AD-A055 277 INDIANA UNIV AT BLOOMINGTON DEPT OF CHEMISTRY F/G 7/4

COMPUTER IDENTIFICATION OF INFRARED SPECTRA BY CORRELATION-BASE--ETC(U)

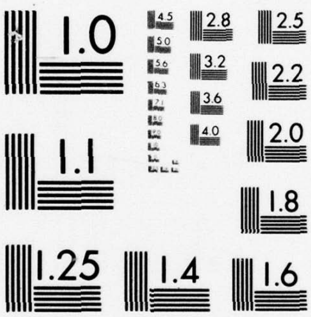
MAY 78 L A POWELL, G M HIEFTJE N00014-77-C-0444

UNCLASSIFIED TR-2 NL

| OF |

AD
A055277

END
DATE
FILMED
7 -78
DDC


1.0
4.5
5.0
5.6
6.3
7.1
2.8
3.2
3.6
4.0
2.5
2.2
2.0
1.1
1.8
1.25
1.4
1.6

12

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUC... BEFORE COMPLET...

1. REPORT NUMBER TR- TWO	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) (6) COMPUTER IDENTIFICATION OF INFRARED SPECTRA BY CORRELATION-BASED FILE SEARCHING,		5. TYPE OF REPORT & PERIOD COVERED
		6. PERFORMING ORG. REPORT NUMBER TR-2
7. AUTHOR(s) (10) Linda A. Powell [redacted] G. M. Hieftje		8. CONTRACT OR GRANT NUMBER(s) ~~N14~~-77-C-0444 N00014
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Chemistry Indiana University Bloomington, Indiana 47401		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS NR 051-659
11. CONTROLLING OFFICE NAME AND ADDRESS Office of Naval Research Washington, D.C. (11) 5 May 78		12. REPORT DATE May 5, 1978
		13. NUMBER OF PAGES 29
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office) (15) NØØØ14-77-C-Ø444 (12) 29p.		15. SECURITY CLASS. (of this report) UNCLASSIFIED
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

SC

16. DISTRIBUTION STATEMENT (of this Report) (14) TR-2

Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

Prepared for publication in ANALYTICA CHIMICA ACTA

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

correlation, infrared spectrum, computer file searching

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

A new method for the computerized search and identification of infrared spectra has been developed and evaluated. Based on cross correlation, the search system utilizes all spectral information in a digitized spectrum when it attempts to match an unknown spectrum to one in a small library of known spectra. To evaluate a spectral match, the search program calculates the cross correlation function between the

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

176685

Block #20
(continued)

unknown and known (library) spectra which indicates their degree of similarity and allows library spectra to be ranked in order of probability of match to the unknown spectrum.

In this study, several small infrared spectral libraries of structurally similar compounds were searched under conditions which examined the sensitivity of the search method to chemical and instrumental variations. Because the correlation technique is slower than conventional file-searching methods, it will probably find greatest use in the search of small collections of similar spectra or as a match-ranking procedure following preliminary selection by a faster search method.

COMPUTER IDENTIFICATION OF INFRARED SPECTRA BY CORRELATION-BASED FILE SEARCHING

Linda A. Powell and G. M. Hieftje*

Department of Chemistry
Indiana University
Bloomington, Indiana 47401

*Author to whom correspondence should be sent.

BRIEF

A method to search small libraries of digitized infrared spectra is described which is based on calculation of the cross correlation function between unknown and library spectra.

ABSTRACT

A new method for the computerized search and identification of infrared spectra has been developed and evaluated. Based on cross correlation, the search system utilizes all spectral information in a digitized spectrum when it attempts to match an unknown spectrum to one in a small library of known spectra. To evaluate a spectral match, the search program calculates the cross correlation function between the unknown and known (library) spectra which indicates their degree of similarity and allows library spectra to be ranked in order of probability of match to the unknown spectrum.

In this study, several small infrared spectral libraries of structurally similar compounds were searched under conditions which examined the sensitivity of the search method to chemical and instrumental variations. Because the correlation technique is slower than conventional file-searching methods, it will probably find greatest use in the search of small collections of similar spectra or as a match-ranking procedure following preliminary selection by a faster search method.

