Selective Data Reduction in Gas Chromatography/Infrared Spectrometry

Dongjin Pyo* and Hyundu Shin

Department of Chemistry, Kangwon National University, Chuncheon 200-701, Korea
Received November 28, 2000

As gas chromatography/infrared spectrometry (GC/IR) becomes routinely avaliable, methods must be developed to deal with the large amount of data produced. We demonstrate computer methods that quickly search through a large data file, locating those spectra that display a spectral feature of interest. Based on a modified library search routine, these selective data reduction methods retrieve all or nearly all of the compounds of interest, while rejecting the vast majority of unrelated compounds. To overcome the shifting problem of IR spectra, a search method of moving the average pattern was designed. In this moving pattern search, the average pattern of a particular functional group was not held stationary, but was allowed to be moved a little bit right and left.

Keywords : Selective data reduction, Infrared data, Pattern search, Fingerprint region.

Introduction

One of great challenges of modern analytical chemistry is to try to make sense of data as fast as instruments can spew it forth. To that end, we have been interested for some time in developing computerized methods of selective data reduction. The goal of this method is the rapid sorting of hundreds of spectra to find the relative few which may be important enough in a given analysis to warrant further attention. For example, if a complex mixture of unknown compounds is to be analyzed, the sample might be chromatographed and the GC effluent be directed into a continuously scanning spectrometer. After 30 minutes of analysis, 900 spectra might have been collected and stored in a data file. Let us assume, in this particular case, that the analyst is interested in finding and identifying all of the chlorinated compounds in the mixture. If the GC effluent had been detected with a chlorine-selective detector, such as the electron-capture detector (ECD), the chlorinated compounds would have been easy to locate; but the ECD contributes little structural information for the identification of unknowns. In the present case, the spectrometer provides full spectral data, useful for identification, assuming the spectra of the chlorinated compounds can be located. Each of the 900 spectra could be visually examined, searching for spectral indications of the presence of chlorine; a tedious and time consuming approach. The data system could perform a library search on each of the 900 spectra-again, time consuming and inefficient. If the data system was sufficiently sophisticated, it could perform some sort of artificial intelligence of automated interpretation scheme as described above on each of the 900 spectra, but the time involved would surely be prohibitive. The sensible approach is to somehow select some of the 900 spectra for further attention, but to select them in a way that maximizes the probability of selecting chlorinated compounds. All of the usual identification routines-manual interpretation, library search, or automated interpretation-could be applied to this smaller subset. This kind of data reduction technique is very valuable in GC/IR analysis since, in most cases, only a small fraction of the data is of interest in any particular analysis. A computer can be employed to selectively reduce the volume of data, thereby improving the efficiency of the analysis. This selection is the process we refer to as selective data reduction.[1]

Selective data reduction is any data processing method that somehow selects certain data as being more "important" or more "interesting" to the analyst. The selection of important data can be based on a number of criteria. The selection criteria often used are the total intensity of the spectra, or the presence of some spectral feature of interest. Commercial spectrometers frequently have simple routines built into the data systems: peak finding routines are just selective data reduction using spectral intensity as a criterion. Routines based on spectral features vary with the type of spectrometer used. In GC-MS, one has mass chromatograms[2] or the presence of isotope clusters[3,4] to use as indicators of compound class. In the example above, the presence of chlorine isotope clusters can be used to indicate chlorinated compounds.[4]

In GC/IR, "chemigrams"[5] are used to indicate the presence of absorptions of interest. There are plots of integrated absorbance in a defined spectral window. Rather than looking at 900 spectra, with the use of chemigrams, the analyst sorts out only those that are likely to contain a particular functional group. Although useful, chemigrams are not always very selective, in that they show only the integrated absorbance over a chosen frequency window.

