Self-Training, Self-Optimizing Expert System for Interpretation of the Infrared Spectra of Environmental Mixtures

Li-Shi Ying[1] and Steven P. Levine*

School of Public Health, The University of Michigan, Ann Arbor, Michigan 48109

Sterling A. Tomellini

Department of Chemistry, University of New Hampshire, Durham, New Hampshire 03824

Stephen R. Lowry

Nicolet Instrument Corporation, 5225 Verona Road, Madison, Wisconsin 53711

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 organics handled on hazardous waste remedial action sites, 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 the final processing subroutine to subtract spectral similarity (PAIRSPLUS). Principal advantages of this system compared to those previously reported are speed, flexibility, and accuracy. For a training set of 62 pure compounds and a data set of 67 four-component mixtures requiring 4154 decisions, the system correctly identified 216 true positive results and 3840 true negatives and incorrectly identified 46 false positives (19.4%) and 52 false negatives (1.2%).

In order to satisfy the requirements of hazardous waste analysis at Superfund and at licensed disposal sites (*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 hazardous waste drum samples (*9*). This approach, which utilizes the speed and sensitivity of Fourier transform infrared (FT-IR) spectrometry meets many of the requirements of a near real-time, principal component screening technique for organic hazardous waste samples (*1*).

Two limitations of PAWMI were that once a training set, consisting of a library of reference of spectra, was defined, the rules for the inference engine (PAIRS) (*10–16*) had to be generated manually. The second limitation was that the PAWMI compound identification software only uses peak location information.

An approach to the automated generation of functional group interpretation rules for PAIRS was previously developed (*17*). This system defines a value "occurrence" as the "number of peaks in a given wavenumber range divided by the number of compounds in the database containing the functionality of interest". This value was used to weight peak position information for the generation of expectation values for the presence of certain functional groups.

Efforts by other investigators have been successful in the interpretation of IR spectra by using computerized interpretation or matching procedures. These program systems include the hierarchal tree (*18*) and table-driven (*19*) programs developed by Munk et al. and the pattern recognition approach of Frankel (*20*). These systems were primarily aimed at identifying functional groups in compounds, as was the original PAIRS program. The success of these approaches were a function of the data set and functional group studied. A related work aimed primarily at identifying compounds in mixtures was that of Lowry (*21*) which used a Boolean logic based search system. Many of these approaches owe their origins to earlier efforts that originated with Jurs, Isenhour, et al. (*22–24*). Recent publications have included investigations of improvements in the PAIRS and hierarchal tree approaches, discussions of multispectroscopy expert systems, and strengths and weaknesses of various computer-aided spectral interpretation systems (*25–27*).

This paper describes a program for the identification of the principal components of mixtures based on computer assited interpretation of the mixture's infrared spectrum. This program (intIRpret), which was developed as a preliminary screening tool for unknown organics handled on hazardous waste remedial action sites, has five main subroutines: the interferogram processing and peak selection subroutine (PUSHSUB) (*8*), the automated knowledge acquisition subroutine (AUTOGEN) (*17*), the system optimization subroutine (STO), the interpretation subroutine (PAIRS) (*7, 10–16*), and the final processing subroutine to subtract spectral similarity (PAIRSPLUS) (*8*).

Many of these subroutines are substantial modifications of the programs previously reported (*7, 8, 17*). 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 interpetation rules are readily accomodated), and accuracy (interpetation is based on peak position, frequency of occurrence, and peak size, each of which is weighted in an optimal fashion).

The method has been evaluated by using the 62 most commonly identified organic compounds on hazardous waste sites (*18*). IntIRpret was designed to be automatic, self-training, and self-optimizing so it could be operated on-site (in a mobile laboratory) during a remedial action project, by personnel with limited training. Other applications of the intIRpret technique would include screening incoming organic waste at licensed disposal facilities or for interpreting gas-phase spectra obtained during industrial hygiene air monitoring.

EXPERIMENTAL SECTION

All solvents were Aldrich Spectrophotometric Grade or equivalent. Mixtures were prepared on a weight basis. Thin-film

[1] Present address: Shanghai Medical University, Department of Occupational Health, School of Public Health, Shanghai 200032, People's Republic of China.

