierarchical tree approaches, multi-spectroscopy expert systems, and various computer-aided spectral interpretation systems [25–27].

In this paper, a program is described for the identification of the principal components of mixtures based on computer-assisted interpretation of the infrared spectrum of the mixture. This program (IntIRpret) has five main subroutines: the interferogram processing and peak-selection subroutine (PUSHSUB) [8], the automated knowledge-acquisition subroutine [16] (AUTOGEN), the system training and optimization subroutine (STO), the interpretation subroutine (PAIRS) [7, 10–16], and final processing subroutine to subtract spectral similarity (PAIRSPLUS) [8].

The principal advantages of this system compared to the previously reported PAWMI system are speed (all spectral information is encoded automatically), flexibility (changes in the data base and in interpretation rules are readily accommodated) and accuracy (interpretation is based on peak position, frequency of occurrence and peak size, each of which is weighted in an optimal fashion).

The method was evaluated for the 62 most commonly identified organic compounds on hazardous waste sites [8,9]. IntIRpret was designed to be automatic, self-training, and self-optimizing.

EXPERIMENTAL

All solvents were Aldrich Spectrophotometric Grade or equivalent. Mixtures were prepared on a weight basis. Film transmission spectra were acquired by placing a drop of sample between two KBr crystals. Spectra were acquired on a Nicolet 20-SX optical bench. Each spectrum was generated with background- and sample-signal averaging over 128 scans. The number of data points collected was 16 384 resulting in a nominal spectral resolution of 2 cm^{-1}. All programming and spectral analysis, including rule writing, compiling and spectral interpretation were done with a Nicolet 1280 computer.

RESULTS AND DISCUSSION

IntIRpret has five main subroutines: the interferogram processing and peak-selection subroutine (PUSHSUB) [8], the automated knowledge acquisition

subroutine [16] (AUTOGEN), the system optimization subroutine (STO), the inference engine (PAIRS) [7, 10–16], and the final processing subroutine which subtracts spectral similarity (PAIRSPLUS) [8]. Figure 1 is a flow chart of the IntIRpret process, where the logic of each of the five major subroutines is outlined.

