Rapid Identification of Aliphatic Ketones from Vapor-Phase Infrared Spectra

Dongjin Pyo[*] and Jongchon Choi

Department of Chemistry, Kangwon National University, ChunChon 200-701, Korea
Received July 8, 2004

The need for the identification of compounds present in complex gaseous mixture or after separation by gas chromatography (GC) using vapor-phase infrared (VPIR) spectrometry is becoming increasingly important. This is primarily due to advances in Fourier transform infrared (FT-IR) spectrometry allowing the acquisition of complete spectra in a few tenths of a second with detection limits of less than 1 ppb in long path infrared gas cells and below 10 nanogarms in the GC/FT-IR interface. With a GC/FT-IR system hundreds of spectra can be generated per chromatogram. Clearly, the limiting processes are the spectral searching and, in cases where the search results are ambiguous, spectral interpretation steps. A powerful digital computer is needed to control the interferometer and perform the Fourier transformation from the time domain to the frequency domain in order to obtain the spectrum. This makes the incorporation of automatic spectral interpretation in the same computer an obvious and sensible capability. It should be kept in mind that the ultimate goal of the chemist is not simply to produce a collection of spectra but rather to determine what chemical compounds are present in the unknown samples and what their significance is to the problem at hand.

The laborious task of spectral identification has been greatly simplified by the use of computer-based library searching routines. Two general classes of algorithms has been developed: procedures that seek to make an exact identification of an unknown by direct comparison with spectra of known compounds (which require the spectrum of the unknown to be in the reference databases), and more general approaches striving to identify all the functional groups in the unknown. A wide range of encoding schemes has been proposed[1] to represent spectra in a digital form suitable for library searching and interpretation. The ideal storage format retains the minimum amount of information necessary for correct identification of unknown compounds in a form permitting rapid numerical comparison. The optimum instrumental parameters for acquiring spectra for spectral searching and identification are a function of the sample, so that practical considerations require some degree of standardization in sampling and measurement. In addition, specific computer characteristics such as word size, amount of semiconductor memory, and types of mass storage devices must also be taken into account.

Buechi *et al.*[2] have discussed the compilation of spectral libraries in terms of five operations: selection, digitization, completion, formatting, and verification. For infrared spectral reference data bases, the problem of digitization has proven to be critical. Ideally the encoding scheme should be equally applicable to condensed-phase and vapor-phase spectra. Historically, however, the compilation of large libraries of condensed-phase spectra was begun over two decades before the commercial introduction of the first FT-IR spectrometer, so that most of these spectra were only available as hard copy of analog data. The problem of converting large databases of hard-copy spectra to a computer-compatible format has been approached by the use of graphics digitizers.[3] On the other hand, most vapor-phase reference spectra were measured using FT-IR spectrometers in the last ten years. Thus not only is their quality higher from an instrumental standpoint, but also they do not suffer from many of the other disadvantages of condensed-phase infrared spectra, such as shifting and relative intensity variation of bands depending on the method of sample preparation, baseline changes due to scattering, etc.

Due to the inherent complexity of infrared spectra, definitive identification of unknown samples usually requires direct comparison with previously collected spectra. Most spectral searching algorithms are designed to step through entire libraries in a sequential manner. These "brute force" methods have become practical only in recent years with the advent of completely digitized spectral libraries and fast, low-cost computers combined with mass storage devices. Much of the work of developing library-based spectral identification programs arose from the need to interpret the large number of spectra generated by GC/FT-IR experiments. Although sequential searches can provide excellent results,[4] it is inefficient to search an entire spectral database to identify an unknown compound. The situation is exacerbated when a good match is not found because of the limited size of the reference library. Investigation of a subset of the database, in the form of a class of known compounds, greatly reduces the amount of time required for the identification. In addition, the chemical functionality of molecules not present in the database can be ascertained. As an example of the success of this approach, "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 the whole 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 absorbance 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 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 algorithms 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.

Experiments

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 form Dr. James de Haseth at the University of Georgia. 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 weigh, 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. Repre-

sentatives 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.

$$M_{SQ} = \Sigma(Si - Ri)^2$$

Where M_{SQ} is the similarity indicator and Si and Ri 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 $M_{SQ} = 0$. ΣSi means the sum of the absorbance values of the sample spectra in a frequency interval i.

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 Discussion

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 group 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.

The structural unit C=O has an excellent group frequency, which is described as a stretching vibration. Since the C=O group is a terminal group, only the carbon is involved in a second chemical bond. This reduces the number of force constants determining the spectral position of the vibration. The C=O stretching vibration usually appears in a frequency range that is relatively free of other vibrations. This reduces the possibility that a vibration of the same species could interact with the C=O vibration. For example, in many carbonyl compounds the double bond of the C=O has a force constant different from those of such structural units as C–O, C–C, C–H, etc.; only structural units such as C=C have force constants of magnitudes similar to that of the C=O group. The C=C vibration could interact with the C=O if it were of the same species, but generally it is not. Almost all carbonyl compounds have a very intense and narrow peak in the range of 1800-1600 cm^{-1}. This is why this region is considered as a very important region by organic chemists.

