Fourier Transform Infrared Spectroscopy without an FTIR Spectrometer: Library Searching and Concise Storage of Spectra†

DONALD F. AVERILL,* KRISTI C. BAIRD, LAURA L. HOPKINS, and MARK J. YERKES

Department of Physical Sciences, Eastern New Mexico University, Portales, New Mexico 88130

Received October 20, 1989

By use of a dispersive infrared spectrophotometer and a minicomputer for data collection and manipulation, a library of liquid and solid sample infrared spectra was constructed by storing FFT data. A library search routine, which compares signs of cosine and sine FFT terms, discriminates between functional groups reasonably well. Accurate matches of spectra have correlation index (CI) values of about 0.9 with a standard deviation of 0.04. The computer can search 255 spectra in less than 2 min. One 8-in. 0.5-MB floppy disk can store 488 spectra. All programming was done in BASIC.

INTRODUCTION

Development of a fast and reliable Fourier transform infrared (FTIR) spectral library search technique has been a concern in the 1980s. A summary of the practical aspects for computer-based spectral search techniques has been presented in the literature.[1] Approaches to spectral library searching have been briefly summarized by Wang and Isenhour,[2] and a more thorough review has been given by Heller and Lowry.[3] Several problems arise when libraries of infrared spectra or libraries of FTIR interferograms are searched with a computer in an attempt to identify unknown spectra. The problems that have been considered are (A) interferometer phase error and other spectrometer-dependent properties,[4] (B) normalization of spectra (closure effects),[5,6] and (C) noise or signal to noise (S/N) ratio.[6,7]

After reviewing the various methods used for storage of infrared spectra obtained from both dispersive and FTIR instruments and the subsequent library search routines used, we began investigating the possibility of using the fast Fourier transform (FFT) to store and search spectral data taken from our dispersive infrared spectrophotometer. Our goals were to devise a method to easily acquire the data with a minicomputer, perform a Fourier transform of the data, store a minimum amount of data that would allow reconstruction of spectra having reasonable resolution, and develop a fast and reliable method of identifying unknown spectra by computer searching our data base.

The search method we describe below utilizes FFT data obtained by taking the Fourier transform of the original infrared spectra acquired from a dispersive instrument. Since a dispersive instrument rather than an FTIR spectrometer was used for collection of data, interferometer phase error was not a problem. However, to maintain good correlations between spectra, the dispersive instrument should be routinely calibrated to prevent peak positions from changing. Since our data were taken from a single instrument, it was not necessary to consider other instrument function problems. Our search method sets limits for the concentrations of unknown spectra for comparison with library entries. It was not necessary to normalize spectra. Normalizing spectra (taken in the percent transmittance mode) to unit maximum absorbance requires at least three manipulations of the data. First, conversion of the data to absorbance, then a search for the maximum, and finally a multiplication (or division) must be carried out. If there is a base-line correction to be performed, a preliminary manipulation of the data must be performed. Although the above procedures are not difficult to perform on a computer, none of these computations are necessary with this method. Thus, the time necessary to make a judgment for base-line correction, for additional programming, and for computer usage are all saved, speeding up the process of identifying a sample. We are able to store Fourier coefficients for 488 spectra on a 0.5-MB, 8-in. floppy disk, and a search of 256 spectra requires slightly less than 2 min. All programs were written in BASIC.