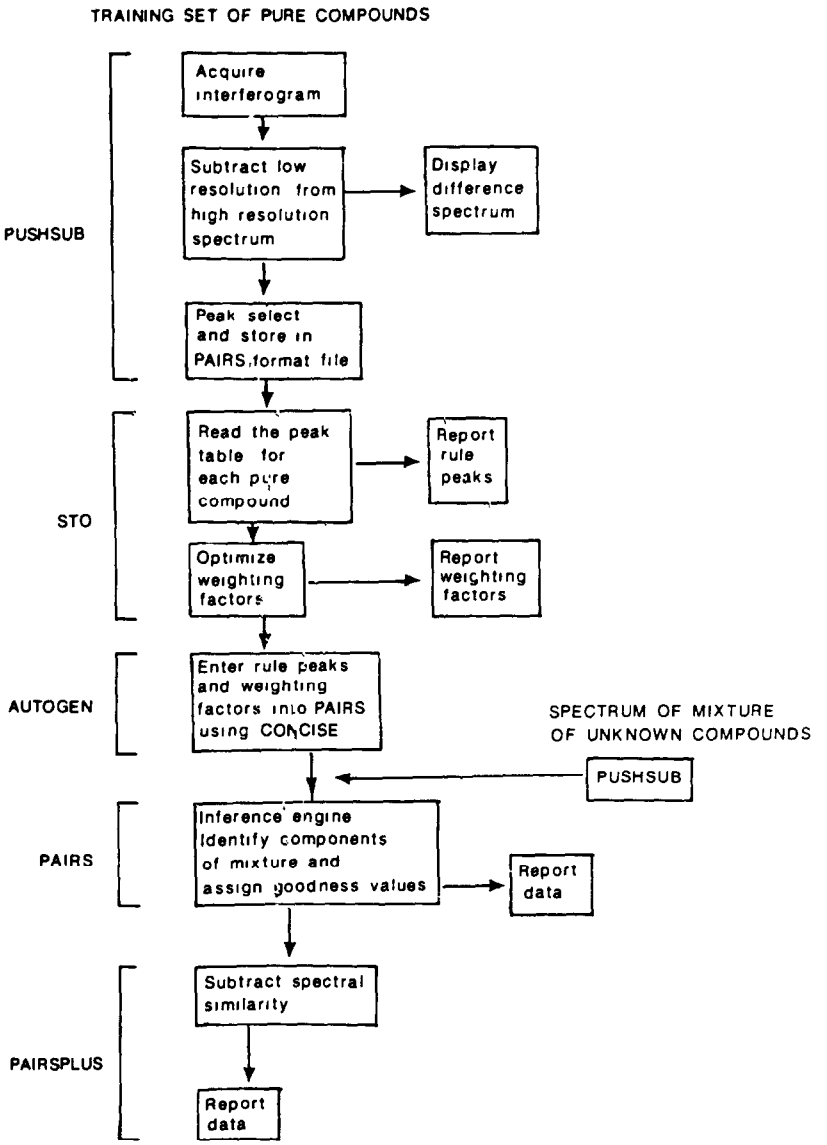

Fig. 1. Flow chart of the IntIRpret process, showing the logic of each of the five major subroutines.

Here, emphasis is placed on describing STO, which is central to the operation of the self-training, self-optimizing mode of operation of IntIRpret.

PUSHSUB

In order to automate PAWMI, a peak-selection subroutine PUSHSUB, was developed that does not require the operator to set a peak-selection threshold, and successfully follows nonlinear baselines [8]. PUSHSUB selects peaks by transforming the first 256 data points right of the center-burst from the original 16 384-point sample interferogram into a threshold curve. PUSHSUB automatically calculates the threshold value from this file. This has been described in detail [8]. PUSHSUB stores the peak file in a format that can be used by AUTOGEN and STO.

STO

This subroutine, the flow diagram of which is given in Fig. 2, accesses the peak tables generated by PUSHSUB. The peaks in a spectrum that are chosen for the purposes of decision-making are called rule peaks. Not all spectral peaks are rule peaks. Each rule peak is assigned a "goodness value" that indicates the probable presence or absence of each compound in the training set. The question of "goodness" has been discussed [7, 8]. It is the purpose of the STO program to utilize the maximal amount of spectral information in an effort to enhance the predictive power of the goodness value.

Three factors are used to weight the goodness values assigned to each rule peak listed by AUTOGEN: $K1$ (frequency of occurrence), $K2$ (intensity), and $K3$ (frequency of occurrence $\times$ intensity). These three factors are designed to follow the logic used by an expert during the interpretation of the infrared spectra of mixtures. In this respect, the underlying intellectual framework is similar to that described earlier [28, 29], in which match factors were automatically calculated for the interpretation of mass spectra.

STO is structured around five subroutines, plus a "main", or driver, program. SUB 1 reads the peak table for each compound that was generated by PUSHSUB. The peak table is compared to the operator-defined window-widths. If there is more than one peak in any given window, only the largest is retained. This results in the loss of potentially useful information, but it greatly simplifies later steps in the program with no apparent degradation of results. SUB 2 reads the peak tables of all spectra in the training set and creates an array consisting of peak position and intensity information. This is used for the calculations done in SUB 4 and SUB 5.

SUB 3 decides which peaks in the peak table of each compound will be used for rule peaks for the PAIRS inference engine. This subroutine is designed to pick the largest peaks in the spectrum, up to a maximum of 20 peaks. If less than 20 peaks are present when the highest threshold value is used, the threshold is automatically lowered incrementally until 20 peaks are chosen. In some