In this work we show that the use of patterns of absorbances provides a much more selective criterion for data reduction. Computer algorithms have been written to search through hundreds of spectra, retrieving only those that display the pattern of interest, and these algorithms have great potential for the analysis of GC/IR data. Although our routines are based on the presence of spectral patterns, they are distinct from pattern recognition methods in both purpose and approach. Pattern recognition statistically sorts a large database into a number of clusters, and assigns a spectrum to

a compound class based on the nearness of some metric representing the spectrum to one of the clustered units. Our approach seeks only to reduce the number of spectra which must be further interpreted by the analyst, and so looks only for similarity within a defined spectral window. The database is not really required, and the fact, one need not know in advance what functional group is responsible for the pattern of interest.

We have also developed computer algorithms to evaluate the selectivity of chemigram method, *i.e.* how selectively chemigram algorithms pick up a specific functional group by monitoring the absorbances over specified IR spectral region. Chemigrams have been used for specific functional group detection since their development in 1979,[5] just as specific ion detection is used in GC/MS. Even though selectivity is vital in chemigram-type approach, it has never been evaluated statistically. The algorithms we developed can give good evaluations for different functional groups, and at different threshold values for each functional groups. To find an optimum threshold value is very important because it affects the selectivity of chemigram, *i.e.*, how many members of a particular compound class are recovered and how many members of other functional group are eliminated. Particularly, when the chemigram is used for quantitative analysis of trace components, its selectivity increases the sensitivity of the analysis of trace components, by being insensitive to interfering spectral contributions.

Our algorithms have been compared with chemigram algorithms in all mid-IR region (4000-400 cm^{-1}). A greater degree of selectivity was observed than with chemigram algorithms, especially in O-H stretching and carbonyl stretching regions. For example, carboxylic acids have O-H stretching and carbonyl stretching regions. One hundred eighty five spectra of carboxylic acids were averaged in the 3800-3400 cm^{-1} window. The resulting pattern showed a single sharp band centering around 3560 cm^{-1}. The same 100 spectra from the database were again considered: this time, M_{SQ} for each was calculated - a measure of similarity of the spectrum to the average pattern for carboxylic acids. The results were striking. Four carboxylic acids (100%) had the lowest M_{SQ} values: 3-chlorobutyric acid, 0.14; butyric acid, 0.26; heptanoic acid, 0.30; isobutyric acid, 0.40. The next lowest M_{SQ} was for 2-bromo-*p*-cresol with an M_{SQ} = 1.70. Not only were the acids located as best matching the average pattern, but there was a large distance between the worst acid (M_{SQ} = 0.40) and the next closest nonacid (M_{SQ} = 1.70). When the experiment was repeated on the entire database, more than 92% of the carboxylic acid spectra had M_{SQ} values of 1.5 or less; fewer than 4% of the non-carboxylic acids had M_{SQ} of 1.5 or less.

Experimental Section

The computer programs described were written in FORTRAN and run on an IBM 370 computer. The database chosen was the EPA Vapor Phase Collection of 3300 spectra, available from Dr. James de Haseth at the University of Georgia. The spectra were stored in digital format at 1600 bpi on nine track magnetic tapes. The format of the vapor phase tapes is the following: each file represents one spectral entry. There is an end of file mark between each file and there is an end of file mark at the beginning of each tape. There are thirteen records in each file. Record 1 is the header record and that is 1148 bytes in length. Records 2 through 5 contain the reference interferogram of 2058 bytes in each. Record 10 is the first part of the spectrum, with 2058 bytes. Record 11 is the second part of the spectrum, with 1646 bytes. Records 12 and 13 contain the inverse Fourier transform of the spectrum. In records 11, 12 the first two bytes contain an integral serial number corresponding to the serial number in the header record. Each spectrum is divided up into record of 1024 data points (20480 bytes). The ordinates are expressed to 0.002 absorbance units from 0.000 to 1.998. In order to save space, the absorbance units are expressed as integers (0 to 1998) by multiplying each ordinate by 1000. The spectra is measured at 2 cm^{-1} resolution from 4000 cm^{-1} to 450 cm^{-1}. The header record includes compound name, formula, molecular weight, Chemical Abstracts Service (CAS) registry number, melting point, boiling point, Wiswesser Line Notation (WLN), etc. More detail on the format of the records has appeared in the literature.[6]