0003-2700/87/0359-2197$01.50/0

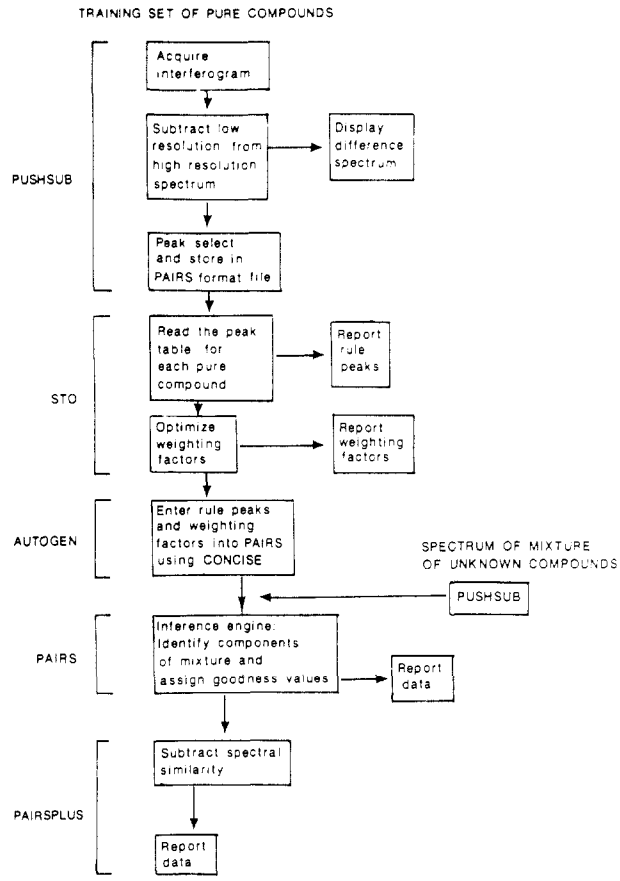

Figure 1. Flow chart of the intIRpret process, showing the logic of each of the five major subroutines.

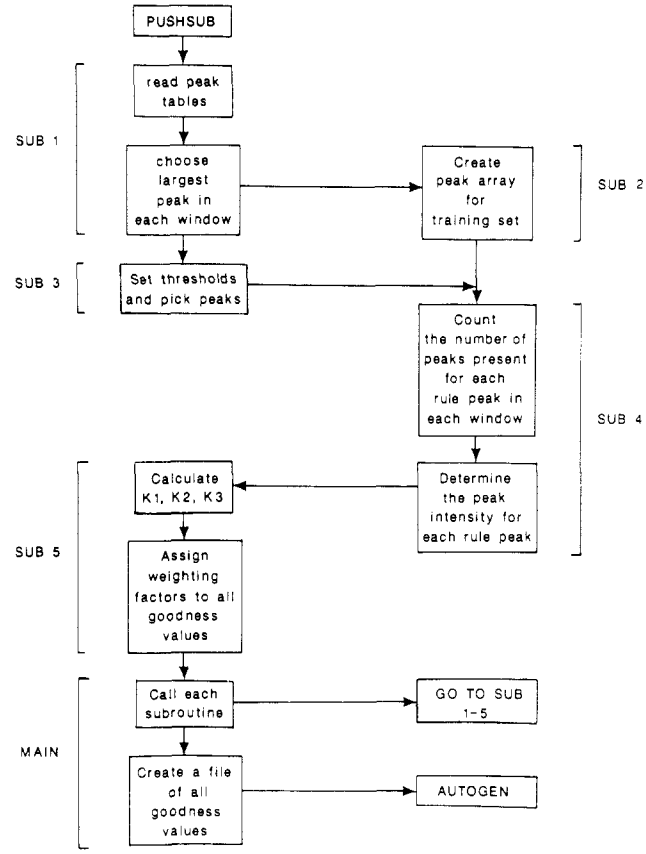

Figure 2. Flow chart of the system training and optimization (STO) process, showing the relationship between each of the five subroutines, and the main driver subroutine.

transmission spectra were acquired by placing a drop of sample between two 13 × 2 mm KBr crystals.

Spectra were acquired on a Nicolet 20-SX optical bench. Each spectrum was generated with a background and sample signal averaging of 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, was performed with a Nicolet 1280 computer equipped with a 160 Mbyte Winchester disk system.

RESULTS AND DISCUSSION

IntIRpret has five main subroutines: the interferogram processing and peak selection subroutine (PUSHSUB) (*8*), the automated knowledge acquisition subroutine (AUTOGEN) (*16*), the system optimization subroutine (STO), the inference engine (PAIRS) (*7, 10–16*), and the final processing subroutine that 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 diagramed.

