Pure & Appl. Chem., Vol. 49, pp. 1827 - 1831. Pergamon Press, 1977. Printed in Great Britain.

COMPUTER-BASED SYSTEMS FOR THE RETRIEVAL OF INFRARED SPECTRAL DATA

Jean-Thomas Clerc and Jure Zupan*

Swiss Federal Institute of Technology, Department of Organic Chemistry,
8092 Zürich, Switzerland

Abstract - Different approaches to the computerised retrieval of infrared spectroscopic data are reviewed.

Comparing the infrared spectrum of a sample of unknown structure with reference spectra in order to identify the sample is an old and well-known technique. One attempts to retrieve from the reference library a compound exhibiting spectral data virtually identical to the spectral data of the sample. If such a reference compound is found, identity between the sample and the retrieved reference compound is assumed. If no such reference data set is found in the library, the reference spectra most similar to the spectrum of the unknown sample are retrieved. To the same extent to which similarities in structures are represented by similarities of spectra, the structures of the retrieved reference compounds will furnish helpful structural suggestions to the analyst.

Direct manual search of a reference library and visual comparison of the spectra are feasible only with rather small reference libraries. Even with moderately sized reference spectra collections, encoding of the spectral data and/or automated search and comparing methods are a necessity. It is the scope of this paper to critically review some encoding schemes, automated search procedures, and computer evaluated similarity measures used in the retrieval of infrared spectral data.

Automated retrieval methods require the infrared spectral data to be available in machine readable form. Earliest attempts to the automated retrieval of infrared spectra date back into the pre-computer era, in which Hollerith type punched cards were the storage medium of choice. The inherent limitations of the electromechanical card sorting equipment available at that time largely determined the codes used to register the spectroscopic data.

Probably the most comprehensive file of computer readable infrared spectra was compiled by the American Society for Testing and Materials (ASTM) (Ref. 1). The file now contains well over 100 000 entries and is still widely used. The ASTM file gives information as to whether an absorption peak maximum does occur in any given 0.1 μm wave length interval. Furthermore, the presence of a strong peak in any given 1 μm wave length interval is also indicated. In addition, the file supplies considerable chemical data for the reference compounds (e.g. molecular formula, name, chemical classification data, etc.). The encoding scheme used severely truncates the spectral data. All information about peak shape is lost, and peak intensity coding is arbitrary and ambiguous, as it is not applied to a specific peak but rather to all peaks within a 1 μm wave length range. Furthermore, differences in sample preparation, spectrometer type, and in the spectroscopist's interpretation of the spectra may lead to quite variable codings for the same compound.

Early attempts to the computerised searching of the ASTM file replaced the card input by magnetic tape (Ref. 2) and disk (Ref. 3) input. For magnetic tape input, Anderson and Covert (2) replaced the non-standard codes used on the ASTM cards by standard Hollerith codes, which results in a considerable increase in file length. To initiate a search the user specifies in which spectral region an absorption peak has to occur and which wave length intervals have to be peak free. These mandatory terms may be supplemented with desirable terms. All reference spectra failing to meet a mandatory term are immediately rejected. After having searched through the complete file the reference spectra fulfilling all mandatory terms are sorted according to the number of desirable terms met and then put out in that order. Non-spectroscopic information may be included in the search.

---

* Chemical Institute Boris Kidric, 61001 Ljubljana, Yugoslavia

A conceptually very similar but more sophisticated system has been described by Sebesta and Johnson (4). Their system is claimed to be able to identify components from multicomponent mixtures.

Erley (3) has adapted the ASTM file for disk input. Here, the absorption data and the non-spectroscopic supplemental information are stored in binary coded form in 160 bit, so that a rather compact representation is achieved. The user enters the absorption band positions for the unknown sample as well as any "no band" regions. Non-spectroscopic information may also be specified. The computer converts the input data into a series of masks, which are then compared with the binary representations of the standard spectra by logical AND and XOR operations. As these are basic computer operations they can be performed extremely fast. Finally, all reference spectra meeting the specified requests are put out.

