Spectrum Comparison of IR Data Taken from Different Spectrometers with Various Precision

H. Kavak* and R. Esen

Physics Department, University of Cukurova, 01330 Adana, Turkey

Received December 14, 1992

A group of algorithms and a computer program based on these algorithms are developed for comparing IR spectra of different origins and/or of different precisions. For this purpose the spectra to be compared matched in precision. Data with smaller wavelength intervals are reduced using different well-known methods, such as moving averages, least squares, and reciprocal exponential convolution. Later baseline correction and normalization procedures, simple noise elimination, and interference checking methods are applied, and the results are compared to decide on the appropriate method for this purpose. One of the most common comparison problems encountered is that the reference and the sample data are usually recorded by different spectrometers and with different scanning parameters. One frequently needs to compare his/her sample spectra with an atlas of IR spectra. Development of an instrument free comparison method therefore will be quite useful for many purposes.

INTRODUCTION

Even though modern digital infrared (IR) instrumentation and commercially available reference libraries with good precision are quite common, researchers have a considerable amount of reference data recorded using analog IR spectrometers. If one encounters such a problem, first he/she will digitize the analog spectrum by manual means or by use of some sort of digitizer (scanner, plotter-digitizer, etc.). After the digitized spectrum is obtained, the comparison process by computer assisted methods is possible.

Since the computer programs can use digitized spectra, one has some advantages in using commercial and/or public domain software. With digitized analog recorded spectra and with the aid of suitable software, one can determine the peak positions precisely, integrate any absorbance band very accurately and quickly, make a fast library search, and have other signal/spectrum processing aids (such as baseline determination, noise reduction, normalization, absorbance–transmittance conversion, etc.).

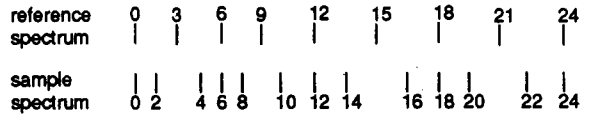
Such an appropriate program will quickly identify and parametrize the spectra under consideration after a simple matching and comparison process with the available library or data base.

PRACTICAL PROBLEMS ASSOCIATED WITH SPECTRA MATCHING

Various problems arise in the IR spectra matching process due to the resolution and/or instrument differences. These differences may cause one or more problems mentioned below.

1. A low resolution (20-cm^{-1}) spectrum will contain 180 points between the 4000 and 400 cm^{-1} interval. On the other hand, commonly used, high precision (1-cm^{-1}) resolution will give 1600 points in the same interval. In this case if the spectrum with more points is reduced simply by deleting, some information will be lost.

2. Even though the number of points between reference and source spectra are about the same, the points still may not coincide. For example consider the case where the reference spectrum is recorded at 3-cm^{-1} intervals and the source spectrum is recorded at 2-cm^{-1} intervals. This case can be illustrated as follows.



In this example even though we have a ratio of 2/3 between the number of points of the reference and the source spectrum, the actual ratio of matching points is about 1/3. This may not cause any problem with visual inspection. However, the computer program will have to use fewer points if some precautions are not taken regarding this problem.

3. Some spectra are recorded with different wavelength interval values. Due to instrumental constraints (grating or prism change for different wavelength regions) some portion of the spectrum (i.e. 4000–2000 cm^{-1}) may be recorded with a 5-cm^{-1} wavelength interval while the remaining portion (2000–625 cm^{-1}) has 20-cm^{-1} precision.

4. The noise level difference is a common problem associated with the instrumentation difference. If this is the case, then some sort of "soft smoothing" is needed for either source or reference spectrum or for both. By soft smoothing it is meant that just enough smoothing is done to reduce the noise considerably, without making any noticeable difference at absorption band shapes.

5. Tilting of the baseline is a common effect in IR spectra when absorption peaks are similar. If not enough precautions are taken, a full spectrum match of the tilted and flat spectra may result in a low similarity index.