Because PUSHSUB, AUTOGEN, PAIRS, and PAIRSPLUS have been explained in detail in the above referenced papers, the reader may have to study those papers in order to understand the details of the operation of those subroutines. However, a summary of those subroutines is given in this publication. In this publication, emphasis is placed on describing STO, which is central to the operation of the self-training, self-optimizing mode of operation of intIRpret. In addition, the linkage of all of the subroutines is explained.

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 base lines (*8*). PUSHSUB selects peaks by transforming the first 256 data points right of the centerburst from the original 16 384 data point sample interferogram into a threshold curve. PUSHSUB automatically calculates the threshold value from this file. The threshold may be set at any value, but the resultant horizontal, linear threshold line is usually set at a value between 1.0 and 10% of the maximal peak height. This value is stored in the threshold register, and peak selection is conducted in the spectral domain by the subroutine, PEAK PICKER. This is described in detail in ref 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 Figure 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" is discussed in detail elsewhere (*7, 8*) but is summarized here. The concept of "goodness" was inferred by Woodruff (*12*) as being a probability that the PAIRS program would achieve the correct identification for all functional groups present in a mixture, regardless of which compound caused the observed absorption pattern. It was observed that, in cases in which a spectral feature was caused by a minor impurity in the mixture, the probabilities were reduced from those expected for pure samples. Puskar (*7, 8*), in the PAWMI program which utilizes a modified PAIRS inference engine, found that all single compounds in the training set returned a goodness value of 0.99 when studied individually. Thus, in the case of pure compounds, goodness and probability were identical. Evaluation of data from the analysis of two-, three-, and four-component mixtures showed that a goodness value greater than 0.60 out of a possible 0.99 indicated the presence of a compound. For each training set, a certain number of false positive and false negative results were obtained for mixtures when a goodness of 0.60 was used to indicate the presence of a compound. The statistics as-

Table I. Comparison of k_1, k_2, k_3 and Goodness Values for Six Peaks in the Spectrum of Benzene[a]

peak position, cm^{-1}	674	1036	1479	3036	3071	3091
rel intens (0–99)	99	9	28	18	5	9
no. 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
k_1 (total for all 3 windows)[b]	6659	2701	5024	5882	4883	8175
k_2 (total for all 3 windows)[c]	19641	1784	5554	3570	991	1784
total intens 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
k_3 (total for all 3 windows)[d]	10682	1009	2726	7763	3828	7317
$k_1 + k_2 + k_3$ for						
±3 cm^{-1}	17968	2519	6109	8324	4545	10537
±5 cm^{-1}	11324	1786	4145	5681	3383	3678
±10 cm^{-1}	7694	1192	3053	3213	1775	3070
total[e]	36986	5497	13307	17218	9703	17279

[a] Values associated with the peak at 674 cm^{-1} are discussed in the text. [b] Apportioned based on 1/number of peaks in window in all spectra × 0.5, 0.3, and 0.2 for the three window widths. [c] Apportioned based on 0.5, 0.3, and 0.2 for the three window widths. [d] Apportioned based on 1/total intensity of peaks in window in all spectra × 0.5, 0.3, and 0.2 for the three window widths. [e] Divide by 1000 for percent contribution.

sociated with the goodness–probability relationship are, therefore, a function of the training set, the mixtures, and the structure of the interpretation rules. 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: k_1 (frequency of occurrence), k_2 (intensity), and k_3 (frequency of occurrence × 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 in the work of McLafferty in which match factors were automatically calculated for the interpretation of mass spectra (*28*, *29*).

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 calculations performed 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 there are less than 20 peaks present when using the highest threshold value, the threshold is automatically lowered incrementally until 20 peaks are chosen. In some 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 a minimum of 3 peaks are not present at a threshold of 3% or greater 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 reexamined by the operator. If the criteria for numbers of rule peaks and threshold are satisfield, the rule peak array is created from the information in SUB 2.

SUB 4 performs the frequency of occurrence and intensity analyses of data in the spectra array created by SUB 2 and

