systems, that of a *good* user interface, also occurs with database use. In neither case have user interfaces been built that really are easy to use. Progress has to be made in this area.

There is no doubt that the impact of expert systems on chemistry will be huge in the future. Part of the impact clearly will be in the use of chemical databases.

SOLID MATERIALS

With the exception of diffraction data, very little work has been done on databases of the properties of solid materials. Yet data on the characterization of surfaces, catalytic properties, corrosion properties, and other areas of solid-state chemistry are very important. The Standard Reference Data program at NBS has begun work on databases of ESCA information, as well as chemical-stability diagrams for corrosion-prediction purposes. However, much more effort is needed in this area. Two examples can indicate the importance of these data. The reliability of microelectronic circuits or "chips" depends on the reactivity of the exposed surfaces to chemicals in their environment. Such microchemistry is very important, yet no organized activity to collect these data is under way. The same type of microchemistry is also important in catalysis, yet again these data are not being systematically collected, evaluated, and made available in databases.

SUMMARY AND CONCLUSIONS

We have outlined in the discussion above some of the future directions of chemical databases. Clearly, many developments will take place, and it is difficult in every case to pinpoint exactly how fast progress will be made. It is equally clear that we can anticipate chemical databases as a dynamic area of development which will eventually lead to the computer revolution in chemical information that has been predicted. Hopefully, the reader will take up some of these challenges and help make computer access to chemical information an everyday reality.

REFERENCES AND NOTES

(1) Hampel, V. E.; Bollinger, W. A.; Gaynor, C. A.; Oldani, J. J. "An Online Directory of Databases for Material Properties"; Lawrence Livermore National Laboratory: Livermore, CA, 1984; UCRL-90276 Rev. 1.
(2) Rumble, J. R., Jr. "Why Can't We Access More Numeric Data via Computers". In "Proceedings of the Fifth National Online Meeting"; Williams, M. E.; Hogan, T. H., Eds.; Learned Information Inc.: Medford, NJ, 1984; p 325.
(3) Lide, D. R., Jr. "Critical Data for Critical Needs". *Science (Washington, D.C.)* **1981**, *212*, 1334-40.
(4) "Standard Reference Data Publications 1964-1980"; Sherwood, G. B., Ed.; U.S. Department of Commerce: Washington, DC, 1981; NBS Special Publ. 612.
(5) "Standard Reference Data Publications 1981-1982 Supplement"; Sherwood, G. B., Ed.; U.S. Department of Commerce: Washington, DC, 1983.
(6) "Workshop on Data Quality Indicators—Summary Report and Recommendations"; Chemical Manufacturers Assoc.: Washington, DC, 1982.
(7) Further information on these meetings can be obtained by contacting: Dr. Henry Kehiaian, Universite Paris VII-CNRS, Institut de Topologie et de Dynamique des Systems, 1 rue Guy de la Brosse, 75005 Paris, France.
(8) Himes, V. L.; Mighell, A. D. "A Matrix Method for Lattice Symmetry Determination". *Acta Crystallogr., Sect. A: Cryst. Phys., Diffr., Theor. Gen. Crystallogr.* **1982**, *A38*, 748-749.
(9) Mighell, A. D.; Himes, V. L.; Rodgers, J. R. "Space Group Frequencies for Organic Compounds". *Acta Crystallogr., Sect. A: Found. Crystallogr.* **1983**, *A39*, 737-740.
(10) Ely, J. F.; Hanley, H. J. M. "Prediction of Transport Properties. I. Viscosity of Fluids and Mixtures". *Ind. Eng. Chem. Fundam.* **1981**, *20*, 323.
(11) Ely, J. F.; Hanley, H. J. M. "Prediction of Transport Properties. II. Thermal Conductivity of Fluids and Mixtures". *Ind. Eng. Chem. Fundam.* **1983**, *22*, 90.
(12) Dessy, R. E., Ed. "Expert Systems Part I". *Anal. Chem.* **1984**, *56*, 1200A-1212A.
(13) Dessy, R. E., Ed. "Expert Systems Part II". *Anal. Chem.* **1984**, *56*, 1312A-1332A.

Data Base Development and Search Algorithms for Automated Infrared Spectral Identification

S. R. LOWRY,* D. A. HUPPLER, and C. R. ANDERSON

Nicolet Instrument Corporation, Madison, Wisconsin 53711

Received February 19, 1985

Specifications and sampling methods for infrared spectral data acquisition are presented. Two spectral search algorithms and some of their special features are described. The relationship between high-quality Fourier-transform infrared reference spectra and good search results is also discussed, and some other applications of large reference libraries are presented.

INTRODUCTION

Infrared spectroscopy has long been the method of choice for qualitative analysis of organic materials. The unique fingerprinting and identification ability provided by an infrared spectrum results from the fact that the peaks in the spectrum correspond to vibrational modes that are characteristic of the complete molecule and to other modes that are directly related to the fundamental vibrations of specific functional groups. This combination of group frequencies and the "fingerprint" region in infrared spectra has made the comparison of an unknown spectrum to a standard spectrum from a reference material a commonly accepted method for compound confirmation, not only in the laboratory but also in a court of law.

Spectroscopists have tried to improve on visual comparison techniques since the early days of infrared spectroscopy. The first methods for automatically retrieving reference spectra that were similar to an unknown involved encoding punched cards with the locations of the major peaks in a spectrum. One early system actually used a series of notches and holes whereby when a needle was inserted into the hole signifying a specific peak location in the molecule, only those cards with spectra containing the peak were captured. This manual technique was replaced by the automatic card sorting machines from the early days of computers. Both of these sorting methods resulted in a set of cards from those compounds containing the specified spectral features.[1,2]

The first computerized infrared spectral data base of significant size was the ASTM spectral file. This was basically a digitized form of the original punch cards used in the card-sorting methods. This file consists of over 100 000 infrared spectra in a binary format. In a binary format the spectrum is broken into a series of equally spaced intervals.