EXPERIMENTAL PROCEDURES

Spectra were acquired from a Perkin-Elmer infrared spectrophotometer, Model 281, with a Digital Equipment Corp. (DEC) MINC 11/23 laboratory computer system interfaced to the cam pulse and the 1-V output terminals on the back panel of the spectrophotometer. Spectra were recorded by using 6-min scan times. Hard copies of reconstructed spectra were either copied from the DEC VT125 terminal screen to a DECWRITER IV graphic printer or plotted on an xy plotter. The computer–plotter interface has been described elsewhere in the literature.[8] Reconstructed spectra have a resolution of approximately 18 cm^{-1} between 4000 and 2000 cm^{-1} and 9 cm^{-1} between 2000 and 600 cm^{-1}. For liquid samples, a few drops of the liquid were sandwiched between NaCl salt plates, and solid samples were made into KBr pellets (1 mg of sample/150 mg of KBr). About 80% of the 325 spectra recorded for this work were of liquids because of the quick and simple sample preparation. Concentration studies were performed with solids by using KBr pellets. An accurately weighed amount of a sample was diluted with KBr and ground to a fine powder with an agate mortar and pestle. Appropriate portions of this concentrated powder were then weighed and diluted further with KBr and reground to give concentrations from 0.1% to 1.3% sample in 0.1% or 0.2% increments.

In the range 4000–2000 cm^{-1} the spectrophotometer produces a 5-V cam pulse every 0.1 cm^{-1}, and from 2000 to 600 cm^{-1} a pulse occurs every 0.05 cm^{-1}. During a scan from 4000 to 600 cm^{-1}, there are 48 000 pulses. Our program for acquiring the data uses the DEC Schmitt trigger routine SCHMITT to count pulses, and a voltage reading from the 1-V output range was taken every 24 pulses. The voltage readings were multiplied by 1000 and stored in an integer array of 2000 points (48 000 pulses/24 = 2000). Since the DEC BASIC FFT routine requires 2^n (n = an integer) data points, the first 48 points (of the 2048-point data array, $n = 11$) were set equal to the first value taken from the spectrophotometer. Then, points 49–2048 were set equal to the infrared spectral data, which gave 48 + 2000 = 2048 points for the FFT. After the transform was completed (the 2048-point FFT takes about 4

†This work was partially supported by an Institutional Research Grant from the Llano Estacado Center for Advanced Professional Studies and Research of Eastern New Mexico University.

0095-2338/90/1630-0133$02.50/0

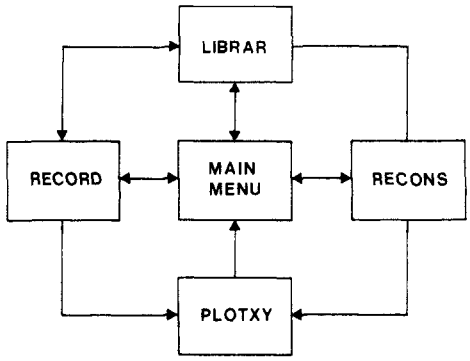

Figure 1. Diagram showing access routes to programs. Line without arrowheads indicates only limited functions are available.

s), the first 255 cosines and 255 sines were stored on a floppy disk. The spectral searching and reconstruction were carried out by using this data.

Reconstruction of a spectrum was accomplished by loading the stored sine and cosine coefficients into 2048 point arrays. The arrays were rebuilt with $C_n = C_{2050-n}$ for the cosine terms and $S_n = -S_{2050-n}$ (n = 1–255) for the sine terms, and all other terms were set to zero. An inverse FFT was performed, and the first 48 data points were omitted when the spectrum was plotted or viewed on the computer terminal. It took about 10 s to rebuild a spectrum from the disk data.

RESULTS AND DISCUSSION

BASIC Programs. Programs were written in DEC BASIC, and because of the small amount of memory available for programs (at most, 8 kB) and large array sizes (2048 sines and 2048 cosines), the programs for data acquisition (RECORD), library searching (LIBRAR), reconstruction of spectra (RECONS), and plotting (PLOTXY) were "chained" together. Figure 1 illustrates how the various programs were accessed.

Following the acquisition of a spectrum with the RECORD program, the user names the spectrum (30 characters or less), and the data are stored on a second "data" disk. Names of spectra are stored in a separate file on the system disk. The data disk has a small directory for eight files labeled FFT(n).DAT (n = 1–8). Each of these files can contain 61 spectra (data values are indexed from 1 to 31 110), for a total of 488 spectra per disk. This procedure used the disk storage space most efficiently.