Ideally, a search should yield just one hit, namely the correct material. However, if the search question is stated too narrow, the correct compound will often not be retrieved because the codes for the two spectra will not be identical. On the other hand, if the search question is stated too broad, an excessive number of reference compounds will be retrieved. A good balance between selectivity and tolerance has therefore to be achieved, which is by no means easy. The most frequent causes for failure to retrieve the correct reference spectra are (Ref. 5) miscoded bands in the file, variations in the choice of how a weak band should be coded, and overuse of "no band" regions by the searcher. To cope with these sources of errors, a tolerance has to be applied to the peak absorption data. This may be effectuated at the time the reference file is generated (Ref. 2) or by "wiggling" the input data for the unknown (Ref. 3). This generally helps to achieve an acceptable recall but tends to result in extremely low precision, unless non-spectroscopic information is extensively used to discriminate against unwanted reference compounds (Ref. 6). There is however an inherent danger of self-deception in this approach. If the user is allowed to restrict the search to compounds belonging to specified chemical classes, he will obviously select only those classes he considers probable for the sample at hand. Thus, the search system is forced to put out only answers conforming the user's expectations. Despite this inherent danger, this method for increasing the search precision is widely used.

Balancing selectivity against tolerance becomes less difficult when the number of search term types and the number of operators to combine them is high. An infrared spectra search system being rather unique in this respect is described by Woodruff *et al.* (7). It uses a minicomputer programme originally developed for text search, featuring truncation and full Boolean logic. It is claimed to be superior to conventional search systems when additional non-spectroscopic information about the unknown sample is available and is included in the search. In addition, investigations more complex than simple searching and comparing are possible, allowing for *e.g.* automatic detection of data set inconsistencies, or preparation of selected subsets for subsequent investigations by pattern recognition or statistical techniques. The price to be paid for these additional capabilities is decreased search speed.

Another group of infrared spectra search systems uses reference file codings modelled after the "Sadtler's Spec-Finder" (Ref. 8). In contrast to the ASTM coding, which lists all absorption bands, only the position of the strongest absorption band within a 1 μm wave-length range is encoded to the nearest 0.1 μm. Thus, for the wave-length range of standard infrared spectral data from 2 to 15 um, at most 13 peaks are encoded. The spectral data for any given compound may therefore be expressed in 13 numbers, corresponding to the wave-length intervals. Each number can have 11 different values, ten corresponding to the strongest absorption in 0.1 μm intervals, and one for "no absorption". There are theoretically $11^{13} < 2^{45} \sim 3 \cdot 10^{13}$ distinct codes possible. Therefore, a bit string of length 45 is adequate to uniquely represent the code for the infrared spectral data for any given compound. Such a compact representation of the spectral code allows for an extremely high search speed as well as for low storage requirements (Ref. 9).

As the number of theoretically possible distinct codes ($\sim 3 \cdot 10^{13}$) by far exceeds the number of reference spectra, the 45 bit representation is still far from the optimum. It is thus feasible to further reduce the length of the code, either by statistical compression (Ref. 9) or by hash coding (Ref. 10). In the former method, two or more bits corresponding to spectral regions with low discrimination power are merged into one common bit. This approach has been used to compress the spectral code to a length of 16 bit, resulting in further significant savings in computer time and storage space without intolerable loss in precision relative to the 45 bit code (Ref. 9). Hash coding is conceptually very similar to statistical compression but has been tested only on simulated infrared spectra (Ref. 10).

Another approach for increasing the search speed uses inverted files (Ref. 11). Normal reference data collections are organised in a file in which to every reference compound corresponds one entry listing the position of the encoded absorption peaks. The inverted file how-

ever has an entry for every wave-length range, containing a list of all reference compounds which exhibit a coded peak in the respective wave-length interval. Practically, each wave-length interval has a byte string assigned, in which every byte corresponds to one reference compound. If the reference compound shows a coded absorption peak in the given interval, the value of the respective byte is set to one. Otherwise, its value is zero. To search for a reference spectrum exhibiting peaks in given wave-length intervals, the respective entries in the inverted file are selected and simply added together. The result is an analogous byte string, in which the value of every byte gives the number of matches for the respective reference compound. Hence, the bytes having the highest values indicate the best matching reference compounds. With this approach a very high search speed may be realised. However, working storage demands of the programme tend to be rather high and updating the inverted file is not trivial.