6. Some thin solid films show interference patterns easily recognizable by visual inspection. These patterns usually give perfect sine curves separated by $\Delta\lambda = 1/nd$ where $\Delta\lambda$ is the wavelength difference between the adjacent maxima, n is the index of refraction, and d is the film thickness.

METHOD

A computer program is developed to deal with the above-mentioned spectra matching problems. All the algorithms which are applied to the source and reference spectra prior to the comparison in the process of identification are illustrated in Figure 1 in a simplified form. The explanations of the processes are given below. The entire computer program is coded in Pascal language, and various steps are prepared as separate units. Several control and checking steps are applied in the program. Among them are the following.

0095-2338/93/1633-0595$04.00/0

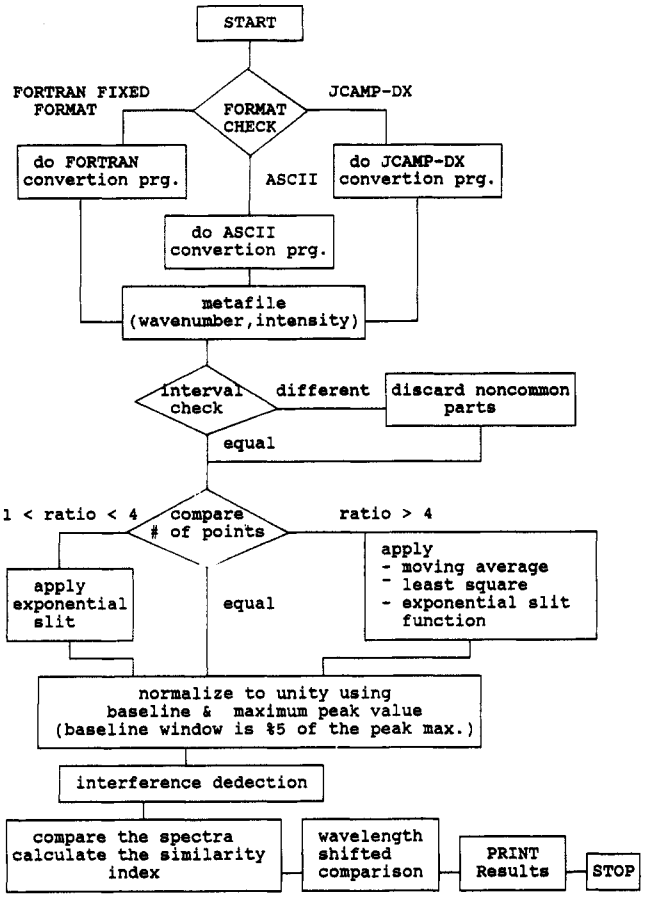


Figure 1. Simplified flow chart of the comparison process.

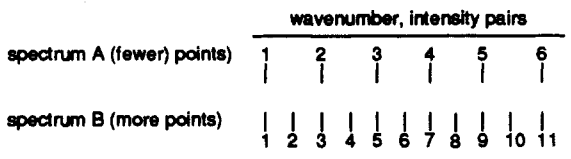
1. Data Format Checking. The source and reference spectra are converted to metafiles which consist of wavenumbers and corresponding intensity values. Data formats recognizable by the data conversion routines are fixed length FORTRAN format, standard ASCII format, and JCAMP-DX format.[1,2]

2. Wavenumber Interval Checking. If a region is not common in the reference spectrum and sample spectrum, then data in this region are not used in the comparison process. When one of the intervals is greater than the corresponding one, the program discards the data points belonging to the non-overlapping wavenumber region.

3. Number of Data Points Checking. The purpose of this subprogram is to check whether or not the ratio of the number of data points is the same in the subintervals. There are three possibilities for this: data points are equidistant and coincident ($r = 1$); the ratio of data points (r) is between 1 and 5, i.e., $1 < r < 5$; and the ratio of data points (r) is greater than 5, i.e., $r > 5$.