Computerized spectral file searching has become an important tool in many analytical laboratories. In such procedures, spectra are ordinarily encoded in some fashion and a file-searching algorithm is used to find the best match between the spectrum of an unknown substance and one in a collection of spectra of known compounds. File searching procedures have been reviewed in several recent articles (1-5).

In general, the goal of file searching methods is to provide unambiguous identification of an unknown spectrum, but to exhibit relative immunity from instrumental artifacts and variations and from the influence of minor impurities or contaminants. Clearly, these goals are mutually exclusive and most procedures attempt to achieve a compromise between them. To simplify a search, an unknown spectrum is ordinarily encoded in some form before being compared with library spectra (6). Unfortunately, this encoding necessarily results in a loss in spectral information, a situation which is particularly undesirable in infrared spectra where peak width and shape are often important characteristics.

In contrast to conventional search methods, correlation-based search techniques require no special encoding process, but employ a <u>complete</u> digitized spectrum, thereby enabling all spectral information to be utilized. The utility of this approach has been demonstrated for identification of ultraviolet absorption spectra (7), atomic emission spectra (8), infrared spectra (9), and γ-ray spectra generated after neutron activation (10). In these applications, correlation has proven to be a useful search technique which is relatively free of influence from instrumental artifacts and sample contamination.

Because correlation is a relatively sophisticated mathematical operation, these past approaches have sought to reduce the information provided by the correlation procedure to a single parameter, expressing the similarity between unknown and library spectra. Often, such "correlation coefficients" are themselves derived from only a small fraction of the available spectral features and are therefore of little more utility than parameters derived from conventional coded searches. In the present study, complete cross correlation functions are calculated between unknown and library spectra; moreover, a large fraction of the resulting function is employed to calculate a "correlation parameter" which is shown to be highly reliable in its ability to differentiate between similar but nonidentical spectra and to correctly identify unknown spectra. In addition, the new correlation search method is shown to be relatively free from the effects of instrumental variation and chemical contamination, making it amenable to use in routine laboratories and with spectral libraries obtained on older instruments or those with slightly different characteristics. Because the technique requires computation of an entire correlation function, it is expected to be somewhat slower than competitive techniques and is therefore expected to find greatest use as a post-searching method to rank probable matches which have been selected by a faster but less accurate method. In addition, the correlation search is predicted to be most useful in the search of spectra produced in Fourier transform infrared spectrometry, because of the utility of the Fourier transform to correlation computation.

CORRELATION-BASED FILE SEARCHING

Correlation involves the evaluation of the averaged product of two signals (here, two infrared spectra) as a function of their relative displacement from each other (12). The magnitude of the cross correlation function at zero displacement indicates the degree of similarity between the two spectra; the greater the number of common spectral features, the greater will be the value. This is illustrated by the cross correlation functions (correlograms) shown in Figure 1. Consequently, the correlation function enables library spectra to be ranked in order of probability of match to an unknown spectrum.

In general, the cross correlation function $C_{ab}(\tau)$ of two real waveforms a(t) and b(t) is represented exactly by the integral

$$C_{ab}(\tau) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{+T} a(t)b(t\pm\tau)dt \qquad (1)$$

where τ is the relative displacement between the two waveforms (12). In computer calculations correlation is often approximated by the following sum, termed the discrete correlation (8).

$$C_{ab}(n\Delta t) = \frac{1}{T}\sum_{t=o}^{T} a(t)b(t\pm n\Delta t) \qquad n = 0,\ 1,\ 2,\ \ldots\ \frac{T}{\Delta t} \qquad (2)$$

Here, Δt is the sampling interval and $n\Delta t$ is the relative displacement between a(t) and b(t).

The correlation of two functions can also be computed through a Fourier transform method (12). According to the discrete correlation theorem, the following two functions form a Fourier transform (FT) pair:

$$C_{ab}(\tau) \overset{FT}{\longleftrightarrow} A^*(s)\ B(s) \qquad (3)$$

where A(s) and B(s) are the Fourier transforms of a(t) and b(t). Thus, the cross correlation $C_{ab}(\tau)$ is equal to the inverse transform of the product of A*(s) and B(s), where A*(s) is the complex conjugate of A(s). It is most convenient and efficient to use the fast Fourier transform algorithm (FFT) to compute the discrete Fourier transform for this correlation method.