The wave-length or wave-number intervals used for encoding the spectral data of course do not have to follow exactly the conventions used in the "Spec-Finder". Any similar technique may be used, as long as the rules defining the code insure that any two operators will obtain the same (or very similar) code numbers when encoding a particular spectrum. If the rules are selected so as to transform the infrared spectral data into a decimal number, the use of a programmable desk top calculator for searching the spectra library becomes feasible. In an approach described by Rann (12), the wave-number scale is arbitrarily divided into 10 sections, the divisions being arranged such that more sampling is provided in the "finger print" regions of the spectrum. Each of the 10 sections is further subdivided into 10. The code digit for each section is obtained by taking the maximum absorbance of the spectrometer trace within that section and assigning to it the number of the appropriate subdivision. Thus, the spectral trace is converted to a ten-digit decimal number. The library of reference spectra, encoded as described, is stored on paper tape. To search through the library, the ten digits characterising the spectrum of the unknown sample are compared with the library spectra codes, which are read sequentially from the library paper tape. For each comparison a figure of merit is calculated which evaluates quantitatively the degree to which the two spectra match each other. This figure is obtained by summing the modulus of the differences between corresponding digits in the two codes compared. A perfect hit results in this sum being zero. However, as several operators may have been involved in coding the spectra, some differences in the codes for identical spectra have to be expected, which will be reflected by a score slightly greater than zero. Thus, a low value of the sum will merit a manual inspection of the spectra involved. The mathematical operations for this search procedure are so simple that they may easily be implemented on a programmable desk top calculator. However, speed and/or storage space limitations of the available peripheral devices for storage and input of the library spectra codes impose severe restrictions on the maximum allowable size of the reference library.

All previously discussed computerised infrared spectra retrieval systems use only information related to the position of main absorption peaks on the wave-length or wave-number scale. They thus neglect any information derived from peak intensity, peak width, and peak shape. The restrictions imposed by this fact were not felt strongly in the age of the prism spectrometer, as the limited spectral resolution and reproducibility of routine prism spectrometers did not allow to determine these parameters with sufficient accuracy. However, with today's modern grating spectrometers and improved photometric techniques this is no more a significant problem. Furthermore, modern infrared spectrometers will be equipped with the necessary hardware for direct connection to a computer or will even contain a microprocessor for preprocessing the spectral data (e.g. Ref. 13), thus making the direct digital acquisition of complete infrared spectra feasible. However, it will take considerable time to accumulate large collections containing the full spectral data, so that the use of the truncated but very comprehensive older spectra files is still justified.

The probably most up-to-date search system for truncated spectra was developed by Zupan et al. (14, 15 & 16). It uses the ASTM data base (Ref. 1) in its most recent form. The spectral data is encoded in 180 bit, the supplemental information in 480 bit. The system allows for the identification of single unknowns as well as for binary mixtures. The input to the system consists of absorption band positions (specified on the wave-length or wave-number scale) with individually selectable tolerance levels of up to 0.9 µm, of "no band" regions, and of various combinations of supplemental information requirements. In the single component search mode the strategy used is similar to the one developed by Erley (3). However, a more sophisticated similarity measure is used. Moreover, the programmes are largely machine independent, as they are written in the FORTRAN IV language. Even when programmed in this high-level language, the search proceeds with adequate speed, more than 1000 spectra per second being processed on a CDC Cyber 60 computer.

For the identification of binary mixtures, the spectrum of the mixture is assumed to be the sum of the spectra of its components. In a first run, all possible mixture components are se-

lected from the reference compilation and their spectra are stored in a highly compressed form. In a second run, all binary combinations are tested against the spectrum of the mixture, the best matching combinations being selected for output. For the second run, the highly compact and suitably preprocessed data generated in the first run is used rather than the complete data sets from the reference library. This results in greatly enhanced search speed, the second run taking on the average only about 15% of the total search time.