The library routine (LIBRAR) allows access to all the names and index numbers (on the system disk) of the spectra saved on the second data disk. The routine also allows changing the name of a spectrum (for the case of occasional misspelling or modifying a "common" name to a correct chemical name) and contains the searching or "correlation" algorithm that is discussed below.

The reconstruction program (RECONS) builds a 2000-point spectrum from the stored FFT data and provides three options for printer hard copies. The entire spectrum can be copied from (1) a single terminal screen image (one of every four data points plotted), (2) two 500-point screen images (every other point plotted), or (3) four 500-point screen images (all data points available).

The PLOTXY program allows the reconstructed infrared spectrum to be drawn on an xy plotter and labeled with markers at 4000, 2000, and 600 cm^{-1}. A border can be added to the spectrum, and tick marks are provided every 100 cm^{-1} to complete the reconstruction. Figure 2 shows the reconstructed spectrum of polystyrene. It can be seen that after the data are acquired and stored and the spectrum is rebuilt, all the salient features of the spectrum are still present.

Library Searching or Correlations. An investigation of the FFT cosine terms for different KBr pellets (containing different concentrations of a known solid sample) showed that the first few terms contained considerable variation in amplitude and occasionally sign. However, we noted that from the 11th through the 50th coefficient there was a surprising consistency of sign. We also noted that when only the first 10 sines and cosines were used to reconstruct a spectrum, very little structural information was present, only very broad peaks and valleys. We found from an investigation of 10 compounds (chosen randomly from the library) that 14 of 20 frequencies (70%) of the signs matched in 50% of the library entries. This is to be expected from the low-frequency terms. Among other general features of an infrared spectrum, these terms provide the information for the construction of the plateau region between 2800 and 2000 cm^{-1} (where little absorption occurs) and the depression in the fingerprint region (where many absorptions occur). The higher frequency terms contain the data that provide the differences between similar spectra. Therefore, we decided to use the signs of the 40 cosine terms (from 11 through 50) for comparison with the spectral library as a spectral search routine. The 40 cosine coefficients (11–50) of the "unknown" were loaded into a test array, and the corresponding 40 terms of each library spectrum were compared to the test array by sign only. Each mismatch was recorded, and a correlation index (CI) was calculated as (40 − number of mismatches)/40, each match contributing 0.025 (1/40) to the CI. An adjustable threshold CI value was usually set at 0.8, and a list of compounds possessing CI values greater than the threshold was obtained. The procedure is rapid and discriminates between different classes of compounds reasonably well. A survey of the search methods reported in the literature indicates that this library searching method is unique.

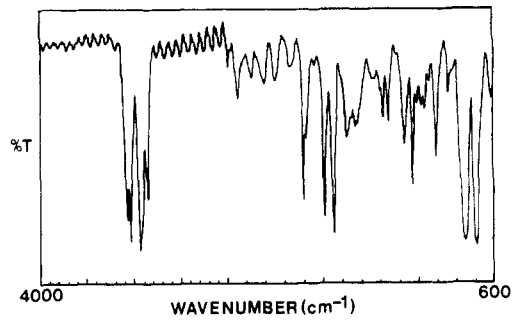

Figure 2. Spectrum of polystyrene reconstructed from 510-point FFT data and plotted on an xy plotter.

In addition to the 40 coefficient CI values, we calculated a second CI (the first search could be considered a filter), which we designated ECI (extended correlation index). The ECI took two forms. The first was like the CI, but the signs of 100 cosine coefficients were compared (11th through 110th terms). The second used the original 40 cosine coefficients and an additional 40 sine coefficients for a total of 80 signed values. The position of the sine coefficients in the data set corresponded to the cosines (11th through 50th).

