An Integrated Method for Spectrum Extraction and Compound Identification from GC/MS Data

S. E. Stein

Mass Spectrometry Data Center

Physical and Chemical Properties Division

National Institute of Standards and Technology

Gaithersburg, MD 20899-8380

Abstract

A method is presented for extracting individual component spectra from GC/MS data files and then using these spectra to identify target compounds by matching spectra in a reference library. It extends a published “model peak” approach which uses selected ion chromatograms as models for component shape. On the basis of this shape, individual mass spectral peak abundance profiles are extracted to produce a “purified” spectrum. In the present work, ion-counting noise is explicitly treated and a number of characteristic features of GC/MS data are taken into account. This allows spectrum extraction to be reliably performed down to very low signal levels and for overlapping components. A spectrum match factor for compound identification is developed that incorporates a number of new corrections, some of which employ information derived from chromatographic behavior. Test results suggest that the ability of this system to identify compounds is comparable to that of conventional analysis.

Introduction

Gas chromatography/mass spectrometry, GC/MS, has long been the method of choice for identifying volatile compounds in complex mixtures. This method can fail, however, when acquired spectra are “contaminated” with extraneous mass spectral peaks, as commonly arise from co-eluting compounds, column bleed and ion-chamber contaminants. These extraneous peaks can pose a serious problem for automated target compound identification methods where they can cause identifications to be missed by reducing the spectrum match factor below some pre-set identification threshold. In addition, the presence of spurious peaks in a spectrum adds to the risk of making false identifications. Perhaps worst of all, this uncertainty in the origin of mass spectral peaks leads to a general loss of confidence in the reliability of making identifications by GC/MS, especially for trace components in complex mixtures, a key application area for this technique.

The most common method of extracting a “pure” spectrum for a chromatographic component is to subtract spectra in a selected “background” region of the chromatogram from spectra at the component maximum. This, however, is only appropriate when background signal levels are relatively constant (ionization chamber contamination, for example). Moreover, a complex chromatogram may not have a suitable background region.

A commonly employed approach for dealing with contaminated spectra is to assume that acquired mass spectral peaks that do not match the reference spectrum originate from impurities. While this method can suggest the presence of trace components embedded in complex background spectra, it can also produce false positive identifications for target compounds having simple spectra (i.e., when target compounds have spectra which are, in effect, embedded in the spectra of other compounds in the analyzed mixture).

This paper presents an integrated set of procedures for first extracting pure component spectra and related information from complex chromatograms and then using this information to determine whether the component can be identified as one of the compounds in a reference library. The practical goal is to reduce the effort involved in identifying compounds by GC/MS while maintaining the high level of reliability associated with traditional analysis. These methods were developed for a specific application, the automated identification of chemical weapons and related compounds, but they are expected to be applicable to any application requiring extraction of spectra from noisy chromatograms and the identification of target compounds by full spectrum matching.

Background

Since the inception of GC/MS, there has been a continuing interest in extracting “pure” component spectra from complex chromatograms. Biller and Biemann [1] devised a simple method in which the extracted spectrum is composed of all of the mass spectral peaks that maximize simultaneously. Colby [2] improved the resolution of this method by computing more precise ion maximization times. Herron, Donnelly and Sovocol [3] demonstrated the utility of Colby’s method in the analysis of environmental samples.

Another recently proposed, computationally facile approach extracts spectra by subtracting adjacent scans (“backfolding”) [4]. An advantage of this approach is that it does not explicitly require maximization. However, it does not account for ion counting noise or peak shape, so is unlikely to adequately identify weak components.

A more computationally intensive approach developed by Dromey et al [5], called the “model peak” method, extracts spectra for individual components from the underlying ion chromatograms based on the similarity of their shapes to a selected model ion chromatogram. As in the Biller/Biemann procedure, this method uses maxima in ion chromatograms to detect chromatographic components. However, to extract abundances, the shape of the most prominent of these maximizing ion chromatograms is taken as that of the actual chromatographic component. A simple least-squares procedure is used to extract individual mass spectral peaks. This method was successfully used for target compound identification in a large-scale EPA study [6]. Rosenthal [7] proposed an improvement to the peak perception logic for this method.

A number of matrix-based approaches have been proposed that make no assumptions concerning component peak shape. These methods generally process an abundance data matrix consisting of mass spectral peak/elution time pairs. Sets of ions whose abundances are correlated with one another are extracted. While diverse approaches have been described [8], to our knowledge none of them have been fully implemented and tested for general-purpose use. The inherent inability to make use of peak shape information is a drawback of this approach.

Method

The model peak method of Dromey et al. [5] was selected as the basis for spectrum extraction (deconvolution) both because it has been shown to produce reliable results in large-scale tests [6] and because it followed an approach similar to that of an analyst. However, its ability to extract weak signals was found to be poor. The origin of this problem was its inability to distinguish signal from noise at low signal levels. This problem was solved in the present work by explicitly considering signal-to-noise values throughout the analysis process. Another problem with the earlier approach was that all extracted peaks were treated the same – there was no way to deal with uncertain peaks.

In the present approach uncertain peaks were flagged and a spectrum match factor described earlier [9] was modified to deal with them. Analysis of test results led to a variety of further refinements in the computation of spectrum match factors.

The overall data analysis process involves four sequential steps: 1) noise analysis, 2) component perception, 3) spectrum deconvolution, 4) compound identification. The first step extracts signal characteristics from the data file for later use in noise processing and threshold setting. The second step perceives the individual chromatographic components and determines a model peak shape for each. The third step extracts “purified” spectra from the individual ion chromatograms using the model shape, explicitly subtracting nearby components when necessary. The final step computes match factors for the extracted spectrum and spectra in a reference library, using a variety of information acquired in the deconvolution step. These match factors are then sorted to produce a traditional “hit list”. Each of these steps is described in detail below.

1. Noise analysis

The first step in this analysis is to extract the following signal characteristics from the GC/MS data file:

(a) Noise Factor (N_f)

Event-counting detectors such as electron multipliers generate signals that fluctuate by an average amount proportional to the square root of the signal intensity [10]. Knowledge of this proportionality factor allows the simple estimation of the magnitude of this type of noise for any signal strength. In the present application this “noise factor” is defined as follows:

$$N_f = \text{average random deviation} / \text{signal}^{\mathbf{1/2}} \qquad (1)$$

In principle, N_f may be obtained from measured levels of random signal fluctuation during instrument tuning. However, this information is not generally available from instrument data systems. Therefore, N_f is derived for each data file from ion-chromatographic regions of relatively constant signal intensity. Non-trivial GC/MS data files invariably contain such regions.

An estimate of the noise factor is made as follows and illustrated in Figure 1. Each ion chromatogram, as well as the total ion chromatogram (TIC), is divided into segments of thirteen scans. If any abundance in a segment is zero, the segment is rejected. For each accepted segment, a mean abundance is computed and the number of times that this mean value is “crossed” within the segment is counted (crossings occur for adjacent mass spectral scans where one abundance is above the mean and other abundance is below the mean). If the number of crossings is less than one-half the number scans in the segment (6 or less), the segment is rejected. For each accepted segment, the *median* deviation from the mean abundance for that segment is found. This deviation is divided by the square root of the mean abundance for that segment to obtain a sample N_f value, which is then saved. After processing the entire data file, the *median* of these sample N_f values is taken as the characteristic N_f value for the entire GC/MS data file. The use of medians in place of means (simple averages) and the crossing criterion serve to reject high N_f values arising from real chromatographic components. In this paper the square root of a signal multiplied by N_f is the magnitude of this signal in "noise units". One noise unit represents the typical scan-to-scan variation arising from ion-counting noise at a given abundance level.

