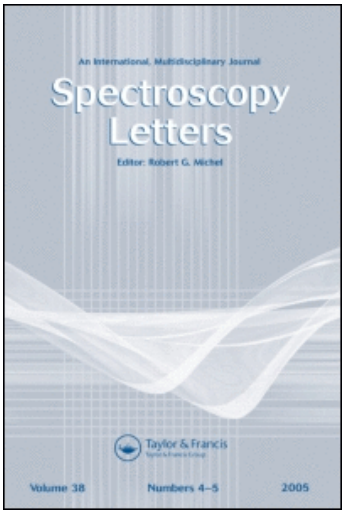

Spectroscopy Letters
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713597299

Description and Performance Analysis of an Infrared Library Search System

P. N. Penchev[a]; A. N. Sohou[a]; G. N. Andreev[a]
[a] Faculty of Chemistry, University of Plovdiv, Plovdiv, Bulgaria

To cite this Article Penchev, P. N. , Sohou, A. N. and Andreev, G. N.(1996) 'Description and Performance Analysis of an Infrared Library Search System', Spectroscopy Letters, 29: 8, 1513 — 1522
To link to this Article: DOI: 10.1080/00387019608007141
URL: http://dx.doi.org/10.1080/00387019608007141

SPECTROSCOPY LETTERS, 29(8), 1513-1522 (1996)

DESCRIPTION AND PERFORMANCE ANALYSIS OF AN INFRARED LIBRARY SEARCH SYSTEM

Key Words: infrared spectra, library search system, computer aided identification of organic compounds.

P. N. Penchev, A. N. Sohou, G. N. Andreev*

Faculty of Chemistry, University of Plovdiv,
24 Tsar Assen Str., 4000 Plovdiv, Bulgaria

ABSTRACT

An infrared library search system is described. The spectral library consists of 608 FT-IR spectra represented with a data point every 4 cm^{-1} in the 3700-500 cm^{-1} range. Four different similarity measures for spectral search were implemented. Performance analysis was carried out in order to estimate the ability of the system to identify organic compounds on the basis of their IR spectra.

* Author to whom correspondence should be addressed

INTRODUCTION

The development of automated systems for structure elucidation and identification of organic compounds with the aid of spectroscopic methods continues to attract the attention of the spectroscopists. In general, these systems can be classified in three categories: knowledge-based systems in which chemical expertise is encoded to assist spectra interpretation, pattern recognition methods which have the ability to recognize and classify spectra patterns, and the most widely used technique - searching in spectral libraries. Each of these approaches has its own advantages and limitations [1-3] but the library search software has demonstrated its usefulness in scientific and laboratory practice [3-12].

In this article we describe an infrared library search system and the results from the test searches.

EXPERIMENTAL

Measurements and Processing of Spectra

608 spectra were registered on a Perkin-Elmer 1750 FT-IR Spectrometer from 4000 cm^{-1} to 450 cm^{-1} at resolution 4 cm^{-1} with 25 scans. The solid samples were measured in KBr pellets and the liquids - as thin films so that the strongest band in the 4000-450 cm^{-1} wavenumber range gave approximately 10% transmittance (T). All spectra were subjected to curvilinear baseline correction and saved as files with Perkin-Elmer spectral format [13] on the PE 7700 Professional Computer. Further, they were off-line converted into JCAMP-DX files and the latter were transferred to an IBM compatible computer with the standard protocol for data exchange KERMIT [14]. The spectra from the received ASCII files were smoothed according to Savitzky-Golay algorithm [15] and

gathered into the library file. In this file each entry is represented with the following data: chemical name, Wiswesser line-formula chemical notation, molecular formula, molecular mass, boiling and melting points, field for comments, and spectral data. The latter are absorption values at 4 cm^{-1} data intervals in the 3700 - 500 cm^{-1} range amounting to 801 points altogether. They are normalized in the range of 0-1 absorbance units (A).

The program code was written and compiled in Turbo Pascal 6.0. The search experiments were carried out on a PC/AT 486 DX/2 (66 MHz) computer.