We compared CI values of the same sample scanned at times of 3, 6, and 12 min. CI values of 3-minute scans compared to the longer scan times were typically less than 0.8. There are two possible explanations for the low CI value. Visual inspection of some spectra taken with 3-min scan times shows a slight peak position change and an alteration of peak shape when compared to data from longer scan times. This might be expected because of the response time of the instrument. The other explanation could come from the computer system. The spectrometer cam pulses would be occurring at a rate of about 300 Hz, and the computer may not be able to acquire the data corresponding to the same wavenumber at which the data were acquired during the longer scan times. We suspect a combination of these effects results in a lower

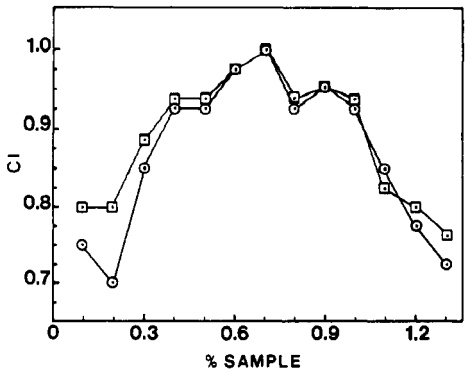

Figure 3. Plot of CI and ECI_{80} of resazurin as a function of concentration (KBr pellet). (⊙) CI; (⊡) ECI_{80}.

correlation index. The 6- and 12-min scans gave the same CI (1.0). This information supported our decision (to save time) to run all spectra with a 6-min scan rather than the default 12-min scan time.

Since normalization of spectra was not performed, it was necessary to study the CI as a function of concentration. Samples were prepared as described under Experimental Procedures. By use of the concentration of 1 mg of sample/150 mg of KBr as the standard (0.7%), CI values were obtained from a computer search of the spectra. Figure 3 shows plots of CI and ECI_{80} versus percent sample for the compound resazurin. It can be seen that the CI and ECI do not decrease significantly until the sample concentration drops to less than 0.4% or becomes larger than 1.0%. This indicates that great care need not be taken to make pellets of a fixed concentration for spectral searching. Other samples exhibited the same general behavior with major deviations attributable to poor-quality pellets.

Reproducibility of spectra was checked with CI values. With the same sample and identical salt plates, six duplicate spectra (one used as a reference) gave a mean CI of 0.875 with a standard deviation (s) of 0.025. The mean ECI_{80} was 0.927 with $s = 0.019$. A similar experiment was performed with KBr pellets containing a solid sample, and the corresponding values were 0.910 and 0.042 for the CI and 0.920 and 0.024 for the ECI_{80}. By use of different sets of NaCl plates and samples of the same compounds from different chemical manufacturers, CI values were typically from 0.8 to 1.0 (eight sign mismatches to a perfect match). We were able to obtain spectra of samples in KBr pellets as reproducibly as with spectra of liquid samples between salt plates. However, when identical samples were compared, deviations of CI values greater than 0.15 indicated poorly made pellets (visual inspection), and good-quality pellets gave CI values comparable to those from samples of liquids mounted between salt plates.

As a test of the method, we found several bottles of reagents without labels in our storeroom so the contents became our unknowns. With the CI threshold set equal to 0.8, a library search was performed for one of these samples, and the largest CI obtained was 0.875 for butyl methacrylate, monomer. An NMR spectrum was obtained for this sample and compared with those of methacrylate compounds. The NMR spectrum was identified as methyl methacrylate. The other bottles had remnants of labels and had been marked "from B names". These unknowns were identified as 1-bromohexane and benzonitrile after each unknown spectrum was compared with spectra of compounds from computer-generated lists of compounds that had CI values above 0.85.