Testing with data files from properly tuned instruments showed that N_f was independent of both signal intensity level and m/z value and that run-to-run consistency for data files acquired on a single instrument was good (N_f variations of less than 10%). Over a wide range of well-tuned commercial mass spectrometers, including quadupole and ion trap instruments, N_f fell in the range 0.5 to 10. However, some dependence on signal strength was noticed at low signal levels in the presence of large amounts of spurious signal. Proper signal threshold setting eliminated this problem. No adverse effects attributable to the averaging of multiplier signals ("centroiding") were noted.

(b) Threshold Transitions

Mass spectrometer data systems typically store only signal intensities that are above a pre-set threshold abundance value, A_T, which is established during instrument tuning. Ion chromatographic regions with an average signal intensity near A_T appear visually as curves whose values suddenly drop to zero when the signal falls below A_T. These zero values prevent simple random statistics from being applied near the detection limit. Moreover, these sudden transitions from zero to non-zero abundance values, common for weak background signals, can be wrongly interpreted as chromatographic components. To avoid the problem, zero abundance values were replaced with estimated values as follows. First, the smallest non-zero ion abundance value in a chromatogram is assumed to be equal to A_T. Then, each ion chromatogram is divided into a fixed number of equal-length segments (10 are presently used). Next, for each m/z in each segment, the number of scans involved in transitions from zero to non-zero abundance values (threshold transitions) is counted and saved. Then, zero abundance values for a given m/z are coarsely estimated as the product of A_T and the square-root of the fraction of scans for that m/z that undergo threshold transitions in the segment. Use of this empirical correction greatly reduced the number of spurious components and mass spectral peaks in noisy analyses. To illustrate a typical correction, for a value of A_T

= 10, with half of the scans involved in threshold transitions, zero abundance values are replaced by 10 x $0.5^{1/2}$ = 7.

(c) m/z Peak Uniqueness

For each m/z value, the fraction of scans with non-zero abundance values is computed in each of the 1/10-th chromatogram segments used for threshold transitions. These values are used to measure the uniqueness of a m/z value. For each m/z value, signal-to-noise thresholds for signal rejection were multiplied by the square-root of the fraction of the scans containing a non-zero value. The key use of this value is to insure that unique m/z values were properly extracted even when they were present at very low signal levels.

2. Component Perception

Some GC/MS instruments, notably those with quadrupole and magnetic sector mass spectrometers, acquire spectra by scanning over a m/z range in a time period of the same order as the time for an individual component to elute. Different mass spectral peaks for a single component may therefore be acquired at distinctly different parts of the elution profile. Colby [2] demonstrated the importance of removing this “skewing” in order to distinguish closely overlapping components. “De-skewing” is done in the present approach by simple three-point quadratic interpolation, with the following three special cases: 1) abundance values in the first and last scans in a data file are not interpolated; 2) zero abundance values are not interpolated (they maintain their zero values); 3) non-zero interpolated values cannot be less than A_T.

Components are perceived when a sufficient magnitude of their ions maximize together, using the following procedure. First, for each ion maximum, the following steps are used to reject any such maxima originating from ion-counting noise. Significant computational effort is expended at this stage, since the false perception of a component could generate spurious results in later calculations. The magnitude of the signal associated with each maximizing ion is determined as follows (Figure 2):

The number of scans on each side of the component used for deconvolution (deconvolution window) is established by sequentially examining scans starting at the scan of maximization and proceeding in the forward and reverse directions up to a pre-set maximum number of scans (12 is the default). If a signal abundance is encountered that is more than five noise units greater than the smallest abundance between that scan and the starting scan (with noise units measured for the smallest abundance), then it is presumed that another component has been found and the window length is set to the preceding scan. Also, if the intensity falls below 5% of the maximum intensity, the window is fixed at that scan.

A tentative baseline is drawn though the lowest abundance on each side of the component maximum. This is adjusted as necessary to ensure that no abundance within these two end points falls below this baseline.

A least-squares baseline is computed using the smallest one-half of all abundance values where abundance values are measured from the baseline established in b).

If the height, in noise units, above this baseline is greater than a pre-set rejection threshold, the peak is marked as a possible component. A default rejection threshold value is 4 noise units was empirically derived. To illustrate, a peak with a maximum signal of 100 and N_f equal to 1.0 would be rejected if its height above baseline were less than $4 \times 1 \times 100^{1/2} = 40$.

This baseline definition was developed for robustness, rather than accuracy. Also, the window is often narrower than optimal for quantification. This narrow window is preferred for deconvolution because it reduces adverse effects of nearby components while providing all necessary shape information for spectrum extraction.

For each ion maximum passing the above test, a precise maximization time is computed by fitting a parabola to the maximum and its two adjacent scans (Figure 3). In addition, a measure of peak sharpness is computed for use in component detection. For this purpose, abundances are first time-shifted to move scans so that the central scan is positioned at the precise maximization time, as described by Dromey et al. [5].

Sharpness values between the maximum abundance, A_{max} , and an abundance value located *n* scans from the maximum, A_n , are defined as:

$$(A_{max} - A_n)/(n * N_f * A_{max}^{1/2}) \tag{2}$$

The maximum sharpness values on each side of the maximum scan are found and then averaged.

Average sharpness values are then used to identify individual components as follows. First, the time interval for each scan is divided into an array of ten sub-intervals (bins). Then, each sharpness value is added to the bin corresponding to its computed retention time, in the general manner recommended by Colby [2] (Figure 4a). After this is done, components are identified by their local maximization of bin values. Specifically, if a bin contains a value larger than all others within a computed range of uncertainty, then a component is associated with the retention time corresponding to that bin. This computed range provides a measure of the uncertainty in retention time arising from random ion counting fluctuations that increases as peaks become broader or less intense. Statistical testing showed that this computed maximum range was inversely proportional to the bin sharpness value and its two adjacent bins. A proportionality factor of 50 was found to be generally effective for estimating this range. To illustrate, if a sharpness value of a maximizing bin (and its two adjacent bins) is 10 (noise units per scan), the computed range would be 50/10 = 5 bins (0.5 scans). This means that if no bin within 5 bins of this maximizing bin has a greater sharpness value, this maximizing bin is assigned to a component. Component peak perception is illustrated in Figures 4b and 4c.

The model shape for each perceived component, used later for deconvolution, is taken as the sum of the individual ion chromatograms that maximize within the range of bins computed above and have sharpness values within 75% of the maximum value. In the original model peak approach by Dromey et al., only the largest ion chromatogram was used to represent component shape. Use of additional ions provides more accurate model shapes for weak components that do not have a single dominant ion.

Maxima in the total ion chromatogram (TIC) were used independently of ion chromatographic maxima for identifying components. This insures the perception of weak components showing a clear maximum in the TIC, but without intense individual ion-chromatogram maxima. This commonly occurred for trace components having many major ions (polychlorinated aromatics, for example). As a result, weak components were sometimes perceived only by TIC models, while stronger components were extracted using two different model shapes (once by its TIC and once by an ion chromatogram model). TIC processing employed the same threshold requirements as used for ion chromatograms.