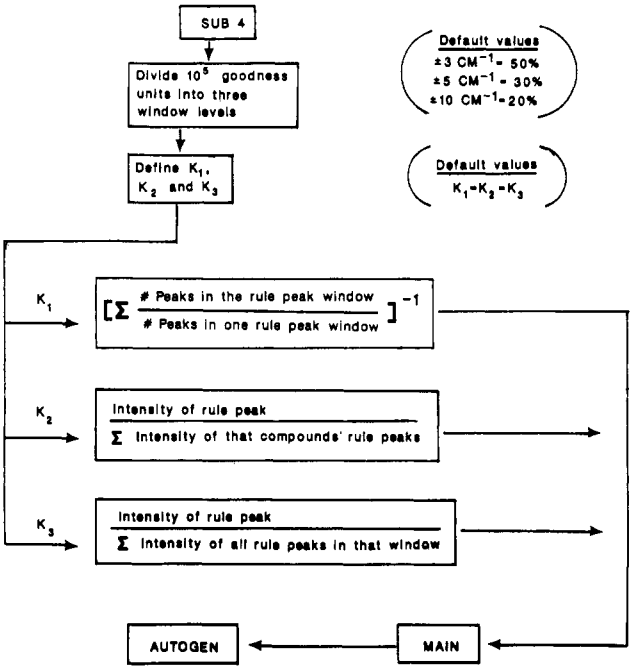

Figure 3. Flow chart of STO SUB 5, showing the relationship between k_1, k_2, and k_3.

SUB 3. The frequency of occurrence analysis 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 11 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 (Table I). This information is utilized to assign weighted goodness values in SUB 5.

A similar calculation is performed 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 (Figure 3). The first division of goodness is between windows, with the default value set at 50% for the tightest

Table II. Comparison of k_1, k_2, k_3, and Goodness Values for Nine Peaks in the Spectrum of Chlorobenzene[a]

peak position, cm^{-1}	468	685	702	740	1022	1083	1445	1477	1584
rel intens (0–99)	31	53	43	99	41	42	30	92	29
no. of peaks in all spectra in									
±3 cm^{-1}	5	6	9	11	12	9	12	8	5
±5 cm^{-1}	6	9	14	16	17	11	22	10	8
±10 cm^{-1}	10	20	22	24	27	22	39	20	20
k_1 (total for all 3 windows)[b]	6345	4454	3118	2667	2446	3352	2114	3728	5095
k_2 (total for all 3 windows)[c]	2245	3840	3114	7172	2970	3042	2172	6666	2100
total intens of peaks in all spectra in									
±3 cm^{-1}	58	221	117	574	129	177	210	357	91
±5 cm^{-1}	88	469	366	741	201	245	352	368	112
±10 cm^{-1}	159	877	697	1188	478	502	968	542	477
k_3 (total for all 3 windows)[d]	7093	2698	3556	2570	3942	3213	1672	4502	4073
$k_1 + k_2 + k_3$ for									
±3 cm^{-1}	7409	5835	5505	5990	4717	4630	3192	6770	5947
±5 cm^{-1}	4706	3158	2490	3695	2808	3022	1704	4684	3668
±10 cm^{-1}	3509	20012	1798	2726	1836	1960	1067	3443	1658
total[e]	15624	10994	9793	12411	9361	9612	5963	14897	11273

[a] Values associated with the peak at 468 cm^{-1} are discussed in the text. [b] Apportioned based on 1/number of peaks in window in all spectra × 0.5, 0.3, and 0.2 for the three window widths. [c] Apportioned based on 0.5, 0.3, and 0.2 for the three window widths. [d] Apportioned based on 1/total intensity of peaks in window in all spectra × 0.5, 0.3 and 0.2 for the three window widths. [e] Divide by 1,000 for percent contribution.

window and 30% and 20% for the remaining two increasingly wide windows. Studies underlying the efforts reported in ref 7 and 8 indicated that these default values gave optimal results for the data set that was studied. These default values can be changed by the operator, if so desired.

Secondly, the factors k_1, k_2, and k_3 are defined by the program. The goodness available to each peak window is divided between k_1, k_2, and k_3, with the default value for the constants set equal. These default values can be changed by the operator.

k_1, which is a measure of frequency of occurrence of peaks in the training set within a given wavenumber window, essentially states that a peak should be given added importance (or goodness) if it is in a region of the spectrum in which there are few peaks in the other spectra in the training set. k_1 is equal to

$$(\text{number of peaks in the rule peak window})^{-1} / \sum(\text{number of peaks for the spectral training set within each window})^{-1}$$

k_2 essentially states that added significance should be attached to the presence of a peak that represents a large fraction of the total peak intensity for the spectrum of a compound. k_2 is equal to

$$(\text{rule peak intensity}) / (\text{total intensity for all peaks in the spectrum of that compound})$$

k_3, which is the cross-term between frequency of occurrence and intensity, essentially states that a large peak should be given added importance if it is in a region of the spectrum in which there are few large peaks in the other spectra of the training set. k_3 is equal to

$$(\text{rule peak intensity}) / (\text{intensity of all peaks in that window in all specra in the training set})$$

Data generated by SUB 5 is accessed by the MAIN or driver program, which 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 useable by AUTOGEN.