Spectral Similarity Measures

Four different measures (hit quality indices, HQIs) [2] for spectral similarity were used. They compare the entire spectral curves and calculate 'the distance' in spectral space between the spectrum of the unknown compound and that of the reference one. These measures are based on the following relations: sum of least squares (Eqn. 1), sum of absolute value differences (Eqn. 2), scalar product (Eqn. 3) and correlation coefficient (Eqn. 4).

$$S_1 = \sqrt{\sum_k (A_k^U - A_k^R)^2 / N} \qquad (1)$$

$$S_2 = (1/N)\sum_k \left|A_k^U - A_k^R\right| \qquad (2)$$

$$S_3 = \frac{\sum_k A_k^U A_k^R}{\left|A^U\right| \cdot \left|A^R\right|} \qquad (3)$$

$$S_4 = \frac{\sum_k (A_k^U - \overline{A^U})(A_k^R - \overline{A^R})}{\sqrt{\sum_k (A_k^U - \overline{A^U})^2 * \sum_k (A_k^R - \overline{A^R})^2}} \qquad (4)$$

Here N is the number of data points, and A_k^U and A_k^R mean k^{th} absorption value in the unknown spectrum and in the library one, respectively.

The relations S_3 and S_4 show the similarity between the spectra, and their maximum value is 1.0 for identical spectra. The lowest theoretical limit of S_3 is 0.0 for orthogonal spectra, whereas that of S_4 is -1.0 for spectra with negative correlation (r = -1). That is why the HQIs are the relations (3) and (4) normalized in the 0-999 range.

The relations S_1 and S_2 show the dissimilarity between the spectra, and their minimum value is 0.0 for identical spectra. The highest theoretical limit of S_1 and S_2 is 1.0 if the spectra are represented in the 0-1 A interval. Thus the HQIs were calculated according to Eqn. (5):

$$HQI_k = C*999/(S_k + C);\ k = 1,2 \qquad (5)$$

C is a constant that prevents the HQIs to become infinite when the spectra are identical. Our experiments show that the C value in HQI_1 must differ from that in HQI_2 in order to obtain reasonable HQIs.

RESULTS AND DISCUSSION

Two spectral sets were composed: (i) 27 Sadtler spectra [16] of compounds, present also in our library, were selected: they represent spectra registered in other laboratories. All Sadtler spectra were normalized in the 0-1 A range by the producer. (ii) 48 spectra from our library were selected and the corresponding compounds were measured once again without obeying the requirement that the maximum band absorption should be about 10% T.

The spectra from both sets were used as unknowns in the library searches and statistics upon the obtained hit lists (20 spectra best matching the unknown) were performed.

The detailed analysis of the results for the first set indicates that the process of compound's identification is always successful if the two spectral curves do not differ significantly. For the sum-of-least-squares' HQI (LS HQI), 21 out of 27 hit lists contain the relevant library spectrum (called hit) on the top position; for the sum-of-absolute-values-differences'

HQI (AV HQI) they are 16; for the scalar-product's HQI (SP HQI) - 22, and for the correlation-coefficient's HQI (CC HQI) - 23.

There are several cases in which the library spectrum of the 'unknown' compound is not at the top position in the hit list. The reasons are:

(1) for Sadtler spectra of dichloromethane and diiodomethane the obtained hit list, when HQI is calculated according to Eqn. 2, contains the relevant reference spectrum in 10th, respectively in 2nd position. This can be explained by the fact that the library spectra of these compounds have little but noticeable baseline deviation.

It is well known that small differences in ordinate values are weighted more heavily in Eqn. 2 compared with the other Eqns. In addition the relative bands' number in both spectra is less than the usual number, thus contributing insignificantly to the sum in Eqn. 2. Both reasons lead to top positioned spectra in the hit lists with nearly the same level of baseline as that of the unknown and, subsequently, to misidentifications.

(2) The search system can not identify the Sadtler spectrum of p-methoxyphenol. The comparison of the spectral curves of both Sadtler and library spectra shows that the former is registered at considerably longer path length. The best result (the hit is in 3rd position) is obtained when the CC HQI is applied; all other hit lists do not contain the relevant library spectrum among the top fifteen positions.