3. Deconvolution

A spectrum for each component is derived from its model peak profile following the least-squares method described by Dromey et al [5]. Each ion chromatogram (m/z value) is individually fit to the model profile, allowing a linear baseline:

$$\mathrm{A}(n) = a + b * n + c * \mathrm{M}(n) \tag{3}$$

where A(n) is the abundance at scan n, a, b, and c are derived constants, and M(n) is the abundance of the model profile at scan n. For components perceived by TIC maxima, the TIC itself served as the model shape. The range of scans used here was the same as described above for component perception. The derived terms a and b describe the linear baseline and are not directly used for spectrum extraction.

The derived abundance for each m/z value is c * M($nmax$), where $nmax$ denotes the scan with the maximum model peak abundance. A(n) values equal to zero were replaced with estimated minimum values as described earlier (1b). This correction was important for eliminating spurious, low abundance mass spectral peaks common to noisy spectra.

As noted by Dromey et al., use of a single model peak was not always effective in removing extraneous signals from closely overlapping components. In such cases, signals from nearby components were explicitly subtracted using their own characteristic model peak profiles as follows,

$$A(n) = a + b * n + c * M(n) + d * Y(n) + e * Z(n) + \dots \qquad (4)$$

where Y(n), Z(n), … represent nearby model peak profiles and d, e, ... are their least-squares coefficients. This expression was also employed in the present method, using no more than two explicitly subtracted components.

Unfortunately, in some cases this adjacent spectrum subtraction method could fail. For instance, chromatographic irregularities could cause a single component to appear as multiple chromatographic peaks, which if subtracted from each other could cause the deletion of genuine mass spectral peaks. This could also happen for incompletely resolved isomers with similar spectra, a common occurrence in some analyses. In addition, this spectral subtraction process could, in effect, extract spurious mass spectral peaks from linear background signals. Therefore, spectra generated without adjacent component subtraction were always produced along with spectra generated with such adjacent component extraction. The benefit of insuring that a component was represented by at least one properly extracted spectrum was found to outweigh the increased risk of false positive identification resulting from additional spectra to compare with library spectra.

Regardless of the method employed for deconvolution, in complex chromatograms some mass spectral signals cannot be reliably assigned to an identified component. Moreover, large background mass spectral peaks could fully obscure signals from small components. To deal with such ambiguities, several rules were devised to find and flag uncertain peaks. These flagged peaks are treated differently than non-flagged peaks in the compound identification process described later.

Criteria for peak flagging and rejection are:

Fraction of signal contained in model envelope

> In complex chromatograms it is not uncommon for a component to be surrounded by too many overlapping components for it to be reliably extracted. In such cases, the least-squares methods described above might extract abundances from ion chromatograms for a target component that, by virtue of their different shapes, an analyst might judge to have originated from other components. The following method was developed to identify such mass spectral peaks.
>
> For each individual mass spectral peak extracted by the above method, the fraction of its total signal within the deconvolution window that did not match the model peak profile in the same window, $\mathbf{F_M}$, was computed,

$$\mathbf{F_M} = \mathbf{\Sigma}\ |\mathbf{I} - M| \qquad (5)$$

where **I** is the extracted signal intensity and **M** is the model intensity, both normalized to unity over the deconvolution window, i.e., $\Sigma\ \mathbf{I} = 1$ and $\Sigma\ M = 1$. Note that $\mathbf{F_M} = 0$ indicates a perfect match and $\mathbf{F_M} = 1$ indicates no overlap.

For strong signals, a value (mismatch) of $\mathbf{F_M}$ greater than 0.2 caused the peak to be flagged. That is, if more than 20% of these normalized signals did not overlap, the corresponding extracted mass spectral peak was suspect. Values with $\mathbf{F_M} > 0.6$ caused the peak to be rejected.

For weaker signals, these threshold values could be exceeded by normal statistical variation. To properly treat these variations, the following empirical quantity was added to the above $\mathbf{F_M}$ threshold value of 0.2:

$$20/[(\Sigma\Delta\ A^{\mathbf{1/2}} / N_f\ + 20] \qquad (6)$$

where ΔA is the absolute magnitude of the extracted abundances of signal **I** that did not match the model peak profile, **M**. The term $\Sigma\Delta\ A^{\mathbf{1/2}} / N_f$ measures the deviation of **I** from the model profile in terms of “noise units”.

Fraction of extracted abundance

When one or more overlapping components were explicitly subtracted, peaks in the extracted spectrum with abundances less than 10 percent of the total extracted value were flagged.

1) *Low S/N*

Extracted peaks with a signal-to-noise level less than 2 were flagged.

2) *Possible noise spike*

When a mass spectral peak at the component maximum was adjacent to scans with zero abundance, the peak was flagged when the peak occurrence probability was greater than 0.1.

Flagged peaks were treated as possible impurities - that is, they were used only if the corresponding peak was in the library spectrum. When they did match, their contribution to the spectrum similarity match factor was reduced by 10% (w=0.9 in equation 7).

In addition, when the computed noise level of a background peak was above the minimum detection level, A_T, the noise level for the peak was saved. This was used later to avoid overly penalizing library peaks that could not have been seen because they would have been within the noise level of the background signals.

4. Compound Identification

Traditional “library search” methods for compound identification find compounds in a reference library whose spectra most closely resemble the submitted (user) spectrum. The submitted spectrum commonly originates from a GC/MS data file, where it can be a single mass spectral scan or an average, with or without simple background subtraction. Each search produces a “hit list” of library spectra, which is ordered by similarity to the target spectrum according to a computed “match factor”. Ideally, this quantity should reflect the likelihood that the user and reference spectrum arose from the same compound.

While the elimination of spurious signals by the methods described above will clearly increase the reliability of library search results, a variety of modifications to the calculation of match factors were made to further improve reliability. Most of these modifications were made after examining results of large-scale tests described later. In this section, all abundances are presumed to be base-peak normalized.

Spectrum Similarity

The central factor in making an identification is, of course, the similarity of the library and user spectra. Two different measures of spectrum similarity are in common use. One assumes that the user spectrum originates from a single compound (pure spectrum match factor) and uses all peaks in both the library and user spectra for match factor computation. The other presumes that impurities may be present (impure spectrum match factor) and ignores peaks in a user spectrum that do not match corresponding peaks in a library spectrum. In the present application these pure:impure factors are combined linearly in a 70:30 ratio. The comparison function shown in equation 7 is the normalized dot product of the spectra being compared [9]:

$$100\frac{\left(\quad wm[A_u A_r]^{1/2}\right)^2}{A_u m \quad A_r m} \qquad (7)$$

Here A_u and A_r are the abundances of peaks in the user and reference mass spectra, respectively, and summations are over all m/z values (*m*) for the pure match factor or over only library m/z values for the impure match factor. A weighting term, w = 0.9, is employed for penalizing flagged (uncertain) peaks identified using criteria described in Section 3.

When a peak in the library spectrum could not have been observed because it was either below the detection threshold or within the noise level of a larger background peak, the penalty for not observing this library peak in the extracted spectrum was reduced. This was done by reducing the abundance of the non-matched library peak by a factor of two.