An example of the generation of the optimized goodness value for the rule peaks of two compounds, benzene and chlorobenzene, are given in Tables I and II, respectively. The values of k_1, k_2, and k_3 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. The peak at 674 cm^{-1} is the largest peak in the spectrum of benzene, therefore the k_2 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 I. 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 k_2 goodness. This is the default value for the allocation of the percent goodness between the windows.

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 k_1 values are correspondingly high. These are set at 3450, 1991, and 1218, respectively, totaling 6659, which is the value found in Table I.

The total intensity, on a scale where the largest peak in a given 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 I.

The total goodness assigned to the peak at 674 cm^{-1} is 36 986, or 37% of the goodness for all of the peaks in the entire spectrum of 6 rule peaks. Goodness is divided into 18% for k_1, 53% for k_2, and 29% for k_3.

For chlorobenzene (shown in Table II), the allocation is made by apportioning the goodness between nine rule peaks. The peak at 468 cm^{-1} is illustrative of the manner in which the system works. The peak at 468 cm^{-1} has a relative intensity of 31, and therfore the k_2 value is the third smallest of that assigned to the nine rule peaks, with a value of 1123, 673, and 449, totaling 2245, the value found in Table II. 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 k_2 goodness.

The peak of 468 cm^{-1} has 5, 6, and 10 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 k_1 values are correspondingly high. These are set at 2845, 2003, and 1497, respectively, totaling 6345, the value found in Table II.

The total intensity, on a scale where the largest peak in a given spectrum has an intensity of 99, of all other peaks in the spectra of the compounds in the training set, is 58, 88, and 159 for the three windows surrounding the 468 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, despite the moderate relative intensity of this peak, the k_3 values are set at the relatively high values of 3441, 2089, and 1563 for the three windows (±3, ±5, and ±10 cm^{-1}, respectively), totaling 7093, the value found in Table II.

The total goodness assigned to the peak at 468 cm^{-1} is 15624, or 16% of the goodness for all of the peaks in the entire spectrum of 9 rule peaks. Goodness is divided into 41% for k_1, 14% for k_2, and 45% for k_3.

As stated earlier, the default values chosen for this study were as follows: window widths of ±3, ±5, and ±10 cm^{-1}; goodness values divided between these windows of 50%, 30%, and 20%, respectively; $k_1 = k_2 = k_3$. 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. Studies are underway to define the optimal window width based on experimentally determined peak shifts.

STO is a very significant departure from the practice previously reported for PAWMI (*7, 8*). In that program, only peak position information was used. Use of 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. The nature of these rules is dealt with in great detail in ref 7 and 10–16, and is described in brief in the section of this text describing PAIRS (below). 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 version of AUTOGEN 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, which has been shown to be effective in the identification of compounds in mixtures (*7*). The intensity algorithm has also been modified to generate information based on a scale of 0–99, rather than the previously utilized 0–9 scale.

At the completion of the running of AUTOGEN for a given training set, a complete set of three level "if–then" rules has been generated for the PAIRS inference engine. Manual entry of rules into PAIRS, using the CONCISE program, is not necessary. 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. The use of the term "goodness" is discussed more completely at the beginning of the discussion of STO.

PAIRS. As previously reported, PAIRS (*10–16*) was modified in the PAWMI program (*7, 8*). The mixture interpretation software uses peak location information and is based on a three-level filter algorithm designed to compensate for potential peak shifts in the mixture spectrum. If a peak falls in a relatively wide frequency window assigned to a certain compound, a percentage of the overall goodness value will be added to the total. "Goodness" is a measurement of closeness of match between the spectrum of the pure compounds used for rule generation and the spectrum of the unknown compound(s). 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. As previously reported (*8*), PAIRSPLUS was developed to limit the effect of spectral similarity. A detailed description of PAIRSPLUS can be found in ref 8. Studies have been performed (*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 spectrally 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 similarly corresponding to the compound with largest goodness value from all the remaining compounds' goodness values. Note that this is a subtraction of goodness values, not of actual spectra. A statistical check is then conducted on the remaining compounds to determine 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 determines there are no goodness values greater than the 95% confidence interval. At that point, the program terminates.

PAIRSPLUS has been modified for this version to recognize the case in which compounds of equal probability, and therefore equal goodness values, may occur in a mixture. In that case, the PAIRSPLUS program runs twice from the point of the equal goodness values. This results in two tables of results, which may differ slightly in the indicated minor, or lower goodness, components. This is an extremely rare occurrence, but the program is designed to recognize and adjust for the situation.