(3) The identity search for the Sadtler spectrum of oleic acid gives hit lists which contain the relevant library spectrum in second position, the top spectrum being octanoic acid. Only when AV HQI is used, does the search give the oleic acid spectrum on the top of the hit list. These results can be explained by the very small differences between the spectra of both acids: the oleic acid spectrum does not contain a band for C=C bond (shoulder on the $\nu_{C=O}$ band), and the =C-H out-of-plane band appears

with too low intensity. As mentioned before, Eqn. 2 weights better the small differences between the compared spectra, and thus succeeds in the identification.

(4) In five cases the relevant library spectra are of poor quality: they have either a noticeable baseline deviation, or were registered with longer (shorter) path lengths or with larger (smaller) amount of compound in the KBr pellet. The HQI most insensitive to these variables is a CC HQI (all hits are on the top), followed by a SP HQI (4 top hits and one in second position), LS HQI (2 top hits, and 3 in second positions), and AV HQI (no top hits).

(5) In two cases the Sadtler spectra do not considerably differ from the relevant library spectra, but the system could not identify successfully the compounds. It is interesting to note that HQIs calculated according to Eqns. 1 and 2 give higher position of the hit than do the rest. A check up of the hit lists shows that compounds with similar structure occupy the top positions.

The results for the second set also show correct identification of the compounds when the unknown spectrum is run under the same experimental conditions as the library spectra. For the LS HQI, 42 out of 48 hit lists contain the hit in the top position; for the AV HQI, they are 35; for the SP HQI - 44, and for the CC HQI - 46.

The cases in which the library spectrum of the 'unknown' compound is not at the top position in the hit list can be explained as follows:

(1) The respective 'unknown' spectrum is with uncorrected baseline, or has pruned and/or overlapped bands as a result of a longer path length or a larger amount of compound in the KBr pellet. Three spectra have bands with low intensity (before normalization), and this fact also upsets the identification. It is interesting to note that the best 'identifying' HQI is the one calculated according to Eqn. 4. The finding can

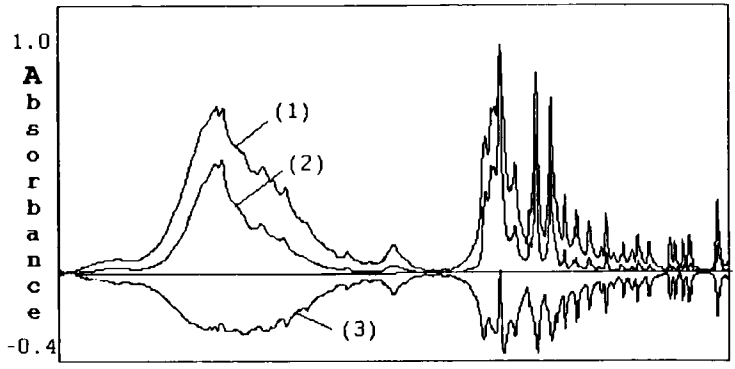

Figure 1. The library (1) and the 'unknown' test (2) spectra of DL-Methionine, and their difference (3)

be explained with the aid of Fig. 1, where the library and 'unknown' test spectra of DL-Methionine, and their difference are given.

As it is seen from Fig. 1, there exists a considerable difference between the two spectra, but the correlation coefficient is 0.961 (HQI = 979). The regression line between absorbance values of the two spectra is $A^U = 0.62*A^R$. The value of intercept, -0.031, is statistically insignificant pointing the same level of the baselines for the two spectra, but the slope, 0.62, indicates the difference between the two spectra. In this case the SP HQI also provides good results because of its meaning as cosine of the angle between both spectra (vectors) in the spectral space. It is well known, that this angle does not depend on the length of both vectors, i.e. on the registered compound's amount.

(2) Some relatively long homologous series are present in our library. In the test runs it leads to three cases of misidentification. The AV HQI gives fourth position of the hit for the test spectrum of 2-octanone, the first three spectra being those of 2-decanone, 2-undecanone, and 3-nonanone. The visual comparison shows that the first two spectra are

almost identical with the unknown, and the spectrum of the third compound is nearly the same as the unknown. This points out that not only the 'similarity' search [1] but also the 'identity' search is deteriorated when the library is not balanced on homologs.