In the calculation of “impure” match factors, an adjustment was made when the abundance of the peak in the user spectrum was larger than the corresponding peak in the library spectrum. In this case, the peak abundance in the user spectrum was reduced by multiplying it by the ratio of the library peak abundance to the user peak abundance. This avoided an unduly large penalty when a small library peak was matched against a large peak in the user spectrum. Otherwise, the penalty for having such a large matching peak in the user spectrum could be greater than the penalty for there being no matching peak at all in the user spectrum.

Spectrum Complexity

A drawback of the simple dot product expression used for the match factor is that it tends to produce higher match factors for spectra with few major peaks than for spectra with multiple major peaks. This tends to produce a disproportionate number of false identifications for compounds with spectra having a single dominant peak. To reduce the severity of this problem, a scaling method for such spectra was devised that decreases the relative importance of only the larger peaks. In this method, each peak abundance value is multiplied by:

$$1/(1 + w\,\mathrm{A}) \tag{8}$$

where A is the observed abundance (assuming a base peak = 1) and w is a weighting factor designed to apply this correction only to spectra with a single dominant peak:

$$w = 1/(a + \Sigma\,\mathrm{A} - 1) \tag{9}$$

Here, Σ A is the sum of observed peak abundances and *a* is a selectable scaling factor. The weighting factor, *w*, ensures that only spectra with few dominant peaks (i.e., Σ A – 1 is small) will be appreciably scaled. In the most extreme case, with a spectrum containing only one prominent peak, setting *a* = 0.5 causes this peak to be diminished by a factor of three while having little effect on the small peaks. This value of *a* was selected as conservative level of scaling for the final version of the method.

Two more obvious scaling methods that increase the relative significance of smaller peaks, namely, logarithmic and fractional power scaling, are unsuitable for this purpose because they uniformly reduce relative peak abundances for peaks at all abundance levels. This leads to the overemphasis of trace impurity peaks in match factor computations.

Additional penalties were applied to match factors for extracted spectra having small numbers of peaks. Such spectra usually arose from components with signal strengths just above the detection limit. Depending on the number of (non-flagged) peaks in the component spectrum, match factors were multiplied by the following empirical values: 0.75 (1 peak); 0.88 (2 peaks); 0.94 (3 peaks); 0.97 (4 peaks).

Other Corrections

Adjacent Peak Deconvolution: For each explicitly subtracted overlapping component, a penalty of 2 units (100 = perfect match) was subtracted.

Component Purity: The uncertainty in identifying components whose signals represented a small fraction of the total signal in the central scan (purity), was dealt with by adding the following modest correction to the match factor;

$$1.0 \log_{10}(\text{purity}) + 0.6 \tag{10}$$

Detection Threshold: To account for the loss of confidence associated with the inability to measure peaks below the detection threshold, A_T, the match factor was multiplied by the following factor (threshold is relative to a base peak of unity):

$$(1 - \text{threshold})^{\mathbf{0.3}} \tag{11}$$

Results and Discussion

Method Development:

As outlined in the Method section, the development of the present method began with the implementation of the “model peak” approximation of Dromey et al. [5] for spectrum deconvolution along with a “dot product”-based match factor [9] for compound identification. Further development was guided by examining false positive and false negative results that might not have been made in a conventional analysis. The underlying reason for each failure was sought and appropriate improvements were made, leading ultimately to the set of procedures presented here. The overall goal was to achieve a level of performance for identifying compounds similar to that of a chemist with no prior knowledge of sample composition or retention time.

Reference Spectra:

Because the reliable identification of chemical weapons and related compounds was a primary goal of this project, these compounds provided the principal reference spectra used for algorithm development and testing. They represent a wide range of spectra and were especially suitable for false positive testing since they should not be present in the environmental samples making up the bulk of the data files used for testing (see below). Most of these reference spectra were the same or equivalent to those in the NIST/EPA/NIH Mass Spectral Library [11].

False Positive Testing:

Because false positive identifications are matrix dependent and often rare, a sizable collection of data files is needed for effective testing. For this purpose a collection of 43,006 data files was amassed, most of which were from environmental analysis following EPA protocols. About half were from waste-water analysis [6]. For all identifications with match factors above 80, the data file was examined to determine whether a human evaluator would have also concluded that it was a sufficiently good match to support identification. This process led to most of the corrections and setting of parameters employed in the present method.

The potential problem that the spectrum extraction processes, and in particular the explicit subtraction of nearby component spectra, might somehow synthesize spectra that matched target compounds did not occur for any of the spectra in the test library of several hundred compounds.

Results of false positive testing for selected compounds that are not expected to be present in the analyzed samples are given in Table 1. The first five compounds are among the most commonly cited chemical weapons related compounds and each has a unique spectrum (as determined by the lack of similar spectra in comprehensive libraries).

Of these compounds, pinacolyl alcohol (3,3-dimethyl-2-butanol), had the least unique spectrum, having some major peaks in common with other, more common aliphatic alcohols and matching about a dozen of them in the NIST/EPA/NIH library [11] with match factors in the range 70 to 75. The extracted spectra that produced the highest match factors for pinacolyl alcohol were separately searched against this library, and pinacolyl alcohol was clearly identified as the best matching compound with a match factors as high as 92. Examination of the data files indicated that the identifications with match factors above 85 were probably correct (pinacolyl alcohol was present in the sample) and most below 80 were probably incorrect.

In general, the degree to which a low match factor indicates that the identification is false depended on signal intensity. Strong signals with low match factors are generally false positives that arise from structurally related, but different compounds. For weak signals, especially in cases where significant peaks are near the detection limit or significant noise is present, correct identifications are likely to involve lower match factors.

False Negative Testing:

Data files from a series of analyses of commercially-available contaminated soil samples which had been spiked with 10 parts-per-million of selected target compounds were analyzed both by a conventional method (manual background subtraction followed by library searching) and by the present method. Results, shown In Table 2, compare match factors from these methods. Using 80 as the identification threshold, out of 80 possible identifications, the present system reported 45 identifications, compared to 34 identifications by the conventional approach. With 60 as the identification threshold, the corresponding numbers are 52 and 38, respectively. In 6 cases, neither approach made an identification. In two cases, only the manual method identified the target compound, but with very low match factors (14 and 40). In 7 cases, the present method provided an identification that was missed by conventional analysis. The most significant failure of the conventional method was for sarin in the TCLP/Pesticides matrix. In this case the signal for sarin at its maximum was less than 0.1% of the total signal (it was submerged beneath an overloaded, co-eluting peak of trimethyl phosphate). In this case conventional background subtraction was unable to remove enough of the overlapping peaks to permit a library-search identification.

Common Compounds:

Numbers of identifications of common compounds expected to be present in the many of the samples analyzed are shown in Table 3. These results provide a general view of distributions of match factors expected in practical analyses. Inspection of results suggested that identifications above 80 are reliable, 70 to 79 are often correct and 60 to 69 are very uncertain. Note that numbers of identifications generally decline by a factor of 2 to 3 as the match factor drops from the 90s to the 80s, and then another factor of 2 to 3 for a drop from the 80s to the 70s. The generally smaller decline from the 70s to 60s arises from the increasing number of false positive identifications at the lower match factors.