0095-2338/85/1625-0235$01.50/0

Steve Lowry is presently a senior research scientist at Nicolet Instrument Corp. Steve received his Ph.D. from Tom Isenhour at the University of North Carolina in the area of computerized spectral analysis. He joined Nicolet after working for Diamond Shamrock Corp. as a research chemist and has been actively involved in application and software development in infrared spectroscopy since that time.

Dave Huppler is presently manager of the spectral data base development group at Nicolet. Dave received his Ph.D. from the University of Wisconsin and has been involved in software development at Nicolet for a number of years. Dave has been responsible for most of the search software presently available on the FT-IR spectrometers.

Chuck Anderson is vice-president of research and development in the Nicolet analytical division. Chuck has been active in FT-IR research for a number of years and has been directly involved in the joint data base project with Aldrich Chemical Co.

If a peak maximum occurs in a particular interval, the corresponding location in the spectral representation is set to one. Otherwise, the location is set to zero. A number of research groups have reported infrared spectral search systems based on this data base, and a number are still in use today.[3-5]

Although the algorithms and software developed for the binary spectral libraries have proven useful, this binary spectral representation does not utilize the intensity and peak-shape information contained in an infrared spectrum. Several search systems have been reported that use peak tables that contain not only the peak location but also some level of intensity and peak-width data.[6-8] Many of these searches used highly sophisticated algorithms and feature selection techniques, and yet, the overall results were often poor. Much of this failure was attributed to the small size of the libraries or the somewhat arbitrary methods used to choose peaks and to determine their locations or intensities. Because the spectra were manually encoded, each person used slightly different criteria for peak selection and coding. If a reference spectrum was encoded differently from an unknown spectrum, the match might be bad even if the two spectra were virtually identical.

A major milestone in the development of computerized infrared spectral searching was the creation of the EPA vapor-phase library by Dr. L. V. Azaraga. This was the first large data base of directly true digital infrared spectral data available to the scientific community. A significant feature of this data base was that all the spectra were acquired on a Fourier-transform infrared (FT-IR) spectrometer. The two advantages of FT-IR spectra over previous spectra are the extremely low noise levels and the high wavelength accuracy of the digitized data points. This accuracy is obtained because all frequencies in the spectrum are referenced to the wavelength of a helium–neon laser in most commercial FT-IR spectrometers. The availability of computerized spectra with accurate wavelength registration has provided the basis for full spectral search methods commonly used today.[9-15]

In this paper we will report on work done in this laboratory to investigate different search algorithms and to develop high-quality infrared spectral data bases utilizing modern FT-IR spectrometers. We will describe two search systems that we have developed[16,17] and discuss some of the tradeoffs involved in compressing a spectral library. We will conclude this paper by discussing other applications of data bases.

DATA BASE DEVELOPMENT

Very early in our research into various spectral identification algorithms, we realized that the results were completely dependent on the basic quality of the original spectral data. Even when data-compresion techniques are employed, the results can be significantly better if high signal to noise levels and precise peak-location information are available. These effects were first observed in working with the EPA vapor-phase data base. Although this library was acquired with an FT-IR spectrometer, we encountered a number of errors in the spectral data. Frequently, we retransformed the interferograms and base-line corrected the resulting spectra. This reprocessing greatly improved the effectiveness of this spectral data base as a reference library.

On the basis of these initial results, we concluded that rigorous attention to quality during the initial development of a data base would pay off in future applications. The remainder of this section will include many of the considerations and procedures we implemented in our joint library development with Aldrich Chemical Co. This is an ongoing project, and the first combined book and digital library has been released recently. On the basis of previous experience with smaller spectral acquisition projects, we realized that establishing a comprehensive quality-control program would be essential to the production of a quality library. We have divided this program into four areas: (A) establish exact specifications; (B) set up a program for instrumentation certification; (C) certify sample purity and identification; (D) set procedure for spectroscopic review.

Specifications. (1) Sampling Techniques. Several definitions of spectral quality have appeared in the literature. Of particular interest are the specifications for condensed-phase spectra proposed by the Coblentz society[18] and the original specifications for vapor-phase spectra proposed by the Coblentz society GC-IR subcommittee.[19] We have based our spectral guidelines on the proposals of these groups. The most important reason for defining and adhering to a precise set of specifications is to ensure the consistency of the resulting library. Computer algorithms can generally be designed to handle consistent changes in a data base. It is only when the variables are poorly defined or random that problems arise.

The first major decision in our library project was sampling techniques. This was particularly difficult in the case of solid samples, where both nujol mulls and KBr pellets are used extensively. After many discussions with experts in the field, we chose to run all solid samples as nujol mulls. The major reasons for this choice were based on our concerns about sample reproducibility between laboratories, potential changes in the sample due to the pressure, and ion-exchange possibilities in preparing the sample. Although the presence of the nujol peaks required us to write special compensation software, this was less of a problem than trying to deal with the laboratory to laboratory variation in KBr sample preparation techniques.

All liquid samples were run neat between KBr plates. We also ran all the liquid samples in a fixed-path liquid cell to provide some intensity information. Certain polymer and some low melting point solids were run as melts.