The basic strategy of our method was to use spectra from the database to identify patterns of absorbance that characterize certain functional groups; and then to search for those patterns in a series of 'unknown' spectra. Representatives of a functional group were identified by computer searching the Wiswesser Line Notation (WLN) in the database header records. The list of spectra retrieved by WLN was checked against the compound names to avoid coding errors. An "average spectrum" was calculated by taking the mean absorbance of all the normalized spectra at each frequency interval (2 cm^{-1}) throughout the range. Since the goal of the project is rapid screening, only a small portion of the full IR range was used, a portion chosen surrounding a characteristic band of that functional group. For example, when searching for carboxylic acids or alcohols, the O-H stretching region was used (3800-3400 cm^{-1}).

The average spectrum was considered to represent the functional group. Other spectra were then tested in the same frequency window to see if they exhibited the same pattern of absorbance as the average. A score was assigned to each spectrum, reflecting the degree of similarity to the average. Since this process is similar to a library search routine, except in that it is applied only to a small region of the spectrum, we used the same metric reported in the literature for library searching.

In most of the work described herein, the "difference squared" metric,[7,8] was used.

$$M_{SQ} = \Sigma (S_i - R_i)^2$$

where M_{SQ} is the similarity indicator and S_i and R_i are the absorbance values of the sample and reference spectra in a frequency interval i. Clearly, the smaller the value of M_{SQ}, the better the match between the unknown and the reference

(or average) spectrum: a perfect match would give $M_{SQ} = 0$. ΣS_i means the sum of the absorbance values of the sample spectra in a frequency interval i.

In the moving pattern search, a continuous incremental comparison of the average pattern was used. In the multiple patterns search, multiple regions were searched, a decision was made on the basis of multiple search results.

The speed of our search algorithms is about the same as chemigrams: both of them take about 100 sec to search 1000 spectra. When the moving pattern search is employed, it takes about three times longer than the stationary pattern search or chemigram-type search.

Results and Discussions

One of the most powerful functions of infrared spectroscopy is establishing conclusively the identity of two samples that have identical spectra when determined in the same medium. The region 1500-600 cm^{-1} contains many absorptions caused by bending vibrations as well as absorptions caused by C-C, C-O, C-N and C-Cl stretching vibrations. As there are many more bending vibrations in a molecule than stretching vibrations, this region of the spectrum is particularly rich in absorption bands and shoulders. For this reason, it is frequently called the fingerprint region. Small differences in the structure and constitution of a molecule result in significant changes in the distribution of absorption peaks in this region of the spectrum. As a consequence, a close match between two spectra in this fingerprint region usually gives a strong evidence for the identity of the compounds yielding the spectra. Since many bending vibrations give rise to absorption bands are thus composites of these various interactions and depend upon the overall skeletal structure of the molecule. Exact interpretation of spectra in this region is seldom possible because of their complexity; however, on the other hand, it is this complexity that leads to uniqueness and the consequent usefulness of the region for final identification purposes.

A number of important group frequencies are to be found in the fingerprint region. These include the aromatic ring stretching vibrations at 1620 to 1470 cm^{-1}, the C-O-C stretching vibration in ethers and esters at about 1200 cm^{-1}, the C-Cl stretching vibration at 800 to 600 cm^{-1}, the C-F stretching vibration at 1400-1000 cm^{-1}, C-O vibrations and C-C vibrations at 1250 to 1050 cm^{-1}, and C-H bending vibration in the range of 1000-670 cm^{-1}.[9]

When the spectrum of unknown material is obtained, there are some questions that can be asked by the analyst immediately: does it contain a carbonyl group? is it alcohol? is it acid? is it aromatic? if so, what is the substitution type? Answers to questions such as these will give many clues for chemical work that can lead to the conclusive identification of the compound. Modern GC/IR instruments produce and enormous number of unknown spectra in a short time. It is amazing to note that these questions can be answered fairly quickly by examining selected fractions of the spectra for those tremendous amounts of data. This is possible by monitoring certain spectral windows with the progress of chromatographic fractionation. Selective data reduction by use of chemigrams and our approach in the fingerprint region, we selected aromatic compounds as the first compound class. This group has a great importance in many practical applications and one of the best group frequencies for recognizing the presence of aromatic ring structures occurs in the 1620-1470 cm^{-1}.