Differences in relative numbers of correct identifications with high versus low match factors depend on spectral uniqueness. Anthracene-d_{10}, which has the most unique spectrum of those examined, shows 20 times fewer identifications with match factors in the 60s than the 90s, while this ratio is near 5 for the other compounds. Consistent with this idea, inspection of results showed that even in the 60s, a large majority of anthracene-d_{10} identifications were probably correct, while the majority of identifications for less unique toluene-d_8 were probably incorrect.

Deconvolution Tests:

The ability of the present method to resolve overlapping components is demonstrated for two cases. One involved two pairs of compounds whose retention time differences were comparable to the time required for a mass spectral scan. Results, shown in Table 4, demonstrate the resolving ability of the algorithms over a range of relative concentrations of the overlapping compounds.

Another test case, examined also by Colby [2], contains what appears to be a single TIC peak with a width at half height of five scans. This TIC peaks was actually composed of three components, each with a width of approximately four scans with less than one scan separating each (Figure 5). The present method correctly identified these components: dibromochloromethane (match factor=91, 19.586 min.), 1,3-(or 1,2-) dichloropropene (match factor=87, 19.607 min), and 1,1,2-trichloroethane (match factor=97, 19.653 min.). As noted by Colby, de-skewing was essential for the successful deconvolution of these components.

Comparison to Other Methods:

Commonly available spectrum extraction methods perform simple background subtraction based on the TIC profile. This method that cannot separate closely co-eluting components or identify trace compounds showing no maximum in the TIC. Ion chromatograms, used for peak perception in the present approach, commonly show components not evident in the TIC. Moreover, the presence of components in the selected “background” region can lead to deletion of valid peaks in the target compound. The method for peak deconvolution developed by Colby [2] and tested by Donnelly et al. [3] avoids these problems, is easy to implement and is very effective in separating spectra of closely eluting components having strong signals. It cannot, however, reliably extract abundances for ions common to a pair of closely overlapping components. This method also extracts the entire abundance of a mass spectral peak – this may not be appropriate for ions with a significant non-zero baseline. The use of a model shape by the present method to extract abundances minimizes these problems. However, the biggest advantages of the present approach for deconvolution over others (including Dromey’s [5] original model peak method and the “backfolding” method [4]) stem from its use of the noise factor to allow signal to be distinguished from ion counting “noise”. This permits the extraction of spectra for trace components without generating of large number of spurious components and also provides an objective means of identifying the maximizing ions associated with a single component.

Limitations:

The use of peak maximization as the only means for perceiving components can cause problems. If, for instance, two components maximize at precisely the same time, even if they have different shapes, the present approach will report just one component and extract a single spectrum. Also, if peak tops are broad and several local maxima are present, a component may be identified more than once. Moreover, if a component is very broad it may be missed entirely. These problems can be reduced in severity by using reverse matching logic (ignoring mass spectral peaks not in the library spectrum), but this would also increase false positive risks significantly.

Another drawback of the present approach is the requirement that a simple yes/no decision be made concerning the existence of a component. In complex chromatograms, the presence of some components will be uncertain.

Also, because of the different models that may be employed for a single component, the present approach may generate more than one spectrum per component. While this works well for target identification, where only the best matching spectrum for a library compound is reported, for non-identified components, an analyst may have to decide which is best among several extracted spectra. A means of ranking the relative reliability of the different extracted spectra for a single component is under development.

Applications:

The present method has been developed and tested specifically for the identification of chemical weapons and related substances in matrices of arbitrary complexity. It has been accepted for automated compound identification by the Organization for the Prohibition of Chemical Weapons in The Hague, the Netherlands. It is, however, expected that it will find use for other applications, particularly those where a substantial number of target compounds need to be monitored down to the limits of detection in matrices of arbitrary complexity. Such an application has recently been reported for the identification of urinary acids for disease diagnosis [12].

Software:

All algorithms described here have been incorporated into a Microsoft Windows program called AMDIS (automated mass spectral deconvolution and identification system) which has been recently reviewed [13] and is available free-of-charge from NIST [11]. On a 200 MHz personal computer, with a 200 compound library, analysis of a 30 min. GC/MS data file generally takes between 10 s and 5 min. depending on sample complexity.

Conclusions

The method described here is capable of automatically extracting pure component mass spectra from highly complex GC/MS data files and then using these spectra for identifying compounds in a reference library. This was built on earlier methods for spectrum deconvolution and library searching with the addition of a variety of factors to account for noise and other features of GC-MS data. Parameter optimization and testing involved the analysis of a very large set of data files. For identifications based solely on mass spectral information, comparisons to results of manual analysis suggest that the overall false positive and false negative performance of this method is comparable to that of an analyst.

Acknowledgment and Disclaimer

This work was supported by the Defense Threat Reduction Agency, U.S Department of Defense (DSWA-IACRO-98-3029). Gary Mallard (NIST) provided valuable assistance in the discussion of algorithms and test methods. Jane Klassen and Edward White V provided careful reviews of this paper. Several organizations provided important GC/MS test data including, the Illinois Institute of Technology Research Institute (Guggilam Srestry), Battelle Memorial Laboratories (Laura Hernon-Kenney), Edgewood Arsenal/CBDCOM (Dennis Reutter, Tom Rosso), DynCorp/EPA (Jim King) and the Finnigan Corporation (Peter D'Arcy). All references to commercial products are given for information purposes only and are not official endorsements.

References

[1] “Reconstructed Mass Spectra, A Novel Approach for the Utilization of Gas Chromatograph—Mass Spectrometer Data”, Biller, J.E.; Biemann, K. *Anal. Lett.* **1974** *7* 515-528.

[2] “Spectral Deconvolution for Overlapping GC/MS Components” Colby, B. N. *J. Am. Soc. Mass Spectrom*. **1992** *3* 558-562.

[3] “Software-Based Mass Spectral Enhancement to Remove Interferences from Spectra of Unknowns”, Herron, N.R.; Donnelly, J.R.; Sovocool, G.W. *J. Am. Soc. Mass Spectrom.* **1996** *7* 598-604.

[4] “Automated Extraction of Pure Mass Spectra from Gas Chromatographic / Mass Spectrometric Data”, Pool, W.G.; Leeuw, J.W.; van de Graaf, B *J. Mass Spectrom.* **1997** *32* 438-443.

[5] “Extraction of Mass Spectra Free of Background and Neighboring Component Contributions from Gas Chromatography/Mass Spectrometry Data” Dromey, R.G; Stefik, M.J.; Rindfleisch, T.C; Duffield, A.M *Anal. Chem.* **1976**, 48 (9), 1368-1375.

[6] “An Evaluation of Automated Spectrum Matching for Survey Identification of Wastewater Components by Gas Chromatography-Mass Spectrometry” Shackelford, W.M; Cline, D.M.; Faas, L; Kurth, G *Analytica Chim. Acta* **1983** *146*, 25-27.

[7] “Improvement of Algorithm for Peak Detection in Automatic Gas Chromatography-Mass Spectrometry Data Processing” Hargrove, W.F.; Rosenthal, D.; Cooley, P.C. *Anal. Chem.* **1981**, *53,* 538-539.

[8] “Data Analysis for hyphenated techniques”, Karjalainen, E.J; Karjalainen, U.P. Elsevier, Amsterdam, 1996.