If in addition to peak position data, peak intensity and/or peak shape data are included in the reference library, it becomes feasible to develop more sophisticated computing procedures to evaluate the degree of similarity between the two spectra compared. Even with very large data collections it can generally not be assumed that the spectral data for a compound identical to a given unknown sample is present in the library. Furthermore, even when the spectral data for such a compound is available, there will be some differences between the two spectra due to unavoidable variations in sample preparation, purity, type of instrument used, etc. One has therefore to assume that no exact replica of the spectrum of the unknown sample will be found in the library. In consequence, heaviest emphasis has to be placed on a similarity measure which is insensitive to instrumental and technical artifacts and which at the same time predominantly reflects structural rather than spectral similarity (Ref. 17). A spectra search system designed to retrieve reference spectra exhibiting the highest values of such a similarity measure can give useful results even when no reference compound with the same structure as the sample is documented in the library. Thus, the limitations imposed by the contents of the reference library become significantly less stringent.

Recently, Zupan and Hadži (18) have reported attempts to develop an algorithm for the quantitative evaluation of the similarity between two structures. Such an algorithm would allow for the unambiguous and unbiased comparison of different search strategies and different similarity measures for the comparison of spectral data. However, as the proposed algorithm is based on the comparison of WLN codes (Ref. 19), it results in a structural similarity measure strongly biased in favour of such structural entities that have a unique representation in the WLN code.

Even a very crude peak intensity code distinguishing only between strong, medium, and weak absorption peaks allows for the evaluation of quite sensitive similarity measures. This may be realised with a simple scoring scheme, where preset positive (or negative) scores are assigned to the different types of matches (or mismatches) (Ref. 20). Furthermore, by varying the assigned values, the search strategy may be optimised to varying types of search problems. If the spectrum of a given unknown sample features peaks that rarely occur within the wide range of reference compounds, it may be advantageous to put heavy weight on the presence of these peaks in the reference compounds. Another search strategy gives an additional bonus to those reference spectra which, besides all desired peaks, have the lowest total number of peaks in the complete spectrum. This strategy will predominantly retrieve simple spectra. It is considered useful in attempts to correlate specific peaks with specific functional groups.

In another approach (Ref. 21) relative peak intensities are recorded with a precision of 1%. On the one hand, this allows for a very precise description of relative peak intensities. On the other hand, this coding requires utmost care in the selection of tolerance ranges to cope with the unavoidable variations in the spectral data of identical compounds.

In addition to peak positions and intensities, peak shapes may also be included in the evaluation of a suitable similarity measure. Penski *et al.* (6) encode for the largest peaks in an infrared spectrum position, intensity, and shape. Intensities are graded as strong, medium, or weak. Peak shapes are coded as sharp, medium, or broad, depending on the half width of the respective band. For the definition of the similarity measure it is firstly assumed that a separation in the wave length of two peaks reduces their probability of being a match by a functional relationship based on the normal distribution (*cf.* equation 1). Secondly, it is assumed that the relative value of a match between two peaks of like intensity and shape decreases with decreasing intensity and increasing half width. Thirdly, if intensity and/or peak shape of two compared peaks do not match, the relative weight for the match of like peaks is suitably reduced. To evaluate the similarity measure every peak in the spectrum of the unknown sample is compared to every peak in the reference spectrum. Each peak pair compared contributes to a match sum, the contribution being the product of three factors. The first factor measures the degree of match in peak position, the second factor W is the relative value for a match of like peaks, and the third factor R reduces the total contribution if intensity and/or shape of the peaks do not match. With x_i being the wave length of the i-th peak in the spectrum of the unknown sample and y_j being the wave length of the j-th peak in the reference spectrum, the match sum M_{ij} is given by equation 1.

$$M_{ij} = \sum_i \sum_j e^{-(y_j - x_i)^2 / 2\sigma_i} \cdot W_i \cdot R_{ij} \qquad (1)$$

The standard deviation σ_i for the wave-length position of the i-th peak is assumed to be $x_i/60$. Obviously, the match sum M_{ij} will strongly depend on the number of peaks in the spectra compared and is therefore not suitable for direct use. However, a good similarity measure is obtained by dividing the match sum M_{ij} between the unknown and known by the match sum M_{ii} between the unknown and itself.