Avoiding "End Effects". To perform a discrete correlation of two waveforms of finite duration such as IR spectra, one must sample both waveforms and thus impose upon each of them an artificial period (N) determined by the length of the sampling window (total spectrum). Under such conditions, the correlation function can be distorted by "end effects" which occur as one waveform is shifted with respect to the other; points which should not overlap are then multiplied by points in the next period of the other waveform. In computing the discrete correlation sum (Eq. 2), end effects can be eliminated by multiplying non-overlapping points by zero. Similarly, for the Fourier transform method, no unwanted overlap will occur if the period N is enlarged by adding zeroes to one end of each finite-duration waveform with a number of points P and Q respectively such that $N \geq P + Q - 1$ (11). If P = Q, the addition of zeroes will produce two new waveforms, each of which has twice as many points as the original.

Significantly, Bendat and Piersol (13) have shown that end effects do not strongly influence values of the correlation near zero displacement (τ=0), provided that the correlation function decays rapidly. Because the search system reported here is concerned only with values of the correlation function near τ=0, end effects can be ignored without affecting the reliability of the search. Consequently, computer storage space need only be adequate to store each digitized spectrum with no added zeroes.

EXPERIMENTAL

A block diagram of the experimental apparatus used to measure infrared absorption spectra, store them on magnetic tape, and perform the correlation-based searches is shown in Figure 2. A conventional infrared spectrophotometer (Model IR5, Beckman Instruments Co., Fullerton, CA) is interfaced to a minicomputer (PDP-12/40, Digital Equipment Corp., Maynard, MA) for data-logging through a 10-bit analog-to-digital converter. These data are then transported to a large, remote, high speed computer (Model 6600, Control Data Corp., Minneapolis, MN) which calculates correlation functions and performs library searches. Spectral data can be transferred between the two computers by means of punched paper tape or electronically through an intermediate, medium-scale computer (Sigma 2, Xerox Data Systems, El Segundo, CA).

Spectral Libraries. Table I lists the compounds in each of the three small spectral libraries compiled for this study. Samples (all liquids) were obtained from a variety of sources but all were of reagent grade or better purity. Experimental spectra of these reagent grade compounds were found to be visually indistinguishable from literature spectra of the same compounds.

Each library contains a number of very similar spectra produced by compounds with similar structures. However, several compounds such as 2,2-dimethoxypropane and benzaldehyde, whose spectra show strong peaks that are not present in other spectra in their libraries, were included to determine what effect such anomalies would have on the correlation-based searches.

Procedures and Search Algorithms. All spectra were digitized at a rate of one point per second for a total of 512 points, the digitization rate being fast enough to preserve all spectral features such as shoulders and very narrow peaks. Only the wavelength region between 5.5 μm and 14.3 μm, which includes the "fingerprint region," was digitized and stored by the computer. The figerprint region (7 to 12 μm) contains bands which are vibrations of the molecule as a whole and are considered to be specific for particular compounds. Spectra were permanently stored in magnetic tape libraries at the remote CDC 6600 computer, where the searches were also performed.

In the first general trials of the correlation-based search techniques, each library spectrum was treated in turn as an unknown and the library was searched for a match. For later tests, other spectra that resulted from mixtures, solutions, compounds with impurities, etc., were added to the various libraries and treated as unknowns. In these latter cases, only the original core libraries, not the additional spectra, were searched. It should be noted that an unknown spectrum does not actually have to be added to the library before a search is performed.

Two alternative FORTRAN IV programs were used to calculate the correlation functions between spectra; one program uses a computational technique to determine the correlation while a second method utilizes the fast Fourier transform approach. After the cross correlation patterns have been calculated, the best match is determined by finding the pair of spectra which gives the greatest "correlation parameter" (see below). The computational program which uses the "covariance" technique first calculates the mean value of the spectrum, then subracts that mean from each data point (14). This subtraction, then, gives each spectrum a zero average, and is the only way in which covariance differs from discrete correlation. This similarity can be appreciated from the formulae used to compute the cross covariance of two functions, a(t) and b(t), after subtracting the means:

$$C_{ab}(\tau) = \frac{1}{n - \tau} \sum_{t+1}^{n-\tau} a(t)\ b(t + \tau),\ \tau = 0,1,2,\ldots,m \qquad (4)$$

$$C_{ab}(-\tau) = \frac{1}{n - \tau} \sum_{t=1}^{n-\tau} a(t + \tau)\ b(t),\ \tau = 0,1,2,\ldots,m \qquad (5)$$