Table I shows the results of comparing the various searches for three compounds, 1-pentanol, 2-heptanone, and 1-bromopentane. These samples were chosen because the library contained several compounds that were similar; i.e., they had the same functional group position but different chain lengths,

Table I. Correlation Index Values for Compounds Found from Library Searches

compound	CI	ECI_{100}	ECI_{80}
Unknown 1: 1-Pentanol			
1-pentanol	0.9	<0.8	0.9125
1-octyne	0.85	<0.8	<0.8
1-pentanol[a]	0.825	<0.8	0.875
1-dodecanol	0.8	<0.8	0.8375
kerosene	0.8	<0.8	<0.8
1-octanol	0.8	<0.8	0.8375
Unknown 2: 2-Heptanone			
2-heptanone	0.925	0.87	0.875
2-octanone	0.875	0.95	0.9125
hexadecane	0.85	<0.8	<0.8
1-octanethiol	0.825	0.91	<0.8
p-methylcyclohexanone	0.825	0.9	<0.8
1-chlorooctane	0.825	0.76	<0.8
isovaleraldehyde	0.825	0.9	<0.8
gasoline	0.825	<0.8	<0.8
Unknown 3: 1-Bromopentane			
1-bromopentane	0.825	0.85	0.85

[a] Duplicated entry explained in text.

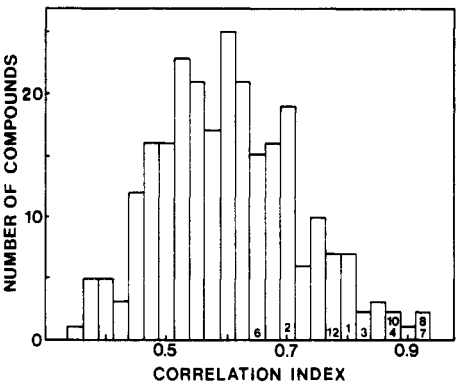

Figure 4. Histogram showing frequency of occurrence of CI values from a library search of 254 liquids using 1-pentanol as the reference (CI = 1.0). The CI values of the primary alcohols in the library are marked with small numbers indicating the number of carbon atoms in the alcohol.

so we could get some idea of the discrimination of the different ECI methods. Six library entries were obtained with CI values equal to or greater than 0.8 for the test compound 1-pentanol. The first and third entries are 1-pentanol, which was the identity of the test compound. The second entry for 1-pentanol was due to an inadvertent duplication since it was entered as amyl alcohol. Of the five different compounds, three were alcohols. The extended correlation index calculated from 100 cosine terms, ECI_{100}, indicated that none of the compounds was a very good match. The ECI calculated from 40 cosines and 40 sines, ECI_{80}, exhibited an increase when it was compared to the CI values for each of the alcohol entries, and the values for 1-octyne and kerosene dropped below 0.8.

The second unknown, 2-heptanone, was correctly identified by the initial 40-term CI as the closest match and 2-octanone a close second. However, both ECI calculations identified 2-octanone as the closest match, which was in error by one methylene group. It is clear that the ECI_{80} discriminated better than the ECI_{100} did since the ECI_{80} values for six of the eight entries dropped below 0.8 and the values for the first two entries changed very little.

The only listing having a CI greater than or equal to 0.8 for the third unknown corresponded with the identity of the unknown, 1-bromopentane. The two ECI calculations coincided with a value of 0.85. In general, we found the ECI_{80} to be the more useful ECI value.

Figure 4 is a histogram showing the frequency of CI value versus CI for 254 liquids used in a library search with 1-

pentanol the reference spectrum. The numbers in the histogram bars are the numbers of carbon atoms in the primary alcohols present in our library, and the position indicates the CI value for each alcohol. Notice that the compounds most similar in structure to 1-pentanol (1-butanol and 1-hexanol) do not have CI values closest to 1-pentanol. We suspect the major factor involved is different grades of chemicals. The sample of hexanol was of practical grade, whereas the butanol and pentanol samples were of reagent grade. We made no attempt to use the highest grade of chemicals when we built our library, and reagent, practical, and technical grade compounds were used. In spite of this lack of uniformity, the compounds listed in the top 10 CI values below 1-pentanol contained five alcohols. It can also be seen that the number of compounds having CI values less than 0.8 rises rapidly, and consequently discrimination will be lost as the value of the CI decreases. Other investigators have used direct interferometric data searching methods such as computation of the dot product between vector representations.[9]