In the case of vapor-phase spectra, we have tried several techniques with mixed success. The first method was a heated 10-cm^{-1} gas cell and an attached heated sample cup. The sample was placed in the cup and allowed to vaporize. This technique worked well for pure volatile samples but failed if the sample contained residual solvents or impurities or was thermally labile. Our second method used GC to volatilize the sample and isolate the pure material. We used our standard GC-IR software with a special lightpipe, which could be isolated with an automatic valve. We monitored the actual IR signal in the lightpipe and trapped the sample when the spectrum in the lightpipe was adequate. Although this technique separated the sample from solvent and impurity and the GC injector minimized sample degradation, we still encountered some sample decomposition in the lightpipe during the trapping. The final technique and the one we are presently

using involves used a packed column injector with a capillary column. This technique uses the GC injector to volatilize the sample and to slowly bleed it into the column. The object is to degrade the chromatographic performance so that the GC peaks are relatively flat and at least 1-min wide. This allows signal averaging with a reasonably constant amount of sample in the cell. We have found this technique to work well with most samples that can be run by gas chromatography.

(2) Resolution. In the choice of resolution for the condensed-phase spectra, we have followed the recommendations of Becker and chosen 2 cm^{-1}. This resolution is obtained by collecting a 16 384-point interferogram and performing a 32 768-point transform. The larger transform size produces better peak shapes. We definitely feel that this is the minimum resolution for representing many spectra where the true peak width may be less than 10 cm^{-1}.

The choice of resolution for the vapor-phase spectra is more difficult because many low molecular weight compounds have rotational splitting of their vibrational modes in the vapor phase. Some of these molecules have rotational peaks with widths of less than 0.1 cm^{-1}. The storage and data requirements for a library at that resolution is beyond the scope of our research. The decision as to what resolution was acceptable was finally influenced by the main application of the data base, which is GC-IR. We chose 1 cm^{-1} as a spectral resolution that should be sufficient for future high-performance GC-IR systems.

(3) Spectral Range. Although we have archived the complete interferogram, we have chosen to save the spectral region between 4800 and 400 cm^{-1}. We keep this extended range in order to retain any potential combination bands in the high end of the spectrum. Even though these peaks are rarely used at this time, they may contain unique information concerning the molecular structure of the sample. In the case of the vapor-phase spectra, we chose a detector that cuts off at 500 cm^{-1}. All spectra in both data bases must have a maximum absorbance between 0.5 and 1.5 absorbance units to ensure an acceptable signal to noise level for the weaker peaks.

Instrument Certification. A key part of ensuring overall spectral quality is the monitoring of spectrometer performance. The following is the procedure followed in our program. (1) A comprehensive set of set spectra is run automatically once a week and retained for reference purposes. (2) Instrument backgrounds are taken twice per day to ensure constant purge calibration and precise spectral ratioing. (3) Water vapor and CO_2 references are produced to certify wavelength accuracy and to remove any residual purge effects from the library spectra. (4) Reference spectra are acquired from the "pure" nujol to check for impurities. The KBr plates are also checked once a week for residual contamination.

By maintaining the data described above, we have a complete record of the spectrometer used in the development of the spectral data base. At any time, the most recent test spectra could be called up and compared to previous spectra with the spectral subtraction and peak pick software on the data system. This ability was often used to quickly resolve any problems or questions concerning instrument performance.

Spectral Purity. One of the major advantages of our joint project with Aldrich Chemical Co. in the development of this data base is the commitment of Aldrich to only retain spectra that truly represent the quality of their chemicals. Each lot of each chemical was analyzed by chromatographic and other quality-control techniques before a spectrum was obtained. This spectrum is then compared to previous reference spectra to ensure that no differences are present.

This concludes our discussion of spectral aquisition. We are commited to the acquisition of high-quality spectral libraries and presently have compiled a number of spectral data bases all containing spectra run on FT-IR spectrometers.

The remainder of this paper will describe several applications of infrared spectral data bases in the chemical laboratory. The major use of digital infrared data bases today is computerized spectral searching. The next section will provide a detailed description of two spectral search systems that we have implemented in this laboratory. We will also describe our research in the area of reverse searching and discuss some of the problems that we encountered in this work.

SEARCH ALGORITHMS

We mentioned earlier that the development of the EPA vapor-phase spectral library by L. V. Azarraga provided the basis for many of the search techniques used today. These full spectral search techniques are derived from some of the pattern-recognition research performed by several groups.[20-22] In the simplest form, the unknown spectrum and each reference spectrum are represented as points in a multidimensional space, where each dimension corresponds to a particular wavenumber location in the spectrum. The similarity of the unknown to each reference spectrum is then computed as the "distance" between the two points in the "hyperspace". This is the basis of the nearest-neighbor technique, which assigns features to an unknown compound based on the features contained in the compounds whose spectra are "closest" to it. The early applications of this method to IR searching were reported by Hanna et al.[9]

We have worked extensively with this concept in our laboratory. Specifically, we have investigated modifications of the basic matching algorithm and have studied the effects of reducing the information content of the spectral representations used for the library. Most of our research has revolved around the spectral library of over 10 000 spectra obtained from compounds listed in the Aldrich catalog. The original spectra have been condensed into a spectral library by deresolving both the intensity and resolution in order to improve search times and to reduce storage requirements. All the spectra were normalized so that the largest peak was 1 absorbance unit. We evaluated the effects of this "deresolving" process by building search libraries with resolutions from 8 to 64 cm^{-1} and intensity information from 20 to 5 bits. We observed that even at the lowest resolutions the search algorithms worked well, but the retrieved spectra were inadequate for visual confirmation of the match. We chose to save the spectral data at 8-cm^{-1} data-point separation and 10 bits of intensity information. This spectral format gives sufficient resolution for plotting and display while reducing the library storage requirements by a factor of 8. We do provide a higher resolution library on the large disk systems. This "Hi Res" library has 2-cm^{-1} data-point separation below 2000 cm^{-1} and 4 cm^{-1} above the scale change. This library requires over 25 megabyte of storage for 10 000 spectra. Figure 1 shows a spectrum of 4-bromofluorobenzene as acquired by Aldrich and the same spectrum in the two compressed formats. This figure demonstrates that even the low-resolution format contains sufficient spectral features for visual confirmation.