τ is the displacement between the two functions, m is the maximum displace- and n is the number of points in each function. The summation limit and normalization factor $(n - \tau)$ account for the fact that during the covariance calculation all points in the two spectra overlap completely only when $\tau = 0$. At other values of τ the non-overlapping points are multiplied by zero and end effects are eliminated. The value of the cross correlation at any value of τ is equal to the sum of all the products divided by the number of products, where the number of products is the number of overlapping points.

The fast Fourier transform algorithm used in the second method is based on that of Brigham (11). For this search method, the spectral libraries contain the normalized Fourier transforms of the original spectra rather than the spectra themselves.

Correlation Parameter. For both methods, the best match for an unknown spectrum should be the library spectrum whose cross correlation with the unknown yields the highest value at zero τ. In practice, this expectation was not usually met, partly because of variations in overall intensity or background level among the library spectra. Also, most spectra in a library have several bands in common, which results in the appearance of a peak at or near zero delay in many cross correlation patterns. However, a simple weighting or "normalization" procedure can be used to increase the likelihood that the true match is the one whose cross correlation value at zero τ is greatest. This weighting procedure involves finding the mean value of the cross correlation for the range $-10 \leq \tau \leq 10$ and subtracting this mean from each point in that range. The value of the resulting function at zero τ is then an excellent indicator of the similarity between two spectra and is termed the correlation parameter in the following discussion.

RESULTS AND DISCUSSION

Correlation parameters calculated during one typical search are shown graphically in Figure 3, where toluene was chosen as the unknown and the spectral library of aromatic compounds was searched for a match by the covariance and Fourier transform methods. In both cases the greatest correlation parameter value occurs when the "unknown" spectrum is compared to the library spectrum of toluene; the next highest correlation parameter occurs for the match with ethylbenzene. In all searches, the ratio of the greatest to the next greatest correlation parameters served as the most reliable indicator of a spectral match. In most cases, such ratios (hereafter termed "match ratios") resulting from the covariance and Fourier transform methods are nearly the same, even though the two methods do not yield the same numerical result for the calculation of any particular correlation function.

The results of all other searches of the three libraries under the conditions just described are tabulated in terms of this match ratio in Table I. In T able I, each compound in turn was treated as an unknown and its particular library searched for a match; match ratios were calculated between this best match (the one yielding the greatest correlation parameter) and the next best. In all cases, both correlation-based spectral search methods were successful in picking the correct match for an "unknown." In addition, for every unknown except chlorobenzene, the two search methods chose the same compounds for the second best match (the compound with the second highest correlation parameter). For chlorobenzene, the covariance method ranked bromobenzene and o-xylene as the second and third best matches, respectively; in contrast, the Fourier transform method ranked o-xylene second and bromobenzene third. For both methods, the absolute values of the second and third choice parameters were within 1.4% of each other, so the difference in ranking might not be significant.

The correlation-based search methods are able to distinquish compounds with very similar spectra and compounds that are geometrical isomers. For instance, the infrared spectra of ethanol and 2-(2-ethoxyethoxy)-ethanol are nearly the same except for some differences in the band structure near 1050 cm^{-1} and a small peak at 925 cm^{-1} that occurs only in the ether spectrum. When ethanol is treated as an unknown, the ether is the second best match, and vice versa. The match ratios which result from the searches are shown in Table I. The correlation of the ethanol spectrum with itself gives a correlation parameter that is more than three times as great as that from the correlation of the ether with itself. However, the correlation parameter between ethanol and the ether is still less than either of the parameters from cross correlation of a spectrum with itself, so the correct match is chosen for both compounds. Other similar pairs of compounds, such as o-xylene and m-xylene, or methyl ethyl ketone and methyl isobutyl ketone, were also readily distinguished by the search methods.