Results obtained by using intIRpret are given for both PAIRS and PAIRSPLUS goodness values. That is, results can be obtained prior to (PAIRS) or after (PAIRSPLUS)

Table III. Results Obtained by Using the PAIRS and PAIRSPLUS Subroutines of PAWMI and intIRpret[a]

compd in mixture	PAWMI (8)	intIRpret
PAIRS (Goodness)		
TCE	0.99	0.999
chlorobenzene	0.90	0.999
toluene	0.82	0.835
benzene	0.32[b]	0.520[b]
PAIRSPLUS (Goodness)		
TCE	100	1000
chlorobenzene	82	835
toluene	47	631
benzene	NP[b,c]	233

[a] The test mixture is chlorobenzene + 1,1,1-trichloroethane (TCE) + toluene + benzene in 1:1:0.5:0.1 ratio (w/w). [b] Reported as a false negative result. [c] NP = not present; goodness value not significantly different from other low goodness matches at the 95% confidence value.

Table IV. Results Obtained by Using the PAIRS and PAIRSPLUS Subroutines of PAWMI and intIRpret[a]

compd in mixture	PAWMI (8)	intIRpret
PAIRS (Goodness)		
TCE	0.99	0.999
benzene	0.97	0.999
toluene	0.74	0.878
chlorobenzene	0.58[b]	0.658
PAIRSPLUS (Goodness)		
TCE	100	1000
benzene	97	1000
toluene	70	847
chlorobenzene	NP[b,c]	138

[a] The test mixture is chlorobenzene + 1,1,1-trichloroethane (TCE) + toluene + benzene in 0.1:1:1:1 ratio (w/w). [b] Reported as a false negative result. [c] NP = not present; goodness value not significantly different from other low goodness matches at the 95% confidence value.

subtracting the goodness due to spectral similarity. These data are given in Tables III and IV for four-component mixtures. In each case, the mixture consists of toluene, chlorobenzene, 1,1,1-trichloroethane (TCE), and benzene. In these tables, the results obtained by using the previously reported PAWMI program (7, 8) are compared to results obtained by using intIRpret.

The results presented in Table III are for such a mixture in which benzene is the minor component (at 3.85% (w/w) of the total). The PAIRS portion of the PAWMI program correctly reported the presence of toluene, chlorobenzene, and TCE. Benzene was reported as a false negative result since the goodness value is 0.32, which is less than the critical value set at 0.60. This is an operator-definable value chosen to represent the dividing line between positive and negative results. For environmental applications, 0.60 is used as the default value, since the probability of reporting a false negative result is very low (7, 8). The intIRpret program also gives a false negative indication for benzene, but the goodness value obtained from PAIRS is substantially improved for benzene and for chlorobenzene.

When PAWMI is used, both PAIRS and PAIRSPLUS incorrectly report the absence of benzene. When intIRpret is used, PAIRSPLUS correctly reports the presence of all four components of the mixture. Note that the PAIRSPLUS goodness value does not have a cutoff for positive results at 0.60. Instead, the program continues to run as long as there

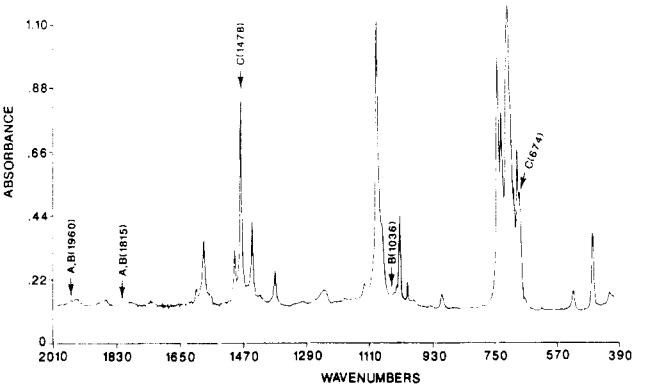

Figure 4. Portion of the spectrum of the mixture chlorobenzene + 1,1,1-trichloroethane (TCE) + toluene + benzene in 1:1:0.5:0.1 ratio (w/w). Peaks used as PAWMI rule peaks but not intIRpret rule peaks are shown by A; peaks missed by PUSHSUB are shown by B; peaks heavily weighted in the spectrum of benzene by intIRpret that fall within the ±3-cm^{-1} window are marked by C.

is a 95% probability that the next compound in the PAIRS goodness list is actually present.