[9] “Optimization and Testing of Mass Spectral Library Search Algorithms for Compound Identification” Stein, S.E. Scott, D.R. *J. Am. Soc. Mass Spectrom.* **1994** *5*, 859-866.

[10] “Signal-to-Noise Ratios in Mass Spectroscopic Ion-Current-Measurement Systems”, Peterson, D.W.; Hayes, J.M. in “Contemporary Topics in Analytical and Clinical Chemistry”, Vol. 3, Plenum Publishing, **1978** 217-251.

[11] Standard Reference Data Program, National Institute of Standards and Technology, Gaithersburg, MD. Standard Reference Database IA. Internet address: http://www.nist.gov/srd/nist1a.htm

[12] “Deconvolution GC/MS of Urinary Organic Acids – Potential for Pattern Recognition for Automated Identification of Metabolic Disorders”, Halket, J.M; Przyborowska, A; Stein, S.E.; Mallard, W.G; Down, S.; Chalmers, R.A., *Rapid Commun. Mass Spec.*, **1999** *13*, 279-284.

[13] “The New Automated Mass Spectrometry Deconvolution and Identification System (AMDIS)”, Davies, T. *Spectroscopy, Europe*, **1998**, 10/3, 24-27.

Tables

Table 1. False Positive Results [a]

	Match Factors			
	60-69	70-79	80-89	90-100
Tabun [b]	2	0	0	0
VX [c]	2	0	0	0
Soman [d]	96	4	0	0
Mustard [e]	111	9	0	0
Sarin [f]	181	63	0	0
MPA-TMS [g]	244	38	8	0
Pinacolyl Alcohol [h]	5513	2776	141 [i]	2 [i]

Numbers of false identifications by the present method for each compound within the specified range of match factors. None of these analytes were expected to be in any of the samples.

[a] Results of searching 43,006 GC/MS data files, all for EPA analysis.

[b] O-ethyl-N,N-dimethyl phosphoroamidocyanidate

[c] O-ethyl-S-2-diisopropylaminoethyl methyl phosphonothiolate

[d] O-pinacolyl methyol phosopnonfluoridate

[e] Bis-(2-chloroethyl)sulfide

[f] O-isopropyl methylphosphonofluoridate

[g] bis(trimethylsilyl)methylphosphonate (methylphosphonic acid – trimethyl silyl derivative)

[h] 3,3-Dimethylbutane-2-ol

[i] some of these may correct identifications (see text)

Table 2: Match factors for the identification of compounds added to contaminated soil samples from the present method and conventional manual analysis.

Sarin	Soman	Tabun	Mustard
-------------------------BNA-Pesticide-----------------------			
98 83	98 90	95 97	90 91
98 90	97 90	92 95	97 98
99 83	97 78	97 94	93 93
-----------------------------TPH------------------------------			
98 90	96 64	85 64	80 91
98 90	94 56	87 68	80 86
99 90	88 50	77 81	85 52
-----------------------TCLP-Pesticides----------------------			
89 NI	74 NI	61 NI	75 NI
67 NI	NI NI	NI NI	NI NI
88 NI	77 NI	NI NI	NI NI

MPA	IMPA	EMPA	IMPAE
----------------------BNA-Pesticide-------------------------			
89 86	97 98	96 98	92 95
81 47	97 97	96 98	94 98
91 86	97 97	96 98	93 98
-----------------------------TPH------------------------------			
64 10	93 91	94 97	NI NI
NI 14	96 98	94 98	91 98
83 49	91 64	93 94	NI 40

Each sample was prepared and analyzed in three separate analyses. The first of each pair of values was obtained by the present method, the second is from manual analysis (manual background subtraction followed by a PBM library search using HP ChemStation software). Maximum match factors are 100, NI: not identified.

Matrices: Commercially available contaminated soils for EPA analysis: BNA/Pesticide: benzene/naphthalene/anthracene complex mixture; TPH: total petroleum hydrocarbons; TCLP/Pesticide: contaminated with complex mixture of pesticides and phosphates. Each sample spiked with 10µg/g of target compounds. Contaminated soil samples were obtained as Certified Reference Materials from Resource Technology Company, Laramie, WY.

Chemicals: See Table 1 for sarin, soman, tabun, and mustard. Others were detected as TMS derivatives of the following acids after added to the soils: MPA = methylphosphonic acid; IMPA= isopropyl methylphosphonic acid; EMPA = ethyl methylphosphonic acid; IMPAE = di(2-isopropylamino)ethylphosphonic acid.

Table 3. Distribution of Match Factors for Identification of Common Compounds in 43,006 GC/MS Data Files.

		Match Factor Range		
	60-69	70-79	80-89	90-100
Benzene	1491	1663	2517	4819
Toluene	3028	2873	4350	12168
Naphthalene	792	973	1763	3986
Methylene Chloride	3203	4249	6853	18441
Anthracene-d_{10}[a]	676	1473	4147	14508
Toluene-d_8[a]	3028	2870	4353	12168

Shown are the numbers of different data files in which a compound was identified within the specified ranges of match factors. There is no reliable record indicating the actual number of samples containing these compounds.

[a] internal standards

Table 4. Match Factors for Deconvolution of Overlapping Components

	Concentration ratios				
	3/1	1/1	1/3	1/10	1/20
C7-sarin/ dichlorvos[a]	92/74	93/94	92/95	89/97	78/98
Bis(2-chloroethylether/ Malathion[b]	93/92	90/95	87/96	81/98	73/98

Scan time = 1.0 s, peaks widths at half height were about 4 scans.

[a] difference in retention time = 0.5 s

C7-sarin = 2-methylcyclohexyl methylphosphonofluoridate

dichlorvos = 2,2-dichlorovinyl dimethyl phosphate

[b] difference in retention time = 1.0 s

Figures

Figure 1.

Illustration of the determination of the noise factor (N_f) from 13-scan ion chromatogram segments. The upper chromatogram is rejected because it has fewer than seven "crossings" of the mean. The lower ion chromatogram crosses the mean eight times, so provides a sample noise factor. The *median* distance from the mean (seventh closest to the mean) is used to generate a sample noise factor, N_f. The final N_f for the analysis is taken as the median of all sample values.