Effect of Chemical Parameters on Match Reliability. The ability of both correlation-based search techniques to match identical spectra was excellent -- both methods always chose the correct match for all trials. In practice, however, an unknown spectrum is seldom identical to one of the library spectra. Impurities present in the compound itself or in solvents can distort either or both spectra; also, the unknown might be a mixture of compounds so that its spectrum would not exactly match that of any single library compound. Various tests of the spectral search systems were performed to determine the significance of chemical variations on the reliability of the correlation-based search.

To examine the effect that an impurity might have on a typical search, a sample of 2-octanol with an unidentified carbonyl-containing impurity was used as an unknown, and the alcohol library was searched. The spectrum of the impure sample showed a distinct band at 1700 cm^{-1} which was absent from the spectrum of pure 2-octanol. Despite this extraneous band, both the covariance and Fourier transform search methods identified the unknown as 2-octanol. In both searches isopentanol was chosen as the second best match.

Further tests were performed to determine the amount of impurity that could be tolerated before the search systems chose the wrong match and to study the ability of the search systems to resolve the components of a mixture. The first such test employed a mixture of equal volumes of pure 1-octanol and pure 2-octanol. When the alcohol spectral library was searched for a match for the mixture, both search methods listed isopentanol -- a compound not present in the "unknown" mixture -- as the best match. However, 1-octanol and 2-octanol were selected as the next best matches. This erroneous selection is not surprising, considering the various spectra involved. The spectrum of isopentanol has strong peaks which match most of the strong peaks in the mixture spectrum. Obviously, the spectra of 1-octanol and 2-octanol are each missing some of the strong peaks which arise from the other compound. Isopentanol therefore appears to be a better match for the mixture than either of the true components. The effect of overall spectral amplitude may bias the results toward the stronger absorbing compound (here, isopentanol) as well, a possibility which will be examined in a later section.

Several mixtures of o-xylene and m-xylene were employed in further studies which sought to determine the response of the methods to mixtures. The studies were patterned after those of Tanabe and Saëki (9). In these experiments, the concentrations of o-xylene and m-xylene in a mixture of the two varies from 10% ortho + 90% meta to 90% ortho + 10% meta by volume. The correlation functions of the spectrum of each mixture with those of pure ortho and pure meta were then calculated. Plots of the correlation parameter versus percent purity of each component are presented in Figures 5A and 5B for the covariance and Fourier transform methods, respectively. The magnitude of the correlation parameter increases with increasing purity of the unknown sample. The variation in our values probably results from an irreproducible path length with the salt discs employed; Tanabe and

Saëki (9) used sealed cells, and noted that an understandably strong dependence of their correlation coefficients on sample thickness existed when transmittance rather than absorbance was recorded.

The quantity of impurity that can be tolerated by the search systems in this particular case can be estimated by treating each xylene mixture as an unknown sample and searching the entire library of aromatic spectra. In this test, the proper compound is identified as the primary component of the mixture when the solution is at least 60% pure. Also, for a 60% pure solution, both search systems rank the impurity as the second best match. As a preliminary estimate, the sample should be greater than 60% pure to ensure that the proper compound will be selected by the searches. Tanabe and Saëki (9) required that the purity of ortho- and meta-xylene be at least 80% and 96% respectively, in order to obtain correlation coefficients greater than 0.95 -- their criterion for a match. The value of their correlation coefficient is independent of the other compounds in the library, whereas the match *ratio* found by the correlation-based techniques depends on the correlations of the unknown with the other library spectra. Accordingly, studies with more and different spectra in the libraries and with sealed cells should be performed to verify the estimate given here.

Effect of Instrumental Parameters on Match Reliability. *Spectral Amptitude.* It has been suggested that sample path length can greatly affect the value of the correlation parameter when spectra are recorded in terms of transmittance (9). This result arises because a longer path length produces more intense peaks -- that is, peaks with lower percent transmittance according to the logarithmic relationship described by the Beer-Lambert Law, $A = -\log T = \varepsilon bc$, where A is absorbance, T is transmittance, ε is molar absorptivity, b is path length and c is concentration. The effect on the correlation function will be less notable for spectra recorded in terms of absorbance (A) because correlation, by its nature, is not sensitive to linearly proportional variations in a waveform.