The C=O stretching vibrations of various carbonyl groups absorb in the region 1800-1600 cm^{-1}; ketones, aldehydes, acids, amides, and carbonates all have absorption peaks around 1750-1650 cm^{-1}, while esters, acid chlorides, and acid anhydrides tend to absorb at slightly higher wavenumbers; that is, 1770-1725 cm^{-1}. Usually, the type of carbonyl group is identified by taking into account vibrations other than the C=O stretching.

We selected aliphatic ketone compound class as the first example in this region, since it has the simplest structure of all carbonyl compounds, *i.e.*, it has only carbon chains on both sides of carbonyl group, for example, 4-decanone, 3-undecanone, 3-methyl-, 2-pentanone, 2,8-dimethyl-, 5-non-
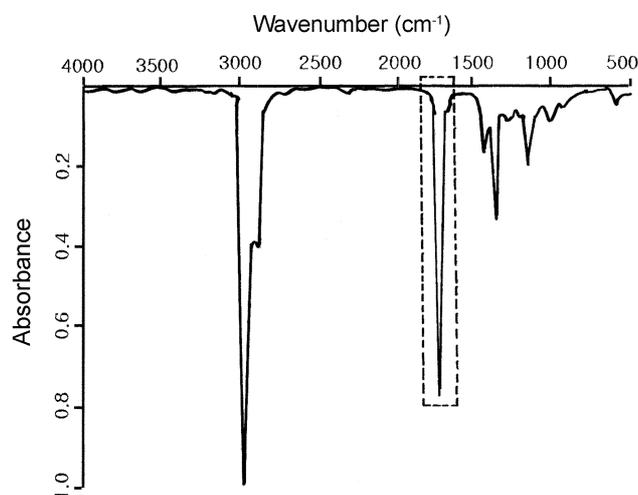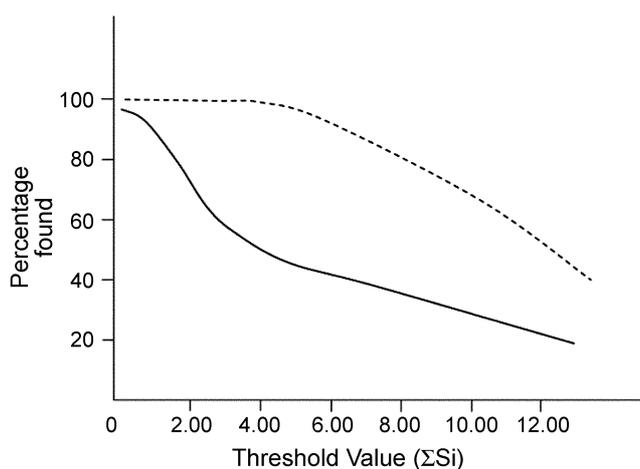

THRESHOLD	A	B
3.0	58.60	100.00
4.0	53.38	100.00
5.0	49.00	96.49
6.0	46.05	91.20
7.0	43.26	85.96

Figure 2. The pattern search results of aliphatic ketones (1800-1600 cm^{-1}). Percentage A; percentage of non-aliphatic ketones (—) with an M_{SQ} value less than the threshold shown. Percentage B; percentage of aliphatic ketones (----) with an M_{SQ} value less than the threshold shown.

anone etc. Aliphatic ketones such as 5-nonacosanone[9] and 7,7-dimethyl-5-ketoanatricontane[10] have been known as flavoring materials with secondary alcohols.[11] In 1985, Maria and Rozsa[12] extracted a series of long-chain aliphatic ketones from the fruits of Evodia hupehensis, and showed honeybee attractant activity. Aliphatic ketones were also found as major components in the analysis of paint vapors. Ross *et al.*[13] sampled the vapors by heating the paints for approximately 30 min., then made derivatives of aldehydes and ketones to form quaternary ammonium salts that were analyzed by secondary ion mass spectrometry (SIMS) and GC/MS. This method enables the selective detection of carbonyl-containing compounds at picomole levels.

In the similar way as before, we selected 32 aliphatic ketones from the database, and calculated the average spectrum of 32 aliphatic ketones (Figure 1). The pattern of aliphatic ketones in the window 1800-1600 cm^{-1} showed a narrow sharp band centering around 1716 cm^{-1}. We calculated the M_{SQ} values on the entire database using the window of 1800-1600 cm^{-1}. The search results were very good (Figure 2). As a threshold level of M_{SQ} = 3.20, 100% of aliphatic ketones are found; *i.e.*, all aliphatic ketones had M_{SQ} value less than 3.20, along with 2.4% of non-aliphatic ketones. Most of the 2.4% of non-aliphatic compounds found were saturated, 6-membered ring ketones, for example, cyclohexanone (Figure 4), 4-methyl-, cyclohexanone, 2-propyl-, cyclohexanone, of which carbonyl stretching