When data points are equidistant and coincident, no change is needed.

If the data point ratio is between 1 and 4, then one only needs to produce equidistant spectrums. In this case the unit program takes the spectrum with the lesser number of points and the other one is reduced to this in the following manner.



Here the ratio r equals 2.

The central points y_c, are computed by using the formula

$$y_c = (\sum_{i=1}^{n} y_i \exp(-|x_c - x_i|/\sigma))/\mathrm{NORM}$$

where y_c is the computed intensity value at the center of the sliding points, y_i is the intensity for every point belonging to the double interval centered at y_c, x_i is the corresponding wavenumber value, n is the number of points in the interval, and NORM is the normalization coefficient defined by

$$\mathrm{NORM} = \sum_{i=1}^{n} \exp((-|x_c - x_i|)/\sigma)$$

where σ is the FWHM (full width at half-maximum) value of the sharpest peak of the spectra.

For the data point ratio exceeding 4, three methods are applied separately: exponential slit function, sliding moving average, and polynomial least squares regression. Exponential slit function convolution is described in the preceding paragraphs.

Sliding moving average is the simplest method applied to the spectra. In this case y_c center intensity at x_c wavenumber is calculated as usual:[3]

$$y_c = (\sum_{i=1}^{n} y_i)/n$$

y_c is the calculated intensity at wavenumber x_c, y_i is the ith point intensity, and n is the number of points falling in this calculation interval.

For polynomial least squares calculation, five data points, one at center, two from the left, and two from right side of the central point (x_c, y_c), are taken and used in the least squares program unit. Coefficients of a third degree polynomial are returned. The intensity value y_c is calculated by

$$y_c = c_0 + c_1 x_c + c_2 x_c^2 + c_3 x_c^3$$

where c_0, c_1, c_2, and c_3 are the coefficients of regression and x_c is the wavenumber of y_c. Due to the precision constraints, not the actual wavenumbers x_i but the differences from the x_i are used in the actual calculation.

A normalization procedure is applied to all of the spectra. For this reason baseline detection is made and the difference between the maximum peak value and the baseline is calculated. The average baseline is subtracted from every y_i, and each y_i is divided by y_m (maximum intensity) as follows.

$$y_i \rightarrow (y_i - y_{\mathrm{baseline}})/y_m$$

For noise reduction the points which are 1, 2, 5, and 10% of the maximum value are discarded.

Later the spectrum is searched for interference effects. If a perfect sine is found, then this curve is used as the baseline for this region.

The spectrum comparison is made by

$$\mathrm{SQI} = (\sum_{i=1}^{n} (y_{1i} - y_{2i})^2/n)^{1/2}$$

where SQI is the similarity quality index (square root of Mahalanobis distance divided by n)[4] y_{1i} and y_{2i} are intensities of the source and reference spectra at wavenumbers x_i. n is the number of points. As seen from the formula above, identical spectra give a zero SQI value.

A calibration shift is a common case where different instruments are used. Identifying this problem is easy by visual inspection. When two spectra of the same material are

Table I. Diaminomaleonitrile Comparison Value

shifting value (cm^{-1})	similarity quality index		
	moving average	exponential slit	least squares
−30	23.56	23.60	23.46
−25	22.69	22.73	22.60
−20	21.78	21.82	21.69
−15	20.63	20.69	20.58
−10	19.20	19.26	19.16
−5	17.70	17.74	17.66
0	17.00	17.01	17.02
+5	17.57	17.58	17.58
+10	18.62	18.64	18.62
+15	19.50	19.53	19.52
+20	20.12	20.15	20.15
+25	20.73	20.75	20.78
+30	21.42	21.44	21.49