In the present investigation, the use of the "correlation parameter" overcomes most of the influence of spectral amptitude, including that portion arising from sample path length changes. A simple test of the search system substantiates this success. Spectra of methyl ethyl ketone and methyl isobutyl ketone with overall amplitudes differing by 20 to 30% T were measured; when each spectrum was subsequently treated as an unknown, the correct match was chosen. The value of the match ratio tends to increase with spectral amplitude even after the correction is performed, but the correction method nevertheless allows both search methods to select the proper match. Further work is necessary with respect to this problem in order to optimize the search speed and reliability. The need for a correction might be obviated by use of absorbance instead of transmittance values, but most conventional infrared instruments record only in %T so an extra computational step would be required.

Wavelength Shift. Shift and imprecision of the wavenumber axis is a common problem with most infrared spectrophotometers. A precision of .03 μm (approximately 3 cm^{-1} at 10.0 μm) is typical for a spectrophotometer such as the Beckman IR-5 employed in this study; furthermore, precision usually deteriorates as the instrument ages. Such a shift could adversely affect a search if the wavenumber axis of the unknown spectrum were different from that of the library spectrum. Some search methods incorporate a "wiggle" option in an attempt to compensate for wavenumber shift (6, 15).

Correction for wavenumber imprecision is readily accomplished in correlation-based search techniques. If the unknown spectrum and its mate in the library do not have identical axes, the maximum value of the correlation which should occur at zero τ will be displaced to another value of τ, as illustrated in Figure 5. The correct match can then be found by searching for the correlation maximum in the vicinity of zero τ. In both our covariance and Fourier transform methods, the user can instruct the program to examine the region $-5 \leq \tau \leq 5$ for a maximum. The interval of $\pm 5\tau$ corresponds approximately to ± 7.5 cm^{-1}, which should be more than adequate to accommodate any reasonable wavenumber shift. Such a shift compares favorably to the 3 cm^{-1} tolerance required by Tanabe and Saëki (9).

The immunity of the search methods to spectral shift was tested using four shifted spectra as unknowns. The digitized spectra of isopropanol and 1-octanol were mathematically shifted by three sampling intervals to simulate the instrumental effect (three sampling intervals corresponds to about 4.5 cm^{-1}). The programs looked for a maximum in the range $-5 < \tau < 5$ for each correlation of an unknown with a library spectrum, and the maximum values were then ranked to determine the best match. Match ratios for this test can be found in Table II. The maximum value in the correlation function calculated between the shifted spectrum and its library counterpart was always found at $\pm 3\tau$. The value for the correlation of 1-octanol and isopentanol which produced a mismatch was located at -5τ. When a mismatch existed in the alcohol library, isopentanol was most often chosen, probably because of its great spectral amplitude.

Influence of Noise. The instrumentation used to record, digitize and store the infrared spectra introduces noise which could affect the success of the correlation-based searches. Figure 6 is an example of a digitized spectrum with a number of spurious values that arose from a dirty potentiometer in the spectrophotometer-minicomputer interface circuitry. The noise appears randomly in time and at smaller percent transmittance values than the neighboring points in the spectrum.

To test the susceptibility of the correlation-based search systems to this rather severe noise problem, a number of noisy ketone spectra were used as unknowns. The ketone library contains similar spectra but which have been rendered noise free by substituting the average of two neighboring points for noisy data points. Match ratios obtained in this study are given in Table III. Correct matches were chosen in all searches except one -- a methyl isobutyl ketone spectrum. However, for this mismatched spectrum, the values of the correlation parameter for the best and second best matches were not significantly different. Another noisy methyl isobutyl ketone spectrum with greater overall spectral amplitude was matched correctly when it was used as the "unknown"; this spectrum's noise-averaged version was actually the library spectrum.

CONCLUSIONS

In most respects the covariance and Fourier transform approaches to spectral search and identification are comparable and yield accurate spectral matches. The correlation functions calculated by the Fourier transform method are susceptible to overlap or end effects unless a correction is included. However, a simple test has shown that neglect of end effects does not harm search reliability. The Fourier transform method is the faster of the two, but it requires twice as much computer storage space as covariance because the libraries for the FT method contain the complex Fourier transforms of the original spectra. So, for the correlation-based search as it is now configured, the choice of the covariance or Fourier transform method is a matter of the user's preference.