The reason for the difference in results is that PAWMI utilized peaks at 1815 and 1960 cm^{-1} as rule peaks, which the STO subroutine of the intIRpret program ignored. These peaks were missed by the PUSHSUB peak picking subroutine. In addition, the peak at 1036 cm^{-1} was also overlooked in the mixture by PUSHSUB but was a peak considered to have minor significance by STO. Lastly, of the four peaks that fell within the ±3- or ±5-cm^{-1} windows in the mixture spectrum, two (674 and 1478 cm^{-1}) were heavily weighted by STO. This is illustrated in Figure 4.

Results for a mixture having chlorobenzene as the minor component (3.23% (w/w) of the total) are given in Table IV. The PAIRS segment of the PAWMI program correctly reported the presence of TCE, benzene, and toluene. Chlorobenzene was reported as a false negative result since the goodness value is 0.58, which is less than the critical value set at 0.60. This difference of 0.02 goodness units may or may not be significant, but PAWMI sets a cutoff of 0.60, with no further interpretation of the data, except that performed by PAIRSPLUS. The intIRpret program correctly identifies all four components of the mixture, and the goodness value obtained from PAIRS is substantially improved for toluene and for chlorobenzene.

After subtracting spectral similarity, PAIRSPLUS still returns a false negative result for chlorobenzene in the PAWMI results, but correctly reports the presence of all four components of the mixture in the intIRpret results.

The reason for the differences in results is that PAWMI utilized a peak at 1068 cm^{-1} as a rule peak, which the STO subroutine of the intIRpret program ignored. This peak was missed by PUSHSUB in the spectrum of the mixture, but all other rule peaks were found. Lastly, of the five peaks that fell within the ±3-cm^{-1} window in the mixture spectrum, four (468, 1083, 1479, and 1584 cm^{-1}) were heavily weighted by STO.

In conclusion, results obtained through the use of PAWMI and intIRpret are shown in Table V 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 difference 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 entered by hand into PAIRS by using the subroutine CONCISE. All peaks are weighed equally, and a three level logic structure is used to compensate for shifts of peak positions from the

Table V. Results Obtained by Using the PAIRSPLUS Subroutine of PAWMI and intIRpret[a]

	PAIRSPLUS results	
	PAWMI[b]	intIRpret
	positives:	
true	200	216
false	77	46
improvement		40%
	negatives:	
true	3809	3840
false	68	52
improvement		24%
total decisions	4154	4154

[a] 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 (TCE), toluene, and benzene. [b] These data do not match those previously reported (8) because the data set has been altered.

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 (k_1), intensity (k_2), and for the cross-term (k_3). Rules are entered automatically by AUTOGEN and compiled into PAIRS. The software system is several orders of magnitude faster than when peaks were entered manually, 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 k_1, k_2, and k_3. However, a certain degree of uncertainty will remain in the direct interpretation of the infrared spectra of mixtures due to 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.

ACKNOWLEDGMENT

The authors thank Greg Kinnes for his help in preparing the mixtures and acquiring the IR spectra and Mary Weed for preparation of figures.