Our first work on search algorithms was based on the least-squares calculation shown in eq 1, where x_i is the intensity

$$M_{sq} = \sum_{i=1}^{460} (x_i - y_i)^2 \qquad (1)$$

of the ith point in the unknown and y_i is the corresponding point in one of the reference spectrum. The library spectra with the lowest values of M are saved and listed when the library search is complete. M is frequently called the match factor or "goodness" value. Those spectra with low values of M are more similar to an unknown than spectra with larger values. Although this computation is related to a least-squares

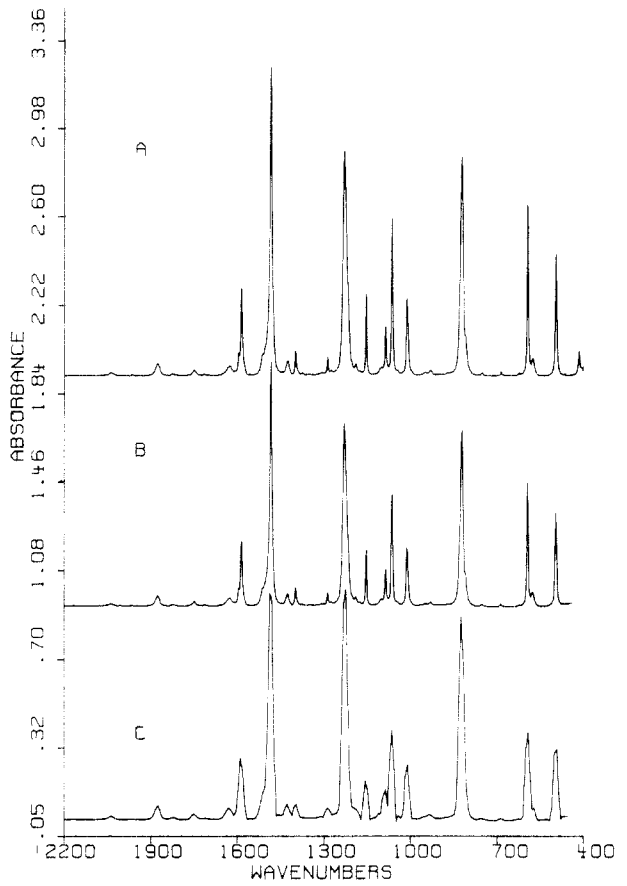

Figure 1. Effects of spectral data compresion: (A) original 2-cm^{-1} resolution; (B) spectrum reduced to $^2/_4$-cm^{-1} data-point separation; (C) spectrum reduced to 8-cm^{-1} data-point separation and 10 bits of intensity scale.

algorithm, no probability should be assigned to a particular value of the match factor.

The first modification that we made to the search algorithm was to compute the absolute value difference rather than the least squares. This calculation is

$$M_{ab} = \sum_{i=1}^{460} |x_i - y_i| \tag{2}$$

where x_i and y_i are the same as in eq 1.

The major difference in this computation is that small differences are weighted more in this computation relative to large differences. Although we chose this algorithm to reduce computation times, we have observed that is seems to give better matches in some cases when the unknown spectrum is not in the library.

A problem that we encountered when we started working with infrared spectra from solids was sloping base lines in the spectra. If either of the methods described above were employed, the base-line offset and slope in the unknown spectrum would contribute to the match factor, resulting in a larger value than might be expected. The following two algorithms were implemented in order to compensate for this problem:

$$M_{sd} = \sum_{i=1}^{460} [(x_i - x_{i+1}) - (y_i - y_{i+1})]^2 \tag{3}$$

$$M_{ad} = \sum_{i=1}^{460} |(x_i - x_{i+1}) - (y_i - y_{i+1})| \tag{4}$$

These metrics involve simply calculating the derivative differences between two spectra. These algorithms not only reduce the contributions of base-line effects but also emphasize the effects of peak shifts or changes in peak shapes. In general,

```
      sx>SRA  = SQ
      sx>SRD
      ALDRICH NEAT LIBRARY
         POSSIBLE HITS
        3287    NEGATIVE CO2
sx>        4752    00000000000000000000000000000000
        3289    NEGATIVE H2O
           5485    00000000000000000000000000000000
        1802    4-FLUOROTHIOPHENOL, 97%
           5524    0128 0000 0000  760 05496 0000 1230 1491
        3285    SEARCH RESULTS ARE QUESTIONABLE
           5820    00000000000000000000000000000000


      sx>SRA  = SQ AB
      sx>SRD
      ALDRICH NEAT LIBRARY
         POSSIBLE HITS
        3287    NEGATIVE CO2
           633     00000000000000000000000000000000
        3289    NEGATIVE H2O
            745     00000000000000000000000000000000
        3285    SEARCH RESULTS ARE QUESTIONABLE
            853     00000000000000000000000000000000
        2030    O-METHOXYPHENYLACETONE, 98%
           1074    0164 0000 0127  010 05250 1054 1247 1496


      sx>SRA  = AB SD
      sx>SRD
      ALDRICH NEAT LIBRARY
         POSSIBLE HITS
        1555    4-BROMOFLUOROBENZENE, 99%
            203     0175 -016 0150  760 05269 1593 1231 1485
        1556    P-FLUOROIODOBENZENE, 99%
           2893    0222 -020 0182  760 05832 1925 1229 1482
        1802    4-FLUOROTHIOPHENOL, 97%
           2930    0128 0000 0000  760 05496 0000 1230 1491
        1554    1-CHLORO-4-FLUOROBENZENE, 98%
           3219    0130 -027 0129  760 04950 1226 0827 1490


      sx>SRA  = SD AD
      sx>SRD
      ALDRICH NEAT LIBRARY
         POSSIBLE HITS
        1555    4-BROMOFLUOROBENZENE, 99%
             86     0175 -016 0150  760 05269 1593 1231 1485
        1556    P-FLUOROIODOBENZENE, 99%
            453     0222 -020 0182  760 05832 1925 1229 1482
        1554    1-CHLORO-4-FLUOROBENZENE, 98%
            458     0130 -027 0129  760 04950 1226 0827 1490
        3285    SEARCH RESULTS ARE QUESTIONABLE
            517     00000000000000000000000000000000
```

Figure 2. Results of searching *p*-bromofluorobenzene spectrum against the Nicolet/Aldrich deresolved library.

if the compound is present in the library, the derivative computations give a larger separation in the match value of the correct spectrum than in the other nearest matches. If the compound is not in the library, the derivative calculation can give worse match values for spectra that have similar spectral features. Figure 2 shows the results of all four search algorithms in searching a spectrum of *p*-bromofluorobenzene. This spectrum was run as a capillary film between old salt plates to create an obvious sloping base line. The results listed in Figure 2 show that the derivitive search algorithms easily compensate for the base-line offset. Figure 3 shows the results of an automatic search and plot subroutine. In this case, the spectrum has been base-line corrected. This ability to overlay the unknown with the best library match is one of the important advantages of digital data. In this example, the derivative algorithms were clearly necessary for correct identification of the original spectrum. Although the problem was reduced by using the base-line correction software available on the FT-IR, this choice of search algorithms adds greater flexibility to the system and provides improved performance.

Up to this point, we have used the complete spectrum for searching. In some cases when the spectrum is known to contain peaks from a contaminant or solvent, better results can be obtained by ignoring portions of the spectrum. We have included a spectral blanking algorithm and an option of scaling the spectrum to a specified value. Figure 4 shows a printout from a search where we have used spectral blanking to focus on the fingerprint region.

One final feature we have added to our full spectral search system is the ability to prefilter the search by using other information about the compound. As part of our Aldrich data

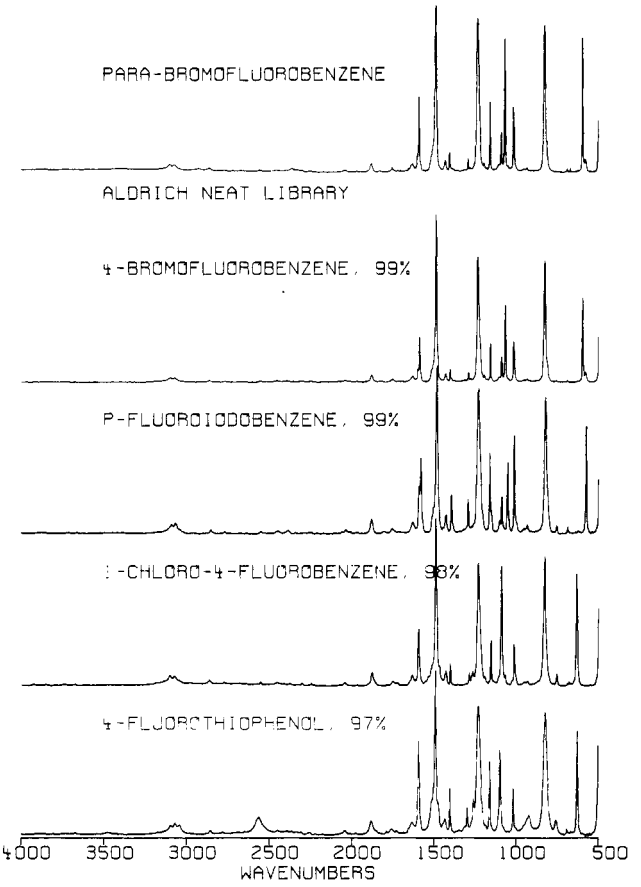

Figure 3. Automatic search and retrieval plot of *p*-bromofluorobenzene and the best matches.

```
!SRQ

SQ
ENTER SKIP REGIONS, NEGATIVE NO. IF DONE
 LOW LIMIT 400

HIGH LIMIT 900

 LOW LIMIT 2000

HIGH LIMIT 4000

 LOW LIMIT -1

MAX FOR SCALING, NEG IF AUTO 1.12

SR

ALDRICH NEAT LIBRARY
   POSSIBLE HITS
  1555    4-BROMOFLUOROBENZENE, 99%
       45    0175 -016 0150  760 05269 1593 1231 1485
  1554    1-CHLORO-4-FLUOROBENZENE, 98%
      128    0130 -027 0129  760 04950 1226 0827 1490
  1556    P-FLUOROIODOBENZENE, 99%
      145    0222 -020 0182  760 05832 1925 1229 1482
   569    PERCHLOROMETHYL MERCAPTAN, 97%
      170    0185 0000 0146  760 05436 1700 0763 0000
   209    PENTACHLOROETHANE, 96%
      171    0202 0000 0161  760 05025 1680 0771 0000
```