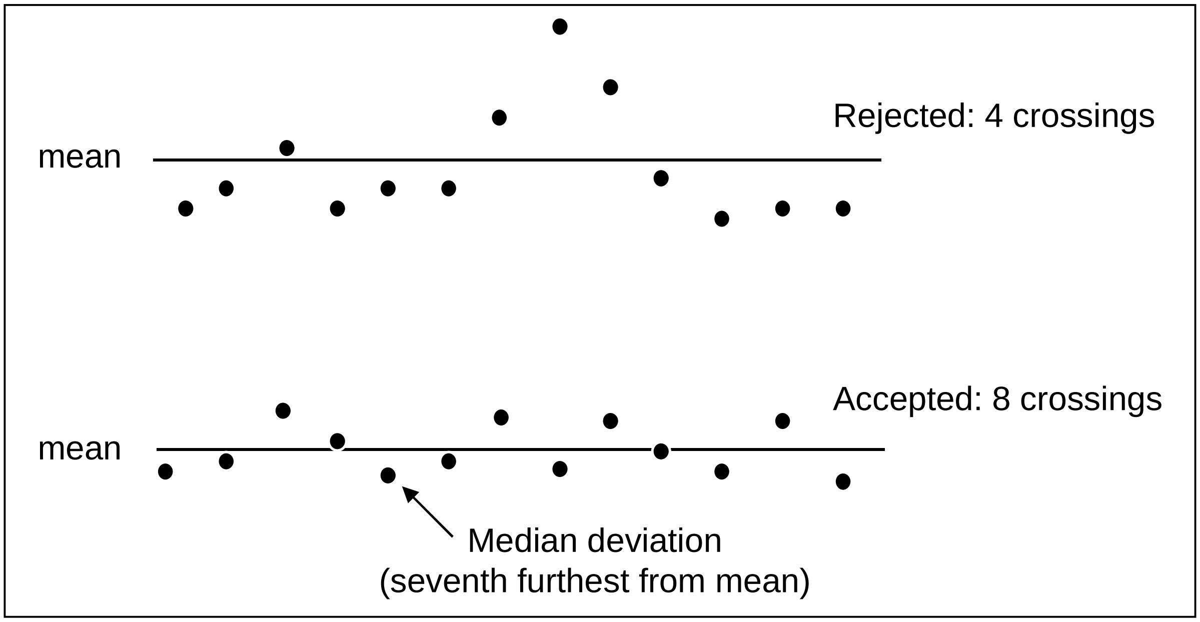
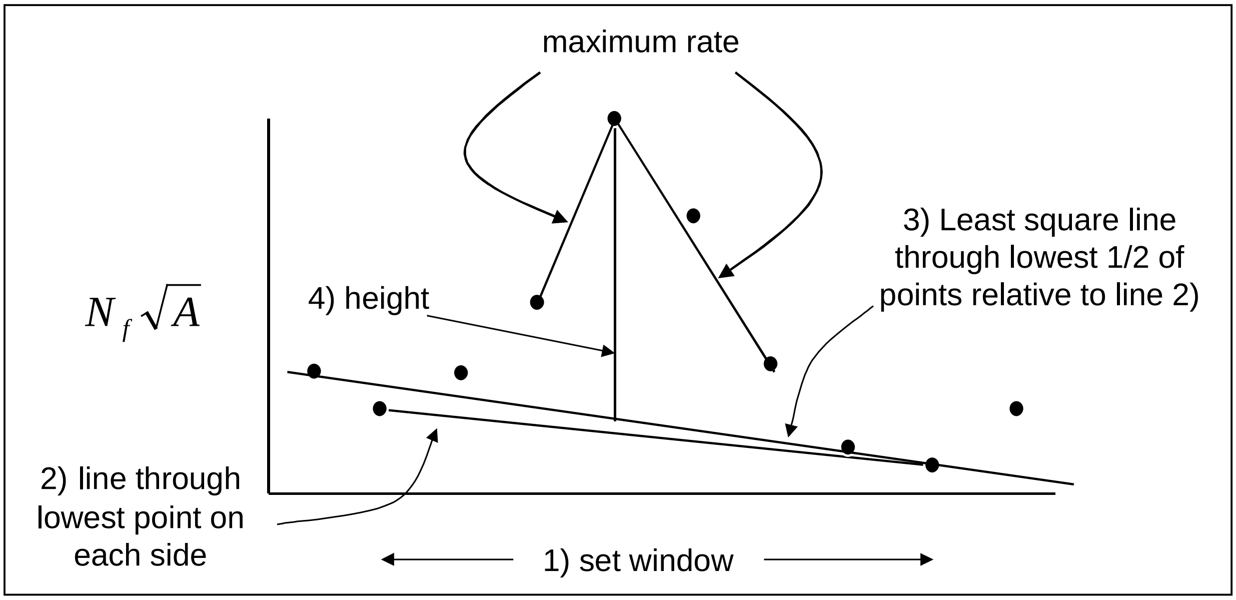


Figure 2.

Four steps for determining whether an ion chromatogram peak is large enough to be used for peak perception. 1) a scan window is set using minima on each side of the peak; 2) a tentative baseline is drawn between the lowest points on each side (readjusted if a point between these end points falls below the line); 3) a least-squares line is drawn using the lowest one-half of points as measured from the baseline in step 2; 4) signal height between the maximum and least squares line is computed. Peaks must have heights larger than 4 noise units ($N_f A^{1/2}$) for use in peak perception (A is the absolute abundance at the peak maximum).



maximum rate
$N_f \sqrt{A}$
4) height
3) Least square line through lowest 1/2 of points relative to line 2)
2) line through lowest point on each side
1) set window

Figure 3.

Time shifting of scans prior to sharpness calculation [5]. The maximum and its adjacent scans are fit to a parabola to find precise retention times (RT). The chromatogram is then time shifted to center the scans at this computed retention time. Sharpness values are the maximum rate of decline in abundance between the central scan and scans on either side.

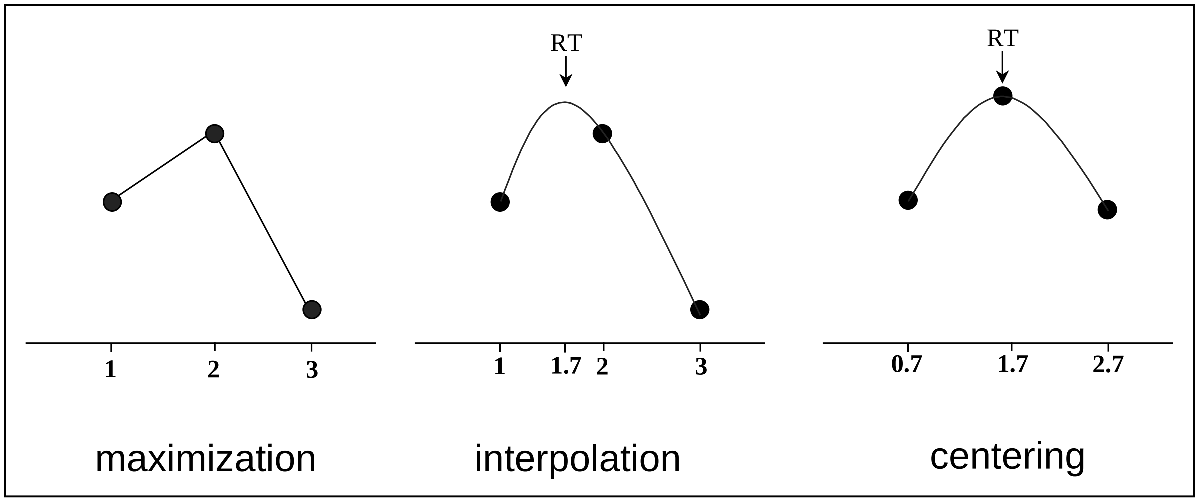


Figure 4a.

Identifying components. Each scan is divided into ten bins (0.0 to 0.9). The sharpness value for each perceived peak is placed in the bin corresponding to the maximization time for that peak (Figure 3). Values for two peaks are given, one with a retention time of 2.4 scans and an average sharpness value of 10 noise units per scan, the other with corresponding values of 2.5 scans and 6 noise units per scan. Sharpness values are averages of two maximum rates of decline (in noise units) from the maximum to points on each side.