Figure 1. The average spectrum of 32 aliphatic ketones Boxed portion shows region used for comparison.
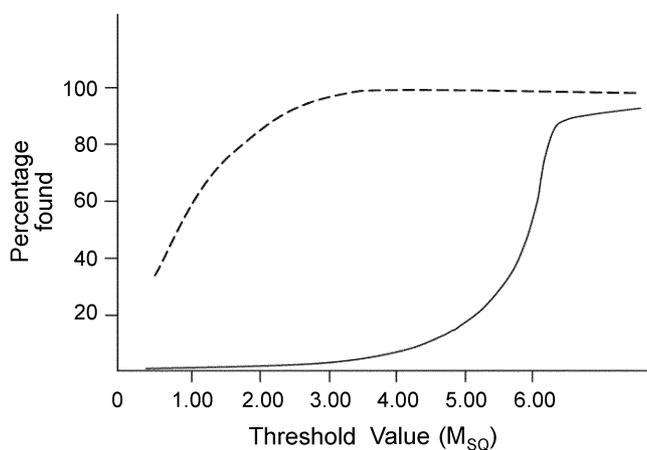
THRESHOLD	A	B
2.8	1.88	94.73
2.9	1.91	95.49
3.0	2.06	96.49
3.1	2.25	98.24
3.2	2.43	100.00

Figure 3. The chemigram search results of aliphatic ketones (1800-1600 cm^{-1}). Percentage A; percentage of non-aliphatic ketones (—) with a Si value greater than the threshold shown. Percentage B; percentage of aliphatic ketones (----) with a Si value greater than the threshold shown.

vibrations overlap exactly with the ones of saturated aliphatic ketones.[14]

These results can be compared with chemigram results (Figure 3). In chemigram results, it is found that all aliphatic ketones had the integrated intensity value greater than 4.0, but more than 53% of non-aliphatic ketones had also the integrated intensity greater than 4.0. Among those 53% of non-aliphatic compounds, most of them were carbonyl compounds including esters, carboxylic acids, aldehydes etc.

Conclusion

Computerized interpretation of vapor phase infrared spectra using a pattern search for spectra/structure correlation for vapor phase spectra is introduced. Rapid identification of aliphatic ketones in gaseous mixture can be achieved using this pattern search. This new approach of
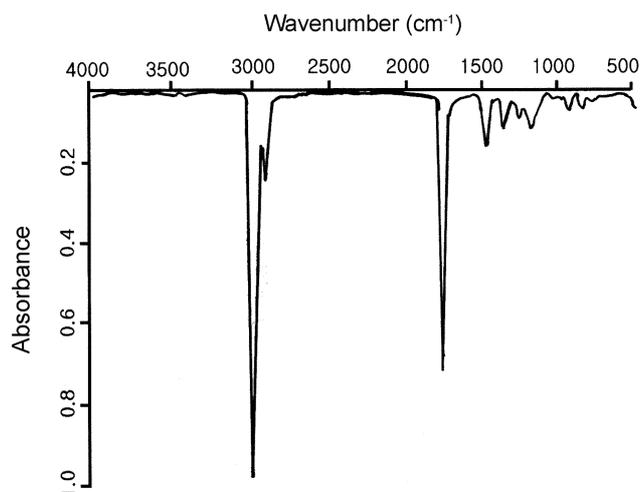

Figure 4. The spectrum of cyclohexanone.

interpretation of infrared data based on the average pattern of infrared spectra has potential for rapid identification of the components of the complex mixture samples. The methods can also be used in generating vapor-phase infrared spectra/ structure correlation rules to obtain a better understanding of the correlation.

References

1. Clerc, J. T.; Zupan, J. *Pure App. Chem.* **1977**, *49*, 1827.
2. Buechi, R.; Clerc, J. T.; Jost , C. *Anal. Chem. Acta* **1978**, *103*, 21.
3. Delaney, M. F.; Uden, P. C. *Anal. Chem.* **1978**, *50*, 2156.
4. Erickson, M. D. *Appl. Spectrosc.* **1981**, *35*, 181.
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 Spectorscopy* **1993**, *5*, 263.
8. Pyo, D.; Lee, J. *Vibrational Spectorscopy* **1994**, *8*, 61.
9. George, V. *Fitoterapia* **1985**, *56*(2), 117.
10. Srivastava, Santosh, K. *Indian J. Pharm. Sci.* **1984**, *46*, 12.
11. Flavor and Extract Manufacturers Associ. of the United Stats (Washington D. C., USA), Report 1984, FEMAUS-84/01, A1.
12. Maria, G.; Rozsa, Z. *Herba Hung* **1985**, *24*, 53.
13. Ross, M. M.; Kidwell, D. A.; Colton, R. J. *Int. Jour. of Mass Spec. and Ion Proc.* **1985**, *63*, 141.
14. Dyer, J. R. *Application of Absorption Spectroscopy of Organic Compounds*; Prentice-Hall, Inc.: New Jersey, 1965.