LITERATURE CITED

(1) Puskar, M. A.; Levine, S. P.; Turpin, R. In *Protecting Personnel at Hazardous* Waste Sites, Levine, S. P., Martin, W. F., Eds.; Butterworths/Ann Arbor: Woburn, MA, 1985; Chapter 6.
(2) Gurka, D. F. "Project Summary: Interlaboratory Comparison Study: Methods for Volatile and Semivolatile Compounds", Environmental Monitoring Systems Laboratory, Las Vegas, NV, June 1984; EPA-600/S4-84-027.
(3) Hallstedt, P. A.; Puskar, M. A.; Levine, S. P. *J. Hazard. Waste Hazard. Mater.* **1986**, *3*(2), 221–232.
(4) Eckel, W. P.; Trees, D. P.; Kovell, S. P. "Distribution and Concentration of Chemicals and Toxic Materials Found at Hazardous Waste Dump Sites", Proceedings of the National Conference on Hazardous Waste and Environmental Emergencies, May, 1985.
(5) Mayhew, J. D.; Sodaro, G. M.; Carroll, D. W. *A Hazardous Waste Site Management Plan*; Chemical Manufacturers Association: Washington, D.C., 1982.
(6) "The Hazardous and Solid Waste Amendments of 1984"; *Congr. Rec.* **1984**, (Oct 3), H11103.
(7) Puskar, M. A.; Levine, S. P.; Lowry, S. R. *Anal. Chem.* **1986**, *58*, 1156–1162.
(8) Puskar, M. A.; Levine, S. P.; Lowry, S. R. *Anal. Chem.* **1986**, *58*, 1981–1989.
(9) Puskar, M. A.; Levine, S. P.; Lowry, S. R. *Environ. Sci. Technol.* **1987**, *21*, 90–96.
(10) Woodruff, H. B.; Munk, M. E. *J. Org. Chem.* **1977**, *42*, 1761–1767.
(11) Woodruff, H. B.; Munk, M. E. *Anal. Chim. Acta* **1977**, *95*, 13–23.
(12) Woodruff, H. B.; Smith, G. M. *Anal. Chem.* **1980**, *52*, 2321–2327.
(13) Woodruff, H. B.; Smith, G. M. *Anal. Chim. Acta* **1981**, *133*, 545–553.
(14) Tomellini, S. A.; Saperstein, D. D.; Stevenson, J. M.; Smith, G. M.; Woodruff, H. B. *Anal. Chem.* **1981**, *53*, 2367–2369.
(15) Tomellini, S. A.; Stevenson, J. M.; Woodruff, H. B. *Anal. Chem.* **1984**, *56*, 67–70.
(16) Tomellini, S. A.; Hartwick, R. A.; Stevenson, J. M.; Woodruff, H. B. *Anal. Chim. Acta* **1984**, *162*, 227–240.
(17) Blaffert, T. *Anal. Chim. Acta.* **1984**, *161*, 135–148.
(18) Zupan, J.; Munk, M. E. *Anal. Chem.* **1985**, *57*, 1609–1616.
(19) Trulson, M. O.; Munk, M. E. *Anal. Chem.* **1983**, *55*, 2137–2142.
(20) Frankel, D. S. *Anal. Chem.* **1984**, *56*, 1011–1014.
(21) Lowry, S. R.; Huppler, D. A. *Anal. Chem.* **1983**, *55*, 1288–1291.
(22) Jurs, P. C.; Isenhour, T. L. *Applications of Pattern Recognition*; Wiley: New York, 1975.
(23) Rasmussen, G. T.; Isenhour, T. L.; Lowry, S. R.; Ritter, G. L. *Anal. Chim. Acta.* **1978**, *103*, 213–221.
(24) de Haseth, J. A.; Woodruff, H. B.; Lowry, S. R.; Isenhour, T. L. *Anal. Chim. Acta* **1978**, *103*, 109–120.
(25) Saperstein, D. D. *Appl. Spectrosc.* **1986**, *40*(3), 344–348.
(26) *Computer Supported Data Bases*; Zupin, J., Ed.; Howard Ltd.-Wiley Co.: New York 1986.
(27) Jurs, P. C. In *Computer Software Applications* in *Chemistry*; Jurs, P. C., Ed.; Wiley: New York, 1986; Chapter 16.
(28) Kwok, K-S; Venkataragahaven, R.; McLafferty, F. W. *J. Am. Chem. Soc.* **1973**, *95*, 4185–4194.
(29) Atwater, B. L.; Stauffer, D. B.; McLafferty, F. W.; Peterson, D. W. *Anal. Chem.* **1985**, *57*, 899–903.

RECEIVED for review March 6, 1987. Accepted May 26, 1987. This work was supported by Grant 1-R01-OH02066-01 from the National Institute for Occupational Safety and Health of Centers for Disease Control.

Low-Pressure Laser Spectroscopy with Flame Atomization

W. B. Whitten,* L. B. Koutny, T. G. Nolan, and J. M. Ramsey

Oak Ridge National Laboratory, Analytical Chemistry Division, P.O. Box X, Oak Ridge, Tennessee 37831-6142

We have developed a low-pressure interface that permits high-resolution laser spectroscopy to be performed with an air–acetylene analytical burner. The reduced pressure in the measuring cell effectively eliminates collision broadening so that laser techniques for Doppler-free spectroscopy can be fully exploited. The interface has been tested with a form of saturation spectroscopy. Spectral resolution of 50 MHz has been demonstrated for sodium.

Flame atomization with an analytical burner is commonly used in conjunction with various spectroscopic techniques, ranging from flame emission (1) and atomic absorption spectroscopy (2) to the more complicated laser-based methods. The latter include laser-induced fluorescence (3), laser-enhanced ionization (4), polarization saturation spectroscopy (5), and degenerate four-wave mixing (6). All of these techniques exhibit excellent sensitivity for trace elements in aqueous solution, typically in the part-per-billion range or less. The spectral resolution, however, is limited by the homogeneous broadening of the optical transitions due to collisions in the atmospheric pressure flame. Line widths of about 5 GHz are

0003-2700/87/0359-2203$01.50/0