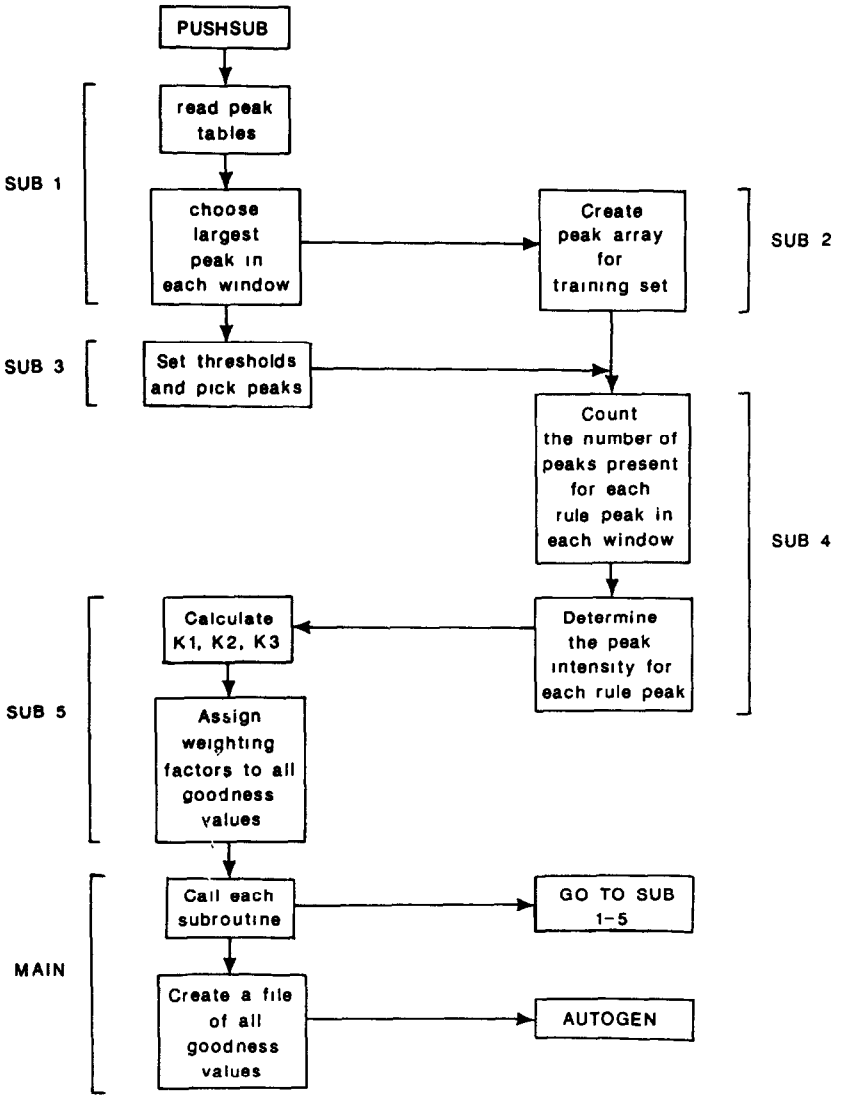

Fig. 2. Flow chart of the system training and optimization (STO) process, showing the relationship between each of the five subroutines, and the main driver subroutine.

cases, 20 peaks will not be present even at a low threshold, so the number of peaks necessary to satisfy this step of the program is lowered along with the threshold. If at least three peaks are not present at a threshold of $\geqslant 3\%$ of the largest peak in the spectrum, then an error message is printed, and the spectrum of that compound in the training set must be re-examined by the opera-

tor. If the criteria for numbers of rule peaks and threshold are satisfied, the rule peak array is created from the information in SUB 2.

SUB 4 analyzes the frequency of occurrence and intensity of data in the spectral array created by SUB 2 and SUB 3. This procedure counts the number of peaks within the window width surrounding each rule peak. The default value of the window widths was set at ± 3, 5, and 10 cm^{-1}, which compensates for peak shifts expected in condensed phase mixtures [7, 8, 10–17]. For example, for a peak at 1036 cm^{-1} in the spectrum of benzene, there are eleven other peaks for spectra in the training set within the tightest window of ± 3 cm^{-1}, 19 other peaks present within the ± 5 cm^{-1} window, and 26 other peaks within the ± 10 cm^{-1} window. This information is utilized to assign weighted goodness values in SUB 5.

A similar calculation is done for the peak-intensity parameter. All peaks within the preset windows are not only counted, but their intensities are summed. This information is also used in SUB 5.

SUB 5 divides the total goodness between peaks and peak windows (Fig. 3). The first division of goodness is between windows, with the default value set at 50% for the tightest window, and 30% and 20% for the remaining two increasingly wide windows. These default values can be changed by the operator, if so desired. Secondly, the factors $K1$, $K2$ and $K3$ are defined by the program. The goodness available to each peak window is divided between $K1$, $K2$ and K3, with the default value for the constants set equal. These default values can be changed by the operator.