CONCLUSION

It might be argued that when this technique is used, memory savings are minimal. However, to obtain 12-cm^{-1} resolution of absorbance data from 600 to 4000 cm^{-1}, storage of 567 data points would be required. Many of the subtleties of infrared spectra (peak shapes and shoulders) are lost when they are reconstructed from data stored as small sets of absorbance data. Also, the data for the search routine would probably require another file to be constructed, and time would be lost writing and implementing a search program and constructing the additional data file (e.g., choosing the peaks and assigning wavenumbers).

The procedure we have developed for searching IR library data is less complicated than many other schemes presented in the literature (see refs 1 and 3). Two primary advantages of this procedure are (1) the elimination of the necessity for the user to select data or use a computer program to construct a search library and (2) the use of very simple programming for the search and correlation method. Many of the complications other researchers have addressed do not occur when the data are acquired from a single dispersive instrument. The method relies on the FFT for storage, and searching is accomplished by sign comparisons of FFT data. Some reconstructed spectra exhibit overshoot oscillations at high and low wavenumbers as a result of truncation of the FFT coefficients, but the rebuilt spectra are more than adequate for comparison with higher resolution spectra. Since 2000-point spectra are stored as 510 integers, disk storage space for spectra has been reduced by a factor of 4. Our investigation of the search routines is admittedly limited because of the lack of an extensive spectral library; however, small laboratories may find this approach a reasonable alternative to acquiring a commercial (usually much larger than necessary) library and the attendant software. Although we used a minicomputer for our study, modern microcomputers with hard disk storage and higher clock frequencies (>16 MHz system clock and compiled programs) should out-perform our system by as much as a factor of 10 in speed. The books by Ramirez[10] and Burrus and Parks[11] are recommended to those not familiar with the FFT properties and programs.

ACKNOWLEDGMENT

We thank undergraduates Brooks Lane and Raymona Turner for testing the system while working on other research activities and Durwin Striplin and high school students Lynden Armstrong and Albert Sae for their help with the acquisition of spectra.

REFERENCES AND NOTES

(1) Coates, J. P.; Hannah, R. W. In *Fourier Transform Infrared Spectroscopy*; Theophanides, T., Ed.; Reidel: Dordrecht, Holland, 1984; pp 167–185.
(2) Wang, C. P.; Isenhour, T. L. *Appl. Spectrosc.* **1987**, *41*, 185–195.
(3) Heller, S. R.; Lowry, S. R. In *Computer-Enhanced Analytical Spectroscopy*; Meuzelaar, H. L. C., Isenhour, T. L., Eds.; Plenum: New York, 1987; Chapter 11.
(4) de Haseth, J. A.; Azarraga, L. V. *Anal. Chem.* **1981**, *53*, 2292–2296.
(5) Owens, P. M.; Isenhour, T. L. *Anal. Chem.* **1983**, *55*, 1548–1553.
(6) Harrington, P. B.; Isenhour, T. L. *Appl. Spectrosc.* **1987**, *41*, 1298–1302.
(7) Olson, M. L. *Proc. SPIE—Int. Soc. Opt. Eng.* **1981**, *289*, 236–239.
(8) Averill, D. F.; Beatty, G.; Cheng, F. A.; Hauser, A. *J. Chem. Educ.* **1986**, *63*, 627.
(9) Richardson, P. T.; de Haseth, J. A. *Anal. Chem.* **1988**, *60*, 386–390, and references cited therein.
(10) Ramirez, R. W. *The FFT: Fundamentals and Concepts*; Prentice-Hall: Englewood Cliffs, NJ, 1985.
(11) Burrus, C. S.; Parks, T. W. *DFT/FFT and Convolution Algorithms*; Wiley: New York, 1985.