Several aspects of the search system could be optimized to provide faster, more efficient searches. One possibility is to completely automate the search from the time the user instructs the computer to read an unknown spectrum to the time the final best match is chosen and printed out. Data-logging, storage, and actual computation could be performed on a single small computer, if the user were careful not to overflow core memory. The practical speed of computation could be improved by using a more efficient Fourier transform algorithm. For example, it is possible to compute the FFT of two real functions simultaneously (11).

A number of other experiments could be performed to explore the utility of the correlation methods for spectral file searching. The ability of the system to handle infrared spectra from other types of infrared instruments (particularly Fourier transform instruments) should be examined. Presumably, normalization of the spectra would then be necessary because infrared instruments differ in wavelength program, spectral response, and other features.

Significantly, the correlation-based search which uses the Fourier transform method should be readily applicable to Fourier transform infrared spectroscopy (FTIR). The nature of the FTIR technique itself and the attendant minicomputer which is an integral part of FTIR instrumentation would allow a number of time-saving modifications of the search process to be implemented. Also, because FTIR instruments produce interferograms which are already the Fourier transforms of infrared spectra, computation of the cross correlation function between two spectra can be accomplished with only a single extra step, consisting of simple point-by-point multiplication. In fact, it can be shown that a value proportional to our "correlation parameter" can be obtained simply by multiplying the two appropriate interferograms and subtracting the resulting zero retardation intensity from the average intensity of the product waveform. This capability will be the subject of a future communication.

Because the correlation-based search techniques utilize all spectral information in a digitized spectrum, they are better able to match or rank an unknown spectrum with those in a library of known spectra. However, most conventional search methods are considerably faster than the correlation techniques and are therefore more suitable for searching large collections of spectra. Consequently, correlation techniques are best used with small spectral collections such as a library of similar compounds of interest to the user or to compile a list of most likely matches which result from a conventional search of a large library.

In general, the correlation-based searches appear to be more sensitive to changes in relative peak position and peak shape than to changes which result from instrumental variations or the presence of impurities. So, they are responsive to spectral differences that are indicative of structural differences, yet they are relatively immune to undesirable experimental variations. This reasonable compromise between high selectivity and freedom from effects of experimental artifacts is one of the greatest strengths of the correlation search techniques.

REFERENCES

1. L. H. Gevantman, *Anal. Chem.*, **44** (7), 30A (1972).

2. R. S. McDonald, *Anal. Chem.*, **46**, 521R (1974).

3. R. S. McDonald, *Anal. Chem.*, **48**, 196R (1976).

4. J. Zupan, M. Penca, D. Hadzi, and J. Marsel, *Anal. Chem.*, **49**, 2141 (1978).

5. L. E. Wangen, W. S. Woodward, and T. L. Isenhour, *Anal. Chem.*, **43**, 1605 (1971).

6. E. C. Penski, D. A. Padowski, and J. B. Bouch, *Anal. Chem.*, **46**, 955 (1974).

7. J. C. Reid and E. C. Wong, *Appl. Spectrosc.*, **20**, 320 (1966).

8. G. Horlick, *Anal. Chem.*, **45**, 319 (1973).

9. K. Tanabe and S. Saeki, *Anal. Chem.*, **47**, 118 (1975).

10. R. MacDonald, A. Robertson, T. J. Kennett, and W. V. Prestwich, *J. Radioanal. Chem.*, **23**, 123 (1974).

11. E. O. Brigham, "The Fast Fourier Transform", Prentice-Hall, Englewood Cliffs, N. J., 1974.

12. G. Horlick and G. M. Hieftje in "Contemporary Topics in Analytical and Clinical Chemistry," Vol. 3, D. M. Hercules, G. M. Hieftje, L. R. Snyder, M. A. Evenson, eds., Plenum Press, New York, N.Y. 1978.

13. J. S. Bendat and A. G. Piersol, "Random Data: Analysis and Measurement Procedures", Wiley-Interscience, New York, N.Y. 1971.

14. "BMD2T - Autocovariance and Power Spectral Analysis", available from Wrubel Computer Center, Indiana University, Bloomington, In.