Data generated by SUB 5 is accessed by the MAIN or driver program, which both calls SUB 1–5 in sequence, and then creates a file for storing the goodness value for each window, peak, and compound in the training set. This data is stored in a form that is usable by AUTOGEN.

An example is the generation of the optimized goodness value for the rule peaks of benzene (Table 1). The values of $K1$, $K2$ and $K3$ are given for each of the peaks. For each compound, a total of 100 000 goodness units are allocated by STO. This is a change from the original PAIRS program in which 100 goodness units were allocated to the spectrum of each pure compound. For benzene, the allocation is made by apportioning the goodness between six rule peaks. The peak at 674 cm^{-1} is illustrative of the manner in which the system works. This peak is the largest in the spectrum of benzene, therefore the $K2$ value is the highest, with a value of 9821, 5892, and 3928, totalling 19 641 goodness units. The value 19 641 can be found in Table 1. These three values are for the ± 3, 5, and 10 cm^{-1} windows, and represent an allocation of 50, 30 and 20% of the total $K2$ goodness.

The peak at 674 cm^{-1} has 4, 8, and 13 peaks in all of the other spectra of the compounds in the training set within ± 3, 5, and 10 cm^{-1} windows. Thus, the peak is in a window in which the frequency of occurrence of potentially interfering peaks is low, and the $K1$ values are correspondingly high. These are set

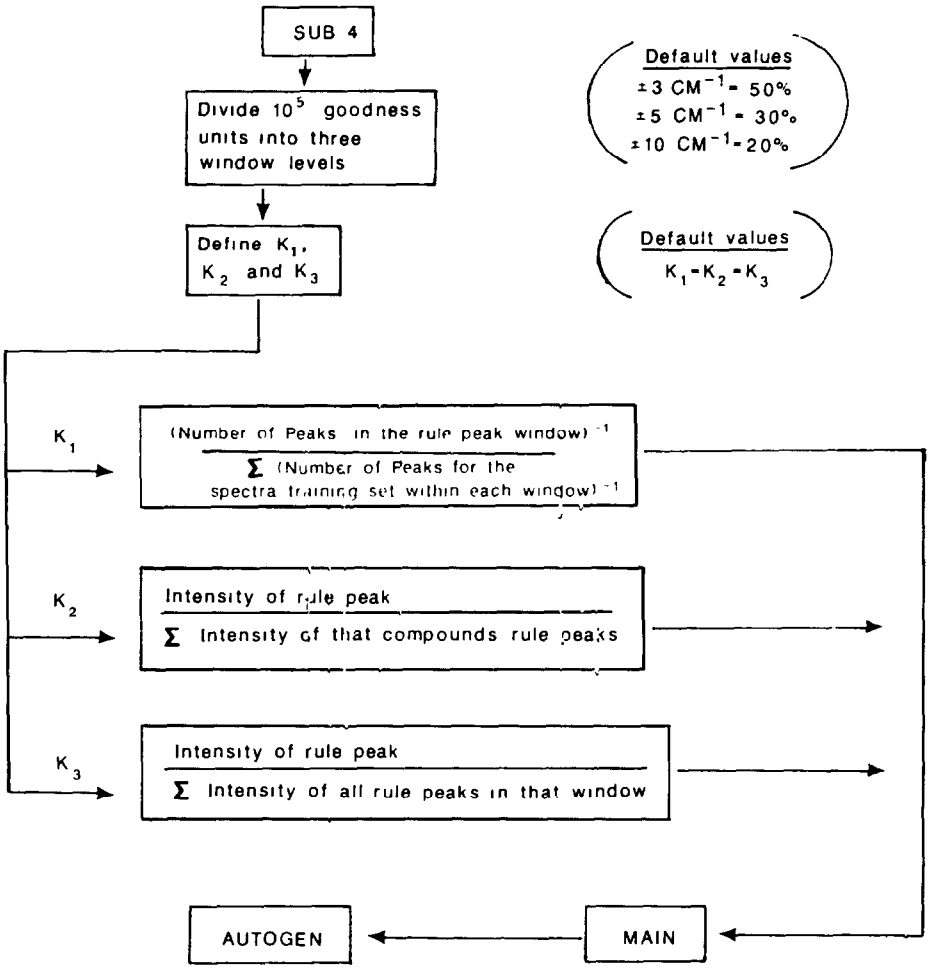

Fig. 3. Flow chart of STO SUB 5, showing the relationship between *K*1, *K*2 and *K*3.

at 3450, 1991, and 1218 , respectively, totalling 6659, which is the value given in Table 1.