Figure 4. Search results using the spectral blanking option to search only on the fingerprint region of the spectrum.

base, we have saved several physical parameters for each entry in the spectral library. These are molecular weight, melting point, boiling point, specific gravity, and refractive index when available. This information is stored in a file linked with the spectral data and can be accessed by the search software. This software checks the value of the parameter for a particular entry, and if it does not meet the defined requirements of the prefilter, that library spectrum is ignored. That is, we can set up the search such that only compounds with a molecular weight between 160 and 170 mass units are searched. Several of these prefilters can be used in series if more information

```
!SRQ
?
!
!SRQ

MW 170,180

MP -22,-10

SO
SR

ALDRICH NEAT LIBRARY
   POSSIBLE HITS
  1058    DIETHYL SUCCINATE, 99%
       0    0174 -020 0217  760 04200 1047 1160 1736
  1099    DIETHYL MALEATE, TECH., 80%
       0    0172 -010 0225  760 04390 1064 1162 1728
  1555    4-BROMOFLUOROBENZENE, 99%
       0    0175 -016 0150  760 05269 1593 1231 1485
```

Figure 5. An example of using the search system to retrieve spectra on the basis of nonspectral information with no search algorithm.

is known about the combound. In fact, one option of the search system allows you to print out all entries in the library that pass the filters without doing any spectral match. This is shown in Figure 5.

The search system described above was designed to perform as an integral part of a spectrometer. The unknown spectrum is automatically converted into a search format, and the spectra from the best matches can be automatically retrieved, displayed, and plotted with the unknown. The results from various search strategies can be quickly compared and all results can be stored for future reference. Although this procedure will generally provide the best results, sometimes the spectrum is not in digital form or we may only have the locations of the major peaks. We have developed an interactive peak search that allows the user to specify peak locations and intensities and to combine these by using Boolean algebra. The key structure of this search algorithm contains two 15 000-bit registers, where each bit corresponds to a spectrum in the library. When a peak location and intensity are entered, the program searches the peak tables and sets the bits for the spectra that contain the appropriate peak. A second bit string can be created on the basis of another peak. These two registers can then be combined on the basis of the chosen Boolean operation to produce a bit string that identifies the spectra remaining. This process can be continued until only a few spectra remain that satisfy all the criteria entered by the spectroscopist. These spectra can be retrieved and the identification confirmed by visual inspection, or the spectra can be plotted and saved for other purposes. We have recently completed the software to include the compound name and molecular formula as part of the user interaction. Figure 6 shows an example the interaction between the user and the software.

An area of ongoing research in this laboratory is the concept of reverse searching. In reverse searching, the software compares the reference to the unknown. In other words, how many features of the reference spectrum are contained in the unknown? These algorithms are designed to handle mixtures. Obviously, a mixture will contain spectral features from all the pure components found in it. This combination spectrum will probably not match well with any members of a library composed of pure spectra. At best, the close matches will be the spectra of the mixture, but more likely, the close matches will be spectra that contain some combination of the functional groups from the pure components of the mixture.

The reverse search that we have developed uses peak tables containing peak locations and intensities of all the peaks in the spectra above 0.1 absorbance unit. The search algorithm computes two metrics. The first calculates the percentage of the peaks in the reference that are found in the unknown. The second metric calculates the percentage of the total intensity in the unknown that is accounted for by the reference. For

```
ex>GRA
ENTER LIBRARY NUMBER
SRM =        7

NICOLET SPECTRAL RETRIEVAL PROGRAM

1.TOTAL RESTART    4. LOGICAL "AND" 7. SAVE SUBSET
2.ENTER PEAK DATA  5. LOGICAL "OR"  8. PRINT ID #
3.ENTER TEXT DATA  6. LOGICAL "NOT" 9. QUIT
 ENTER NUMBER OF CHOICE 2

ENTER PEAK WINDOW AND INTENSITY: W1-W2,I1-I2 1230-1235,70-100

SUBSET CONTAINS  74 SPECTRA
1.TOTAL RESTART    4. LOGICAL "AND" 7. SAVE SUBSET
2.ENTER PEAK DATA  5. LOGICAL "OR"  8. PRINT ID #
3.ENTER TEXT DATA  6. LOGICAL "NOT" 9. QUIT
 ENTER NUMBER OF CHOICE 7

1.TOTAL RESTART    4. LOGICAL "AND" 7. SAVE SUBSET
2.ENTER PEAK DATA  5. LOGICAL "OR"  8. PRINT ID #
3.ENTER TEXT DATA  6. LOGICAL "NOT" 9. QUIT
 ENTER NUMBER OF CHOICE 2

ENTER PEAK WINDOW AND INTENSITY: W1-W2,I1-I2 593-599,50-100

SUBSET CONTAINS  23 SPECTRA
1.TOTAL RESTART    4. LOGICAL "AND" 7. SAVE SUBSET
2.ENTER PEAK DATA  5. LOGICAL "OR"  8. PRINT ID #
3.ENTER TEXT DATA  6. LOGICAL "NOT" 9. QUIT
 ENTER NUMBER OF CHOICE 4

SUBSET CONTAINS   1 SPECTRA
1.TOTAL RESTART    4. LOGICAL "AND" 7. SAVE SUBSET
2.ENTER PEAK DATA  5. LOGICAL "OR"  8. PRINT ID #
3.ENTER TEXT DATA  6. LOGICAL "NOT" 9. QUIT
 ENTER NUMBER OF CHOICE 8

      1    1555
SUBSET CONTAINS   1 SPECTRA
1.TOTAL RESTART    4. LOGICAL "AND" 7. SAVE SUBSET
2.ENTER PEAK DATA  5. LOGICAL "OR"  8. PRINT ID #
3.ENTER TEXT DATA  6. LOGICAL "NOT" 9. QUIT
```