The average of 100 spectra of aromatic ring containing compounds from the database is shown in Figure 1. The pattern consists of two peaks which occur near 1586 and 1495 cm^{-1}. The moving pattern search method was used with a 20 cm^{-1} window and a 2 cm^{-1} increment. The search results are shown in Figure 2. More than 76% of the aromatic compounds spectra had the lowest M_{SQ} value of 16.0 or less; fewer than 15% of the non-aromatic compounds had the lowest M_{SQ} of 16.0 or less.

These results were compared with the stationary pattern search (Figure 3) and the chemigrams (Figure 4). At the threshold level of finding 15% of non-aromatic compounds, chemigrams found 74% of aromatic compounds and the stationary pattern search found 75% of aromatic compounds while the moving pattern search 77% of aromatic compounds. In the fingerprint region, the pattern search results were much less satisfactory than in the O-H stretching vibration region or C=O stretching vibration region, although still better than chemigram result. Another thing to be noted was that the moving pattern search did not make a big difference in this region. All these undesirable results came from the fact that in fingerprint region, the spectra showed much variation in peak positions, peak intensities and peak widths, and as a result, did not show a constant pattern of absorbances. Examination of non-aromatic compounds found and aromatic compounds which were not found would be helpful in understanding these results.

The spectra of the non-aromatic compounds that were assigned $M_{SQ} < 16.0$ were examined to see why they had

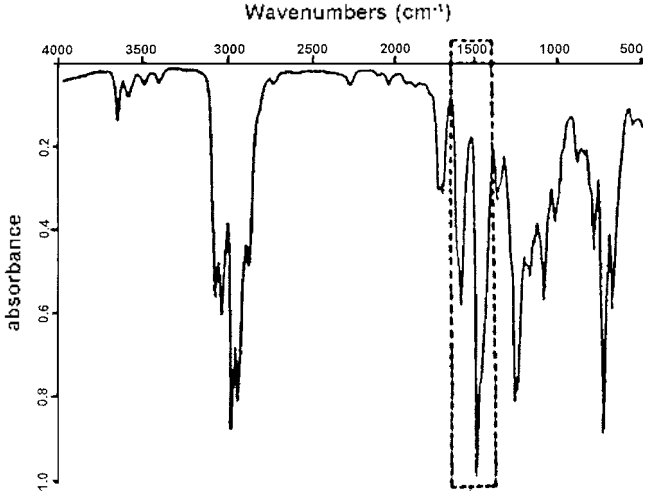
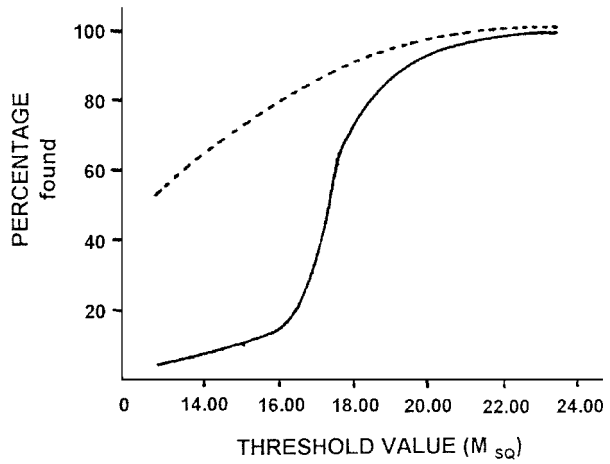
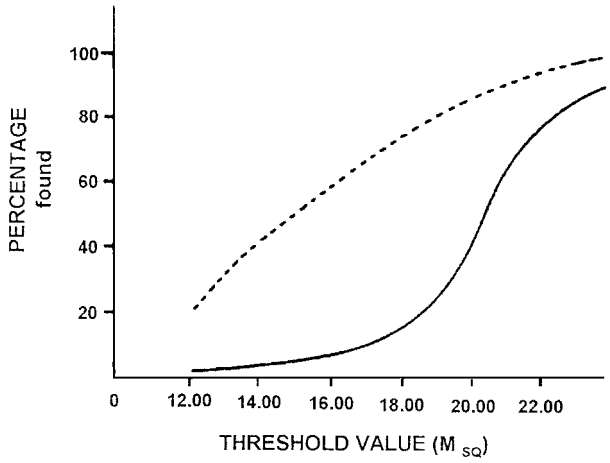


Figure 1. The average spectrum of 100 carbon aromatic compounds (1620-1470 cm^{-1}). Boxed portion shows region used for comparison.