The total intensity, on a scale where the largest peak in a spectrum has an intensity of 99, of all other peaks in the spectra of the compounds in the training set, is 177, 335, and 502 for the three windows surrounding the 674 cm^{-1} peak. Thus, not only does this peak occur at a location where there are few other peaks in the spectra of other compounds in the training set, but those other peaks are relatively small. Therefore, the *K*3 values for this peak are set at the relatively high values of 4696, 3439, and 2547 for the three windows, totalling 10 682, which is the value found in Table 1.

The total goodness assigned to the peak at 674 cm^{-1} is 36 986, or 37% of the

TABLE 1

Comparison of $K1$, $K2$, $K3$ and goodness values for six peaks in the spectrum of benzene. Values associated with the peak at 674 cm^{-1} are discussed in the text

Peak position (cm^{-1})	674	1036	1479	3036	3071	3091
Relative intensity (0–99)	99	9	28	18	5	9
Number of peaks in all spectra in						
± 3 cm^{-1}	4	11	6	5	6	3
± 5 cm^{-1}	8	19	10	8	10	8
± 10 cm^{-1}	13	26	14	14	16	10
$K1$ (total for all 3 windows)[a]	6659	2701	5024	5882	4883	8175
$K2$ (total for all 3 windows)[b]	19 641	1784	5554	3570	991	1784
Total intensity of peaks in all spectra in						
± 3 cm^{-1}	177	203	228	40	24	15
± 5 cm^{-1}	335	254	368	80	39	91
± 10 cm^{-1}	502	515	446	170	110	103
$K3$ (total for all 3 windows)[c]	10 682	1009	2726	7763	3828	7317
$K1+K2+K3$ for						
± 3 cm^{-1}	17 968	2519	6109	8324	4545	10 537
± 5 cm^{-1}	11 324	1786	4145	5681	3383	3678
± 10 cm^{-1}	7694	1192	3053	3213	1775	3070
Total[d]	36 986	5497	13 307	17 218	9703	17 279

[a]Apportioned based on the reciprocal of the number of peaks in window in all spectra $\times 0.5$, 0.3, and 0.2 for the three window widths. [b] Apportioned based on 0.5, 0.3, and 0.2 for the three window widths. [c]Apportioned based on the reciprocal of the total intensity of peaks in window in all spectra $\times 0.5$, 0.3 and 0.2 for the three window widths. [d] Divide by 1000 for % contribution.

goodness for all of the peaks in the entire spectrum of six rule peaks. Goodness is divided into 18% for $K1$, 53% for $K2$, and 29% for $K3$.

As stated earlier, the default values chosen for this study were: window widths of ± 3, 5, and 10 cm^{-1}; goodness values divided between these windows of 50%, 30%, and 20%, respectively; and $K1 = K2 = K3$. It is not known if these are the optimal values for this training set, for all possible mixtures that can be prepared for compounds in this training set, or for other training sets.

Using the STO portion of IntIRpret allows the optimization of goodness values for each rule peak in each training set.

AUTOGEN

The automated generation of rules for a defined training set is essential to the success of this approach. Without AUTOGEN, PAIRS and PAWMI are hampered by the potential for errors that always occurs when data is manually encoded, and by the constraints imposed by the length of time it takes to enter data for new or modified training sets. Because of these problems, such a system is inherently inflexible. AUTOGEN solves these problems.

This subroutine has been modified from the program first reported [16]. The earlier version generated single level rules, plus a value for each functional group called "occurrence". Occurrence was used to generate a "maximum expectation value" which related the probability of the presence of a peak that was associated with a given functional group in a given wavenumber range. The present version generates a three-level filter algorithm. The intensity algorithm was also modified to generate information based on a scale of 0–99, rather than the previously utilized 0–9 scale.

Completion of the running of AUTOGEN for a given training set generates a complete set of three-level "if-then" rules for the PAIRS inference engine. If STO had not been used, goodness values, which are a measure of the probability of the presence of an unknown compound in a mixture, would be assigned on an equal basis to each peak in each spectrum of the training set. The use of STO allows the optimized goodness values to be entered in the rules by AUTOGEN for use by PAIRS.