Figure 6. Results from an interactive peak-search algorithm that incorporates accurate peak-location information with intensity data.

a pure unknown, both of these values should be 100%. If the unknown is a mixture, all the peaks in a reference may be present in the unknown, but these peaks may account for a small percentage of the total intensity of the unknown spectrum. Although we have obtained some surprisingly good results on certain samples, particularly in the vapor phase, in general this technique has many problems. A problem with liquid samples is the fact that a mixture of compounds frequently has a spectrum that is not a linear combination of the pure spectra. This is due to the sensitivity of the vibrational modes of a molecule to the local environment. A second problem occurs when a component is a minor consituent of the mixture and its peaks are difficult to identify. If several of the key peaks in the spectrum are missed by the peak-pick algorithm, the search is likely to fail. A third problem involves interference from reference spectra that only contain peaks frequently found in many compounds. This is particularly true for the simple hydrocarbons, which have spectral features that are common to many complex compounds. With certain unknowns, all the low molecular weight alkanes gave high matches even though the unknown was a long-chained ester.

Because of the great demand to identify the components of unknown mixtures, we have not given up on reverse-searching methods, but at this time, we are not confident that we have a solution to the problems encountered.

This concludes our discussion of various search algorithms that we have completed. We are continuing our work on improved and new algorithms for spectral identification and hope to refine our existing search systems.

OTHER APPLICATIONS FOR SPECTRAL DATA BASES

Up to this point we have discussed the uses of infrared libraries for spectral searching. While the automatic identification of unknown compounds can be done by searching a library, other researchers are investigating systems that do not require a reference library for identification. Much of this work falls into either the area of pattern recognition or artificial intelligence. Both of these techniques require a training or test set of spectra to iterate improved performance. Obviously, an interpretation method that cannot perform well on known spectra will be of questionable use for true unknowns. One specific application of spectral libraries in this area is in the evaluation of the Merck spectral interpretation program developed by Woodruff et al.[23,24] This program was designed to take a set of peak data including intensity and width information and to identify possible functional groups in the molecule. As part of this project, they developed an automatic rule generator that used the vapor-phase library to refine a set of interpretation rules. We are presently involved in a project to use this software package to develope an automatic interpretation scheme for environmental monitoring.

Another major use of spectral libraries is simply as reference spectra. The ability to rapidly retrieve, display, and plot spectra can be extremely valuable in day to day laboratory work. Frequently, a spectroscopist is asked if a particular spectrum could possibly be from a certain material. If the reference library is available, the name can simple be entered and the library searched for the compound. This is often all that is needed to solve an industrial problem.

A final area where very little interest has been shown is the application of high-quality data bases to basic research in vibrational spectroscopy. The controlled procedure for acquisition and the high accuracy of the spectra from closely related compounds and homologous series should provide an excellent resource for developing a better understanding of intensity effects and functional group effects in spectra.

Although the major use of spectral libraries today is spectral searching, as the libraries get larger and the computer software becomes more sophisticated the major uses of spectral libraries in the future may be completely different from our applications today.

In this paper we have described our research into both spectral searching and data base development. We feel that the two projects are tightly linked and any success will require equal effort in both areas. This is an area of spectroscopy that is just beginning to mature. The future of research in search algorithms, data compression, and automated interpretation should be both challengeing and productive.

REFERENCES AND NOTES

(1) Kuentzel, L. E. *Anal. Chem.* **1952**, *23*, 1413–1418.
(2) Baker, A. W.; Wright, N.; Opler, A. *Anal. Chem.* **1953**, *25*, 1457–1460.
(3) Sparks, R. A. "Storage and Retrieval of Wyandotte-ASTM Infrared Spectral Data Using an IBM 1401 Computer"; ASTM: Philadelphia, PA, 1964.
(4) Anderson, D. H.; Covert, G. L. *Anal. Chem.* **1967**, *39*, 1288–1293.
(5) Erly, D. S. *Anal. Chem.* **1968**, *40*, 894–898.
(6) Rann, C. S. *Anal. Chem.* **1972**, *40*, 1669–1672.
(7) Schaarachmidt, K.; Reimer, R.; Steger, E. *Z. Chem.* **1974**, *14*, 374–375.
(8) Tanabe, K.; Saeki, S. *Anal. Chem.* **1975**, *44*, 118–122.
(9) Hanna, A.; Marshall, J. C.; Isenhour, T. L. *J. Chromatogr. Sci.* **1979**, *17*, 434–438.
(10) Erickson, M. D. *Appl. Spectrosc.* **1981**, *35*, 181–184.