The retrieval systems discussed above still do not use the full information content available in an infrared spectrum. This would only be possible if the complete curve trace is digitised with a resolution equal to or better than the resolution of the spectrometer used. However, this results in a number of data points too large to be efficiently stored and processed. Thus, a compromise becomes necessary. Tanabe and Saeki (22) have performed pilot studies on the comparison of complete infrared spectra, using the correlation coefficient between two digitised spectral traces as the similarity measure. They conclude that for identification purposes a wave-number resolution of ≤ 10 cm^{-1} in the 1200 - 650 cm^{-1} range is adequate. Furthermore, they recommend to record the spectral intensity on the absorbance scale, either by directly measuring in the absorbance mode or by conversion of the transmittance data. This eliminates the problems connected with variable sample thickness. Shifts on the wave-number scale of up to 3 cm^{-1} have been found to be of little influence on the correlation coefficient. Shifts in the intensity base line are of course without effect.

In order that infrared spectral data retrieval systems using the full curve trace become feasible for practical use, two conditions have to be met. First of all, sufficiently comprehensive collections of fully digitised spectral traces have to be available. Furthermore, it has to be easy for the analyst to get the full infrared spectral data for an unknown sample in machine readable form. To adequately meet the second condition, infrared spectrometers with direct digital output are needed. Recent trends in instrument design clearly point into this direction (e.g. Ref. 13). As to the first requirement, the retrospective digitisation of existing conventional data collections may be more efficient than building up completely new collections (Ref. 23). In any case, recent trends in the development of computer-based systems for the retrieval of infrared spectral data indicate beyond any doubt that future systems will use non-truncated data digitally stored in a form allowing for the reconstruction of the original spectral trace.

Acknowledgement - This work has been supported by the Schweizerischer Nationalfonds zur Förderung der wissenschaftlichen Forschung and by the Swiss Department of Commerce (Commission of the European Communities Project COST 64b).

REFERENCES

1. "Codes and Instructions for Wyandotte-ASTM", ASTM, 1916 Race St., Philadelphia, Pa. (1964).
2. D.H. Anderson and G.L. Covert, Analyt. Chem. 39, 1288-1293 (1967).
3. D.S. Erley, Analyt. Chem. 40, 894-898 (1968).
4. R.W. Sebesta and G.G. Johnson, Jr., Analyt. Chem. 44, 260-265 (1972).
5. D.S. Erley, Appl. Spectroscopy 25, 200-202 (1971).
6. E.C. Penski, D.A. Padowski and J.B. Bouck, Analyt. Chem. 46, 955-957, (1974).
7. H.B. Woodruff, S.R. Lowry and T.L. Isenhour, J. Chem. Inf. Comput. Sci. 15, 207-212 (1975).
8. Sadtler Research Laboratories, 1517 Vine St., Philadelphia, Pa.
9. F.E. Lytle and T.L. Brazie, Analyt. Chem. 42, 1532-1535 (1970).
10. P.C. Jurs, Analyt. Chem. 43, 364-367 (1971).
11. F.E. Lytle, Analyt. Chem. 42, 355-357 (1970).
12. C.S. Rann, Analyt. Chem. 44, 1669-1672 (1972).
13. Infrared spectrophotometer model PE 283, Perkin-Elmer, Norwalk, Conn.
14. J. Zupan, D. Hadži and M. Penca, Kem. Ind. 5, 275-277 (1974).
15. J. Zupan, D. Hadži and M. Penca, Comput. Chem. 1 (1976), in press.
16. J. Zupan, D. Hadži and M. Penca, Europ. Spectr. News (1976), in press.
17. J.T. Clerc and F. Erni, Topics in Current Chem. 39, 91-107 (1973).
18. J. Zupan and D. Hadži, III International Conference on Computers in Chemical Research, Education and Technology, Caracas, Venezuela (1976).
19. E.G. Smith, The Wiswesser Line-Formula Chemical Notation, McGraw-Hill, New York (1968).
20. R.C. Fox, Analyt. Chem. 48, 717-721 (1976).
21. V.A. Koptjug, Z. Chem. 15, 41-47 (1975).
22. K. Tanabe and S. Saeki, Analyt. Chem. 47, 118-122 (1975).
23. J.T. Clerc, R. Knutti, H. Könitzer and J. Zupan, Z. analyt. Chemie (1976), in press.