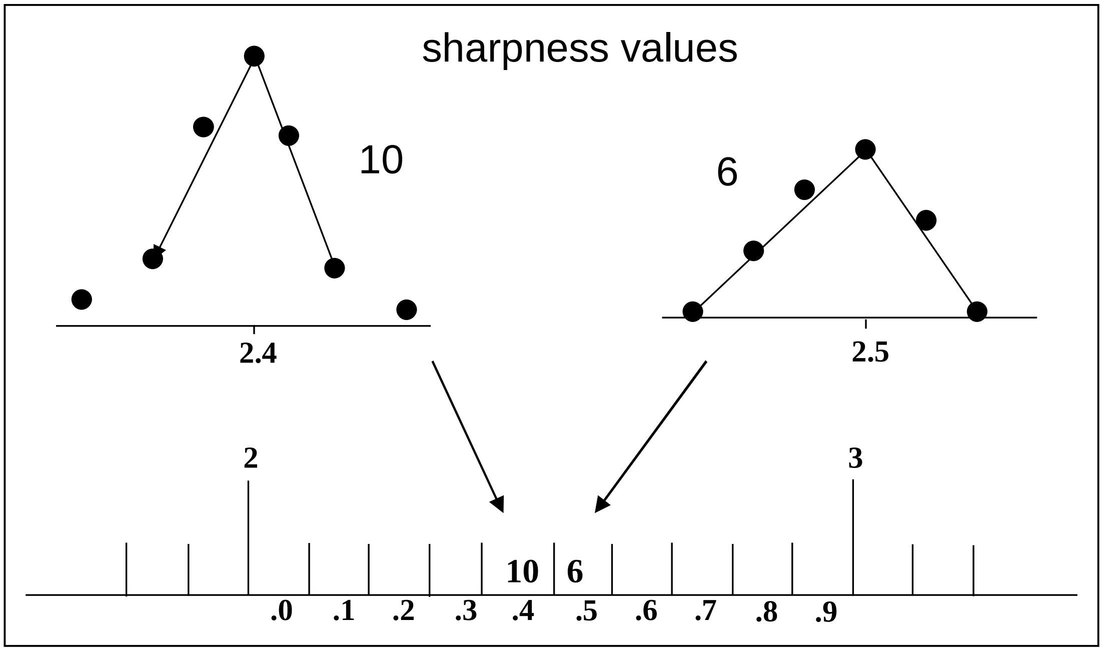


Figure 4b.

An illustration of a set of bins filled with sharpness values for 13 different ion chromatograms maximizing at retention times in the vicinity of 2 to 3 scans. The maximum at 2.4 scans and its adjacent bins contains 26 noise units per scan, corresponding to a range of maximization uncertainty of 10/26 = 0.4 scans (4 bins). Since no bin within 4 bins of the central bin contains a larger value, a component is identified at 2.4 scans.

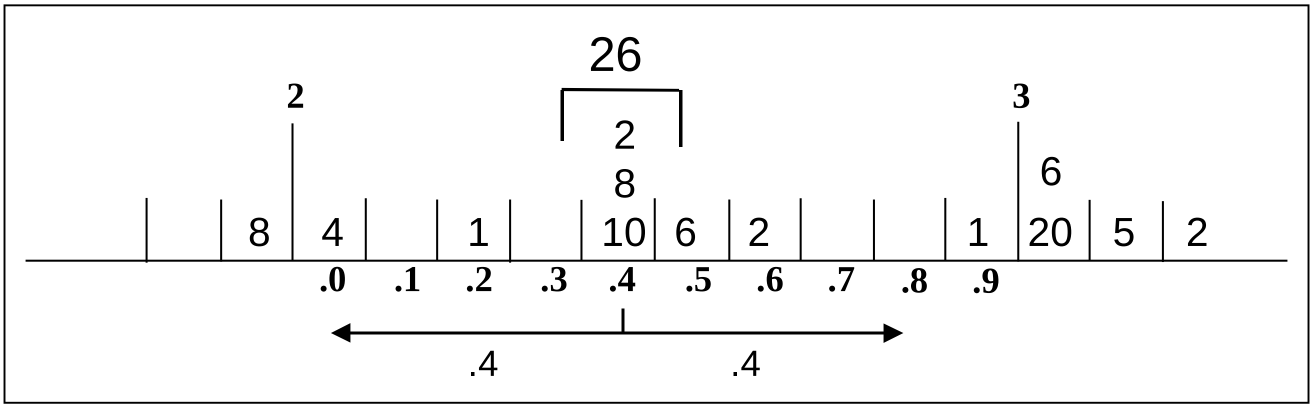


Figure 4c

Two other local maxima in Figure 4b are examined. One, at 1.9 scans has a range of uncertainty of 10/12 = 0.8 scans. Since a larger maximum in this range occurs at 2.4 scans, this maximum is discarded. Another maximum at 3.0 scans has an uncertainty range of 10/32 = 0.3. Since no larger maxima occur within 3 bins, it is marked as a separate component.

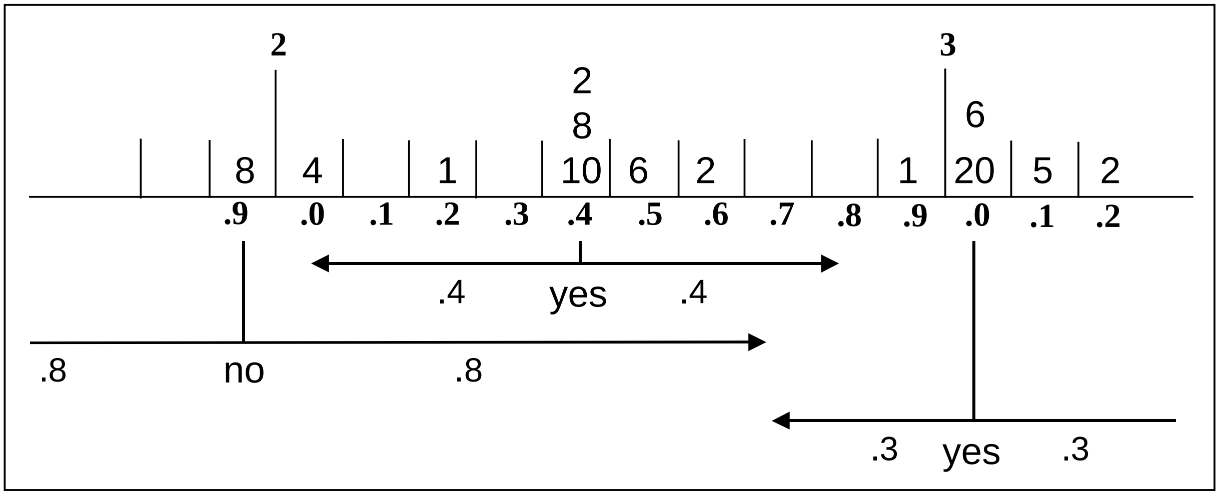


Figure 5.

Example of three overlapping components identified by present method. Numbers correspond to the most prominent model m/z peak for each component; arrows correspond to component maxima; scans are filled circles (lines are for clarity only). Dibromochloromethane (m/z = 129, left arrow), 1,3(or 1,2)-dichloropropene (m/z = 75, middle arrow), 1,1,2-trichloroethane (m/z=83, right arrow).