THRESHOLD	A	B
14.0	8.91	62.02
15.0	11.11	66.13
16.0	14.91	76.89
17.0	30.11	84.49
18.0	66.22	94.30

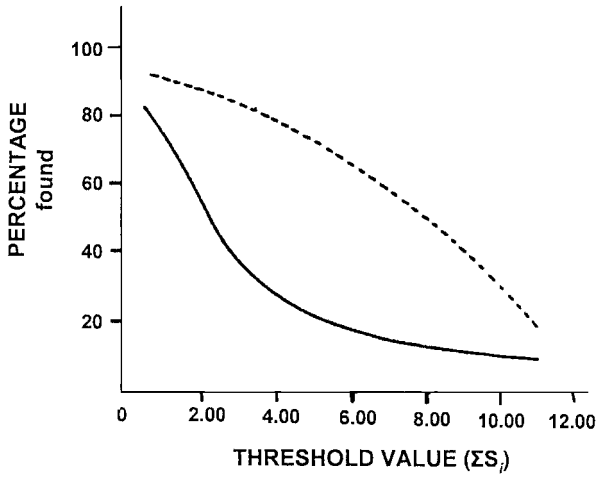
Figure 2. The moving pattern search results of aromatic compounds (1620-1470 cm^{-1}). Percentage A : percentage of non-aromatics (——) with an M_{SQ} value less than the threshold shown; Percentage B: percentage of aromatics (----) with an M_{SQ} value less than the threshold shown.



THRESHOLD	A	B
18.0	11.12	65.70
19.0	13.64	71.41
20.0	17.67	77.35
21.0	27.04	84.73
22.0	45.93	91.24

Figure 3. The stationary pattern search results of aromatic compounds (1620-1470 cm^{-1}). Percentage A: percentage of non-aromatics (——) with an M_{SQ} value less than the threshold shown; Percentage B: percentage of aromatics (----) with an M_{SQ} value less than the threshold shown.

been found. In most cases, these were nitrogen-containing heterocyclic ring compounds such as 4-lutidine, 2-methoxypyridine. 2,3-dihydroindole. 6-methoxyquinoline. Our average pattern was generated only from carbocyclic aromatic compounds. *i.e.*, heterocyclic aromatic compounds were not



THRESHOLD	A	B
2.0	41.10	93.19
3.0	27.80	86.83
4.0	21.71	80.60
5.0	19.01	75.60
6.0	17.05	72.35

Figure 4. The chemigram search results of aromatic compounds (1620-1470 cm^{-1}). Percentage A: percentage of non-aromatics (——) with a ΣS_i value less than the threshold shown: Percentage B: percentage of aromatics (----) with a ΣS_i value greater than the threshold shown.

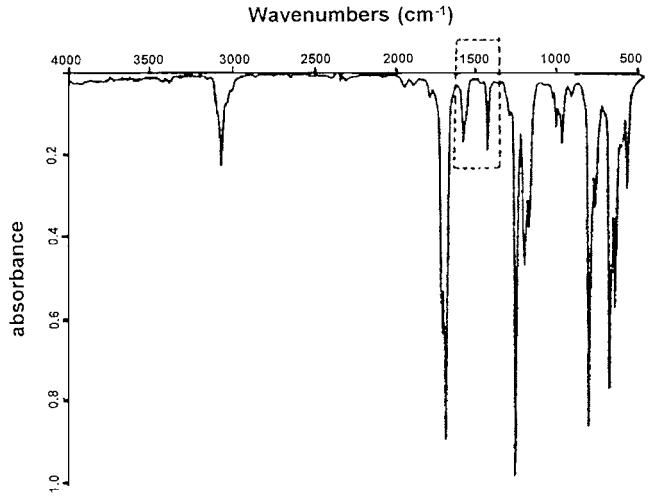


Figure 5. The spectrum of 2,2-dichloroacetophenone. Boxed portion shows region used for comparison.

included. However, these results seem to be all right because it has been known[10] that C=C bands in nitrogen heterocyclic compounds absorb infrared light at the same frequencies as in the carbocyclic aromatic compounds.