PAIRS

As previously reported, PAIRS [10–16] was modified in the PAWMI program [7, 8]. The goodness scale ranges from 0.001 for a complete mismatch to 0.999 for a complete match.

In the IntIRpret program, peaks in the library spectra are picked by PUSHSUB, the goodness values are weighted by STO, and the three-level rules are written by AUTOGEN. A peak table is then created for the unknown mixture by PUSHSUB. PAIRS accesses that table and generates goodness values that indicate the probable presence of compounds in the mixture of unknowns.

PAIRSPLUS

PAIRSPLUS was developed to limit the effect of spectral similarity and has been described in detail [8]. Statistical studies have been conducted [7] to evaluate the quality of the final goodness value reported by PAIRS for actual compound assignment. Based on these results, a goodness value greater than 0.60 out of a possible 0.99 indicated the likely presence of the compound in the unknown spectrum. However, if many compounds in the training set are spec-

trally similar, then goodness values greater than 0.60 may be returned for these compounds as well. Because these compounds are not actually in the sample, but are predicted to be there, they are considered false positives.

PAIRSPLUS accesses both the complete array of known spectra and the PAIRS interpretation results and subtracts the percentage of spectral similarity corresponding to the compound with the largest goodness value from all the goodness values of the remaining compounds. Note that this is a subtraction of goodness values, not of actual spectra. A statistical check is then conducted on the remaining compounds to establish if another compound should be reported as present in the unknown sample. This is accomplished by calculating the mean and standard deviation of the remaining goodness values. If the next largest goodness value in the remaining spectra is greater than the 95% confidence interval, it is reported, the array is accessed, and the percentage of spectral similarity corresponding to its goodness value is subtracted from the goodness values of all the remaining compounds. This is repeated until the statistical check establishes that there are no goodness values greater than the 95% confidence interval. At that point, the program terminates.

In conclusion, results obtained through the use of PAWMI and IntIRpret are shown in Table 2 for the training set of the spectra of 62 compounds frequently found at hazardous waste sites [3] and 67 four-component mixtures of those compounds. As stated previously, the differences between PAWMI and IntIRpret are the subroutines STO and AUTOGEN, and a minor improvement in PAIRSPLUS. Thus, in PAWMI, rule peaks are operator chosen and

TABLE 2

Results obtained with the PAIRS and PAIRSPLUS subroutines of PAWMI and IntIRpret. The training set consisted of 62 compounds frequently found at hazardous waste sites [3]. The test mixtures consisted of 67 four-component mixtures of chlorobenzene, 1,1,1-trichloroethane, toluene, and benzene

	PAIRSPLUS results	
	With PAWMI[a]	With IntIRpret
Positives		
True	200	216
False	77	46
Improvement	—	40%
Negatives		
True	3809	3840
False	68	52
Improvement	—	24%
Total decisions	4154	4154

[a]These data do not match those previously reported [8] because the data set has been altered.

entered by hand into PAIRS using the subroutine, CONCISE. All peaks are weighted equally, and a three-level logic structure is used to compensate for shifts of peak positions from the spectrum of the pure compound to the spectrum of the mixture.

In IntIRpret, rule peaks are chosen by STO, and weighted for frequency of occurrence ($K1$), intensity ($K2$), and for the cross-term ($K3$). Rules are entered automatically by AUTOGEN and compiled into PAIRS. The software system is several orders of magnitude faster than when peaks are entered manually, is immune from mistakes made when complex data is entered manually, and is based on results that are consistently applied regardless of the operator or data set.

These data show a 40% decrease in false positive results and a 24% decrease in false negative results when IntIRpret is compared to PAWMI. Some additional improvements in results can be expected after completion of a study of the optimal values of window widths, window weighting factors and the relative weights of $K1$, $K2$ and $K3$. However, a certain degree of uncertainty will remain in the direct interpretation of the infrared spectra of mixtures because of peak shifts in solution, the similarity of the spectra of structurally similar compounds, and the inability of the peak-picking routines to recognize the presence of peaks that appear as unresolved shoulders or in poorly resolved envelopes.