Table II. Polystyrene Comparison Values

shifting value (cm^{-1})	similarity quality index		
	moving average	exponential slit	least squares
−30	19.99	20.11	20.55
−25	18.53	18.64	19.05
−20	16.62	16.77	17.31
−15	13.92	14.10	14.72
−10	12.25	12.28	12.42
−5	15.60	15.52	15.45
0	20.12	20.15	20.46
+5	22.73	22.77	23.03
+10	24.12	24.18	24.45
+15	25.19	25.23	25.36
+20	26.43	26.47	26.64
+25	27.31	27.38	27.65
+30	27.90	28.97	28.31

viewed in a superposed form, one is seen shifted with respect to other spectrum. To check if there is this kind of shift in the spectra under consideration, one is held fixed, the corresponding sample or reference spectrum is shifted with steps of 5 cm^{-1}, and the corresponding SQI is calculated.

RESULTS AND DISCUSSION

For checking the program, the diaminomaleonitrile spectra used are taken from an IR atlas that is hand digitized (The Aldrich Library of IR Spectra)[5] and by an IR spectrometer (Perkin Elmer Model 983). Afterward the same process is applied to the IR spectra of polystyrene from the same IR atlas and by a BOMEM-FTIR Series 100 spectrometer.

The results are shown at Tables I and II. In these tables, the first column shows shifting values of the sample spectrum relative to the reference spectrum. The remaining columns show SQI values of moving average, exponential slit, and least squares corrected cases, respectively. As seen from the tables, all the smoothing processes give similar results. Because of its computing simplicity, the moving average method seems suitable for this process.

Data in the vicinity of the baseline (about 5%) are discarded for noise reduction. This process gave 5.14% reduction of SQI on the average, with a standard deviation value of 2.94. Discarding 10% gives a SQI reduction of 7.19%, with a standard deviation of 5.35. Discarding 10% or more is not preferred since it will cause a loss of the small peaks, and also the standard deviation is higher for this case.

Without doubt the wavenumber shifted comparison is dependent on the instruments used. In general it is questionable if we have the right to shift the spectra to be compared. There may be absorption peaks in the vicinity of the shifted value, so we may have erroneous results. A group of algorithms is developed/collected[6,7] for full spectrum matching, and a computer program is written employing these algorithms. It is shown that this kind of program is useful for comparisons of this sort. But the test data used are very limited. To gain better understanding of the workings of the program, a significant number of spectra need to be used.

Much of the digitizing of the reference is made by hand, which is very time consuming. Software for scanning of the spectra and recognizing the spectra from the scanned data would be extremely useful for this purpose.

REFERENCES AND NOTES

(1) McDonald, R. S.; Wilks, P. A. JCAMP-DX: A Standard Form for Exchange of Infrared Spectra in Computer Readable Form. *Appl. Spectrosc.* **1988**, *42*, 151–62.
(2) Gasteiger, J.; Hendriks, B. M. P.; Hoever, P.; Jochum, C.; Somberg, H. JCAMP-CS: A Standard Exchange Format for Chemical Structure Information in Computer-Readable Form. *Appl. Spectrosc.* **1991**, *45*, 4–11.
(3) Savitzky, A.; Golay, M. J. E. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Anal. Chem.* **1964**, *36*, 1627–39.
(4) George, W. O.; Willis, H. A. *Computer Methods in UV, Visible, and IR Spectroscopy*; Royal Society of Chemistry: London, 1990; pp 59–64.
(5) Pouchert, C. J. *The Aldrich Library of IR Spectra*, ed. III; Aldrich Chemical Co. Inc.: Milwaukee, WI, 1970; pp 514 E, 1593 F.
(6) Shah, N. K.; Gemperline, P. J. Combination of the Mahalanobis Distance and Residual Variance Pattern Recognition Techniques for Classification of Near-Infrared Reflectance Spectra. *Anal. Chem.* **1990**, *62*, 465–470.
(7) Proctor, A.; Sherwood, P. M. A. Smoothing of Digital X-ray Photoelectron Spectra by an Extended Sliding Least-Squares Approach. *Anal. Chem.* **1980**, *52*, 2315–2321.