(11) Small, C. W.; Rasmussen, G. T.; Isenhour, T. L. *Appl. Spectrosc.* **1979**, *33*, 444-448.
(12) Delaney, M. F.; Uden, P. C. *Anal. Chem.* **1979**, *51*, 1242-1243.
(13) de Haseth, J. A.; Azarraga, L. V. *Anal. Chem.* **1981**, *53*, 2292-2295.
(14) Azarraga, L. V.; Hanna, D. A. "ERL GC/FT-IR Software and User's Guide (USEPA/ERL)"; GIFTS: Athens, GA 1979.
(15) Milne, G. W.; Heller, S. R. *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 204-208.
(16) Lowry, S. R.; Huppler, D. A. *Anal. Chem.* **1981**, *53*, 889-893.
(17) Lowry, S. R.; Huppler, D. A. *Anal. Chem.* **1983**, *55*, 1288-1291.
(18) "The Coblentz Society Specifications for Evaluation of Research Quality Analytical Infrared Spectra (Class II)". *Anal. Chem.* **1975**, *47*, 945A.
(19) Griffiths, P. R.; Azarraga, L. V.; de Haseth, J. A.; Hannah, R. W.; Jakobsen, R. J.; Ennis, M. M. *Appl. Spectrosc.* **1979**, *33*, 543-548.
(20) Kowalski, B. R.; Bender, C. F. *Anal. Chem.* **1972**, *44*, 1405-1408.
(21) Cover, T. M.; Hart, P. E. *IEEE Info. Theory* **1967**, *IT-13*, 21.
(22) Leary, J. J.; Justice, J. B.; Tsuge, S.; Lowry, S. R.; Isenhour, T. L. *J. Chromatogr. Sci.* **1973**, *11*, 201-206.
(23) Woodruff, H. B.; Smith, G. M. *Anal. Chem.* **1980**, *52*, 2321-2327.
(24) Tomellini, S. A.; Stevenson, J. M.; Woodruff, H. B. *Anal. Chem.* **1984**, *56*, 67-70.

Performance Analysis of a Simple Infrared Library Search System

MARTIN RUPRECHT and JEAN T. CLERC*
Department of Pharmacy, University of Berne, Berne, Switzerland

Received January 31, 1985

The performance of a commercial microcomputer-based IR library search system has been evaluated. In particular, the effects of sample preparation, concentration and path length, base-line correction, and impurities on the similarity score were studied.

INTRODUCTION

Identification of organic compounds through comparison of infrared spectra is a well-known and often used technique. Many infrared library search systems are offered by instrument vendors or have been described in the literature.[1,2] In order to be useful in practical applications a library search system has to meet several criteria. First of all, if the spectrum of the unknown at hand is part of the library, the system should be able to retrieve the respective reference spectrum. If the unknown is not documented in the reference library, suitable reference compounds structurally similar to the unknown should be retrieved. This, of course, is only possible if suitable reference compounds are part of the library. If the reference library does not contain any reference compound sufficiently similar to the unknown, the system should be able to inform the user about this fact. The similarities between the unknown and the reference spectra retrieved by the system should thus be quantified by an appropriate similarity measure. The system compares spectra, whereas the user thinks in chemical structures. Therefore, the similarity measure has to map similarities in the spectra domain into the structure domain. The similarity measure used by the system should thus as far as possible conform to the user's similarity measure for chemical structures. Furthermore, the system should be insensitive to variations in the spectral data due to different sample preparations. It should also tolerate slightly impure samples. The handling of the system should be easy; the search results should become available within reasonable time and should be presented in a form easily interpreted by the user.

Some aspects of the performance of a library search system are quite easy to specify. The search time for instance can be accurately measured, and whether a reference compound identical with the unknown at hand has been retrieved is easily seen. Other aspects of the performance, however, are extremely difficult to quantify objectively, if this is possible at all. The user's own similarity measure for chemical structures depends heavily on the problem at hand; it is extremely context sensitive. Whether the similarity index given by the system to the top ranking reference compound should be interpreted as indicating identity with the unknown sample or rather as a high degree of similarity is a matter of subjective judgement. Evaluation of the tolerance against slight variations due to sample preparations and/or impurities requires an arbitrary decision as to what should be considered as "slight". User comfort and presentation of the results are again highly subjective. Despite these difficulties, we attempted to evaluate the possibilities and limitations of a simple infrared spectra search system. The results are given and commented in the following.

DESCRIPTION OF THE SEARCH SYSTEM

The system evaluated in this study was the Infrared Library Search Software Package marketed by Pye Unicam Ltd., Cambridge, U.K. It operates on the Pye Unicam SP3-080 data console. For this study the data console was connected to a Pye Unicam SP3-300 grating infrared spectrometer. In order to get full control over the reference library and to have unlimited access to the source spectra, a specially prepared reference library was used in this study rather than the library tapes supplied with the system. This reference library consists of 270 spectra of relatively simple organic compounds covering a wide range of compound classes. All spectra were recorded in KBr wafers at concentrations giving about 10% transmittance for the strongest band in the 4000–600 cm^{-1} wavenumber range, with instrument parameter settings commonly used in routine work. The samples were all of analytical grade and were used as received from the supplier. As the data console uses tape cartridges for mass storage, space for the reference library is somewhat limited (a library tape will hold about 1000 reference spectra), and the search time is determined by the access time to the tape. The spectral data may be entered automatically from digitized spectra obtained on the data console or manually from a chart recording. The data is then normalized to a standard form and encoded. The vendor does not give information about the algorithms used for encoding, searching, and similarity calculation. It is claimed, however, that the algorithms used provide compensation for wavenumber or transmittance inaccuracies associated with data read from photoreduced chart records, chart readings from instruments from other manufacturers, or poorly maintained instruments.

USER COMFORT AND PRESENTATION OF RESULTS

Once the spectrum of an unknown compound is recorded and stored in the data console memory, there are few operations to be performed for searching the library. The system assumes that all "cosmetic" operations on the data set (e.g., smoothing, compensation for sloping base line) have been performed before entering the search mode. First of all, the system selects spectrally significant data from the fully di-

0095-2338/85/1625-0241$01.50/0