The authors thank Greg Kinnes for his help in preparing the mixtures and acquiring the IR spectra, and Mary Weed for preparation of manuscript figures. This work was supported by grants 1-R01-OHO2066-01 and OHO2404-01 from the National Institute for Occupational Safety and Health of Centers for Disease Control.

REFERENCES

1 M.A. Puskar, S.P. Levine and R. Turpin, in S.P. Levine and W.F. Martin (Eds.), Protecting Personnel at Hazardous Waste Sites, Butterworths/Ann Arbor, Woburn, MA, 1985, Chap. 6.
2 D.F. Gurka, Project Summary: Interlaboratory Comparison Study: Methods for Volatile and Semivolatile Compounds, Environmental Monitoring Systems Laboratory, EPA-600/S4-84-027, Las Vegas, NV, June, 1984.
3 P.A. Hallstedt, M.A. Puskar and S.P. Levine, J. Haz. Waste Haz. Mater., 3(2) (1986) 221.
4 W.P. Eckel, D.P. Trees and S.P. Kovell, Distribution and Concentration of Chemicals and Toxic Materials Found at Hazardous Waste Dump Sites, Proc. Natl. Conf. Haz. Waste Environ. Emergencies, Washington, DC, May, 1985.
5 J.D. Mayhew, G.M. Sodaro and D.W. Carroll, A Hazardous Waste Site Management Plan, Chemical Manufacturers Association, Washington, DC, 1982.
6 The Hazardous and Solid Waste Amendments of 1984, Congressional Record, H11103, Washington, DC, Oct., 1984.
7 M.A. Puskar, S.P. Levine and S.R. Lowry, Anal. Chem., 58 (1986) 1156.
8 M.A. Puskar, S.P. Levine and S.R. Lowry, Anal. Chem., 58 (1986) 1981.
9 M.A. Puskar, S.P. Levine and S.R. Lowry, Environ. Sci. Technol., 21 (1987) 90.

10 H.B. Woodruff and M.E. Munk, J. Org. Chem., 42(10) (1977) 1761.
11 H.B. Woodruff and M.E. Munk, Anal. Chim. Acta, 95 (1977) 13.
12 H.B. Woodruff and G.M. Smith, Anal. Chem., 52 (1980) 2321.
13 H.B. Woodruff and G.M. Smith, Anal. Chim. Acta , 133 (1981) 545.
14 S.A. Tomellini, D.D. Saperstein, J.M. Stevenson, G.M. Smith and H.B. Woodruff, Anal. Chem., 53 (1981) 2367.
15 S.A. Tomellini, J.M. Stevenson and H.B. Woodruff, Anal. Chem., 56 (1984) 67.
16 S.A. Tomellini, R.A. Hartwick, J.M. Stevenson and H.B. Woodruff, Anal. Chim. Acta, 162 (1984) 227.
17 T. Blaffert, Anal. Chim. Acta, 161 (1984) 135.
18 J. Zupan and M.E. Munk, Anal. Chem., 57 (1985) 1609.
19 M.O. Trulson and M.E. Munk, Anal. Chem., 55 (1983) 2137.
20 D.S. Frankel, Anal. Chem., 56 (1984) 1011.
21 S.R. Lowry and D.A. Huppler, Anal. Chem., 55 (1983) 1288.
22 P.C. Jurs and T.L. Isenhour, Applications of Pattern Recognition, Wiley, New York, 1975.
23 G.T. Rasmussen, T.L. Isenhour, S.R. Lowry and G.L. Ritter, Anal. Chim. Acta, 103 (1978) 213.
24 J.A. de Haseth, H.B. Woodruff, S.R. Lowry and T.L. Isenhour, Anal. Chim. Acta, 103 (1978) 109.
25 D.D. Saperstein, Appl. Spectrosc., 40(3) (1986) 344.
26 J. Zupan (Ed.), Computer Supported Data Bases, Howard/Wiley, New York, 1986.
27 P.C. Jurs, in P.C. Jurs (Ed.), Computer Software Applications in Chemistry, Wiley, New York, 1986, Chap. 16.
28 K.-S. Kwok, R. Venkataragahaven and F.W. McLafferty, J. Am. Chem. Soc., 95 (1983) 4185.
29 B.L. Atwater, D.B. Stauffer, F.W. McLafferty and D.W. Peterson, Anal. Chem., 57 (1985) 899.