15. H. B. Woodruff, S. R. Lowry, and T. L. Isenhour, *J. Chem. Inf. Comput. Sci.*, **15**, 207 (1975).

CREDIT

Supported in part by the National Institute of Health through grant GM17904

TABLE I . MATCH RATIOS RESULTING FROM CORRELATION-BASED SEARCHES

Unknown spectrum	Match Ratios	
	Covariance Method	Fourier Transform Method
Library A. Alcohols		
methanol	2.93	2.24
ethanol	3.83	3.39
1-propanol	1.88	1.78
isopropanol	5.82	4.94
t-butanol	6.12	5.80
1-pentanol	1.43	1.34
isopentanol	2.73	2.06
1-octanol	1.32	1.26
2-octanol	1.79	1.17
2-(2-ethoxyethoxy) ethanol	1.04	1.01
Library B. Ketones		
acetone	1.94	1.99
methyl ethyl ketone	1.22	1.27
methyl isobutyl ketone	1.29	1.34
cyclohexaone	1.34	1.36
2,4-pentanedione	1.81	1.95
pentanal	1.64	1.74
2,2-dimethoxypropane	8.16	8.01
Library C. Aromatics		
benzene	4.15	4.43
toluene	1.45	1.43
ethylbenzene	1.50	1.56
isopropylbenzene	1.80	1.69
o-xylene	2.11	2.13
m-xylene	2.77	2.51
benzaldehyde	4.73	4.45
cyclohexylbenzene	1.53	1.46
chlorobenzene	2.54	2.54
bromobenzene	2.81	2.83

TABLE II. MATCH RATIOS FOR SPECTRA WITH SHIFTED WAVENUMBER AXES

Spectra shifted 3 sampling intervals ($\sim$4.5 cm^{-1}) to greater wavenumber

	Covariance Method	Fourier Transform Method
isopropanol	1.83	1.91
1-octanol	1.09	1.16

Spectra shifted 3 sampling intervals ($\sim$4.5 cm^{-1}) to smaller wavenumber

	Covariance Method	Fourier Transform Method
isopropanol	1.77	1.85
1-octanol	Mismatched with isopentanol	

TABLE III. MATCH RATIOS FOR UNKNOWNS WITH NOISY SPECTRA

Unknown Spectrum	Match Ratios	
	Covariance Method	Fourier Transform Method
acetone	2.06	2.01
methyl ethyl ketone	1.33	1.32
methyl isobutyl ketone	1.00	1.00
methyl isobutyl ketone*	1.36	1.34
ciclohexanone	1.41	1.37
2,4-pentanedione	1.70	1.61
2,4-pentanedione*	1.86	1.90
pentanal	1.73	1.83
2,2-dimethoxypropane	8.19	8.23

*These spectra had greater overall spectral amplitude than the other spectrum of the same compound.

FIGURE CAPTIONS

Figure 1. Cross correlation function between the spectra of A) methyl isobutyl ketone and itself, B) methyl isobutyl ketone and methyl ethyl ketone. Vertical axis in arbitrary units. Tau ≡ τ, the delay parameter in the correlation function.

Figure 2. Block diagram of experimental arrangement for identification of infrared spectra.

Figure 3. Results of searches by covariance (discrete correlation) and Fourier transform methods for the spectrum of toluene. The magnitude of the corrected correlation function at zero τ (correlation parameter) is plotted for the correlation of the unknown (toluene) with each library spectrum. O covariance method. Δ Fourier transform method.

Figure 4. Values for the correlation parameter calculated by A) the covariance method and B) the Fourier transform method between the spectra of pure o-xylene and various mixtures of impure o-xylene (O) and pure and impure m-xylene (Δ). The concentration is in terms of percent purity. For example, for m-xylene, a solution of 80% purity contains 80% m-xylene and 20% o-xylene by volume.

Figure 5. Effect of wavelength shift on the correlation search procedure. A shift in wavelength axes between unknown and library spectra causes a shift of the correlation maximum away from zero τ. A correct match can be found by searching for the correlation maximum in the vicinity of zero τ. A and B are the original and shifted infrared spectra of methyl isobutyl ketone. C and D show the shift in correlation functions which results from this spectral shift.

Figure 6. Noisy digitized infrared spectrum of methyl isobutyl ketone used to test the response of the search methods to unknowns with noise (circled points).