Similarly. the spectra of the aromatic compounds that were not assigned $M_{SQ} < 16.0$ were examined. In many cases, the problem was a very low intensity, *i.e.* almost no intensity. These include isobutyl benzoate and the ethyl ester of 4-phenyl-2-butanone-*p*-toluic acid. Other cases. such as 2,2-dichloroacetophenone (Figure 5). and *m*-nitrotoluene, showed a different pattern, *i.e.*, the interval between two peaks was large. It should be noted that all these negative interferences

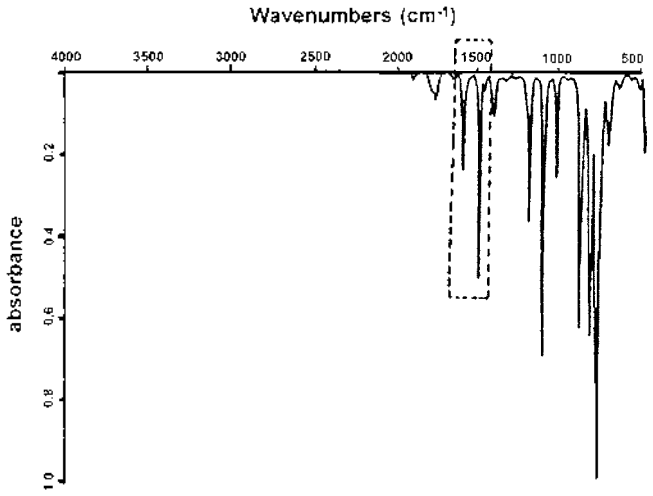


Figure 6. The spectrum of A, A, A, 4-tetrachlorotoluene. Boxed portion shows region used for comparison.

would not be reduced by employing the moving pattern search.

Another serious problem in our algorithms was found in this experiment. In the case of A, A. A. 4-tetracholrotoluene (Figure 6). surprisingly. even though two absorption peaks showed at exact same positions as the average pattern of aromatic compounds, a very high M_{SQ} value was assigned. The difference between two patterns was the peak width: A, A, A, 4-tetrachlorotoluene showed very narrow peaks while the average pattern showed somewhat broad peaks. This illustrates very well that our algorithms have no tolerance for variation in peak width.

References

1. Andregg, R. J. *Amer. Lab.* **1985**, *17*(9), 29.
2. Hites, R. A.; Biemann, K. *Anal. Chem.* **1970**, *42*, 855.
3. Andregg, R. J. *J. Chromatog.* **1983**, *275*, 154.
4. LaBrosse, J. L.; Andregg, R. J. *J. Chromatog.* **1984**, *315*, 83.
5. Mattson, D. R.; Julian, R. L. *J. Chromatog. Sci.* **1979**, *17*, 416.
6. Griffiths, P. R.; Azarraga, L. V.; Dehaseth, J. A.; Hannah, R. W. *Appl. Spectros.* **1979**, *33*, 543.
7. Pyo, D. *Vibrational Spectroscopy* **1993**, *5*, 263.
8. Pyo, D.; Lee, J. *Vibrational Spectroscopy* **1994**, *8*, 61.
9. Silverstein, R. M.; Morrill, T. C. *Spectrometric Identification of Organic Compounds*, 4th ed.; Wiley: New York, 1981; pp 166-172.
10. Nakanishi, K.; Solomon, P. H. *Infrared Absorption Spectroscopy*, 2nd ed.; Holden-Day: San Francisco, 1977.