For 2-nonanone only CC HQI (seventh hit position), and to a certain extent the SP HQI (tenth hit position) give interpretable results: the other HQIs fail. It is interesting that the first four spectra in the hit lists obtained, those of 2-undecanone, 2-decanone, 3-dodecanone, and 2-octanone are almost identical with the unknown.

The other long homologous series is that of 1-phenyl-2-alkanones causing problems when only AV HQI and LS HQI are applied.

CONCLUSIONS

An infrared library search system has been created. The results from the identity searches performed, using four different similarity measures with test spectra registered on one hand in our laboratory and on the other in outer laboratories, pointed out that:

(1) when the registration conditions are kept nearly constant (about 10% T of the most intense band) the process of identification is successful; moreover for such kind of spectra there are no differences between the results for both test sets.

(2) if there are some deviations from the conditions accepted by registration of the test sample, the correlation-coefficient's HQI gives the best results, followed by the scalar-product's HQI. The most sensitive on baseline presence is the sum-of-absolute-values-differences' HQI, followed by the sum-of-least-squares' HQI, both performing worse in such cases.

(3) when the library is not well-balanced on homologs the 'identity' search, like the 'similarity' search, are to some extent hindered to give homologous compounds on the top positions in the hit list.

ACKNOWLEDGMENT

The authors wish to thank the Bulgarian National Fund of Research at the Ministry of Education and Science for the partial financial support through Grant No X-447/94.

REFERENCES

1. Clerc J.T. In: Meuzelaar, H.L.C., Isenhour T.L. eds. *Computer-Enhanced Analytical Spectroscopy.* New York: Plenum Press 1987: 145-162.

2. Luinge H.J. Automated Interpretation of Vibrational Spectra. Vib. Spectrosc. 1990; 1: 3-18.

3. Zupan J. ed. *Computer-supported Spectroscopic Data Bases*, Chichester: Ellis Horwood, Inc, 1986.

4. Lowry S.R., Huppler D.A. Boolean Logic System for Infrared Spectral Retrieval. Anal. Chem. 1983; 55: 1288-1291.

5. Williams S.S., Lam R.B., Isenhour T.L. Search System for Infrared and Mass Spectra by Factor Analysis and Eigenvector Projection. Anal. Chem. 1983; 55: 1117-1121.

6. Ruprecht M., Clerc J.T. Performance Analysis of a Simple Infrared Library Search System. J. Chem. Inf. Comput. Sci. 1985; 25: 241-244.

7. Clerc J.T., Pretsch E., Zuercher M. Performance Analysis of Infrared Library Search Systems. Microchim. Acta 1986; II: 217-242.

8. Wang C.P., Isenhour T.L. Infrared Library Search on Principal-Component-Analyzed Fourier-Transformed Absorption Spectra. Appl. Spectrosc. 1987; 41: 185-194.

9. Kawata S., Noda T., Minami S. Spectral Searching by Fourier-Phase Correlation. Appl. Spectrosc. 1987; 41: 1176-1182.

10. Yu J., Friedrich H. Odd Moments of the Cross-Correlation Function for Library Searching of Infrared Spectra. Appl. Spectrosc. 1987; 41: 869-874.

11. Lo Su-Chin, Brown C.W. Infrared Spectral Search for Mixtures in Medium-Size Libraries. Appl. Spectrosc. 1991; 45: 1621-1627.

12. Lo Su-Chin, Brown C.W. Infrared Spectral Search for Mixtures in Large-Size Libraries. Appl. Spectrosc. 1991; 45: 1628-1632.

13. CDS-3 Applications Software for FT-IR Spectrophotometers, Perkin-Elmer, Norwalk, Connecticut, USA, 1986

14. Perkin-Elmer, Norwalk, Connecticut, USA

15. Savitzky A., Golay M.J.E. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. Anal. Chem. 1964; 36: 1627-1639.

16. Sadtler IR Search 2.25 University, Sadtler Research Laboratories, Division of Bio-Rad Laboratories, Inc.

Received: April 17, 1996
Accepted: May 21, 1996