Bulgarian Chemical Communications, Volume 40, Number 4 (pp. 556–560) 2008

Implementation and testing of routine procedure for mixture analysis by search in infrared spectral library

P. N. Penchev, V. L. Miteva, A. N. Sohou, N. T. Kochev, G. N. Andreev*

Department of Analytical Chemistry, Faculty of Chemistry, University of Plovdiv, 24 Tsar Assen St., Plovdiv 4000, Bulgaria

Dedicated to Academician Ivan Juchnovski on the occasion of his 70th birthday

Receved February 4, 2008; Revised February 14, 2008

A routine procedure for mixture analysis by searching in infrared spectral library using spectra subtraction was implemented and tested. The peak search parameters were optimized to improve the mixture components identification. Four new heuristics have been devised that improved the identification of mixture components.

Key words: infrared spectra, library search, mixture analysis.

INTRODUCTION

The infrared (IR) spectrum reflects to a great extent the compound structure, therefore, it is well suited for the process of structure elucidation [1]. Library search systems are currently commercially available from both the instrument manufacturers and chemical software businesses. Also, spectral libraries containing several hundreds to thousands of spectra can be obtained [2]. Despite that, the number of collected spectra is much smaller than the number of currently known compounds, in some cases a correct identification of a compound or mixture components can be achieved which makes the further study in this field sensible.

The library search in spectral databases has two main goals [3]:

(1) identification of an unknown compound if its spectrum is among the reference spectra (identity search), or

(2) obtaining a list of compounds (called hitlist) whose spectra are most similar to that of the unknown (similarity search). Since the IR spectrum is a function of the structure of the compound, the result hitlist can be used for deriving some conclusions for the unknown's structure, which is usually done through manual inspection of the hit structures by the chemist or with the aid of some computer algorithms as that of maximum common substructure [4].

In this work the identification of mixture components through a search in an IR spectral library is studied. A well-known procedure [5, 6] that uses spectra subtraction is implemented and studied. The procedure is optimized and several heuristics improving considerably the identification of mixture components have been defined.

EXPERIMENTAL

Measurements and Processing of Spectra

The IR spectra were registered on a Perkin-Elmer 1750 FT-IR Spectrometer from 4000 cm^{-1} to 450 cm^{-1} at a resolution 4 cm^{-1} with 16 scans. The solid samples were recorded 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 were transferred to an IBM compatible computer with the standard protocol for data exchange KERMIT [7]. The original spectral data were converted by a smoothing procedure based on weights from a normal distribution.

SPECTRAL LIBRARY AND SOFTWARE

SpecInfo Library

The IR database of this system contains 1000 full-curve spectra together with chemical structures and was available for this work in JCAMP-DX format [8]. The original spectral range is 4000 to 400 cm^{-1}, with a sampling interval of 1.93 cm^{-1}. The IR spectra, structural data, molecular formulas, and compound names were converted for use in the software product IRIS.

* To whom all correspondence should be sent:
E-mail: andreev@uni-plovdiv.bg

IRIS

This is a Windows-based program for searching in libraries of IR spectra [9, 10]. Six different algorithms for comparison of IR spectra are implemented: two methods for peak matching and four methods for comparing full spectral curves. IRIS uses the spectral range from 3700 to 500 cm^{-1}, with a sampling interval of 4 cm^{-1}, corresponding to 801 data points. Furthermore, IRIS contains software tools for the import of IR spectra in JCAMP-DX format, for peak picking, and for an interactive analysis of IR spectra of mixtures based on multiple linear regression.

METHODS

Peak Matching

Peak search algorithms described in Ref. [3] can be generally divided into two types: (1) forward one used for identification of pure compounds, and (2) reverse one applied for identification of the components of organic mixtures. The corresponding spectra similarity measures, or hit quality indices (HQIs), were implemented as it was described in Sadtler IR library [6]:

forward: $HQI_F = ABC$

reverse: $HQI_R = BAC$;

where A, B and C are calculated as it is described below.

If the unknown spectrum contains N peaks, the reference spectrum contains M peaks, and K is the number of matched peaks, then

$A = 9K/N$; $B = 9K/M$;

$C = 9[1 - \Sigma \left| \nu_U^K - \nu_R^K \right| /(K\,\Delta\nu)]$;

where $\Delta\nu$ is the tolerance in peak shift at the abscissa (wavenumber), the sum in C is taken for all matched peaks, and A, B and C are rounded to integers.

In the applied spectroscopy a full coincidence of peaks is rather accidental even for two different IR spectra of the same compound. That is why the peak matching is done with tolerances defined by the user. In IRIS program two tolerances are used: one along the abscissa (wavenumber), $\Delta\nu$, and the other along the ordinate (absorbance), ΔA. The coincidence of a peak from unknown and a peak from reference spectrum means that the peak in the unknown is within a rectangle with sides' lengths $2\Delta\nu$ and $2\Delta A$ and a center in the top of the reference peak (Figure 1).

For comparative purposes, four additional measures were applied to describe the similarity of IR spectra: (1) sum of the squared absorbance differences, (2) sum of the absolute absorbance differences, (3) normalized scalar product of two spectral vectors, and (4) correlation coefficient between spectral vectors. These four measures compare full spectral curves, and the corresponding HQIs (HQI_1 to HQI_4) were described earlier [4].

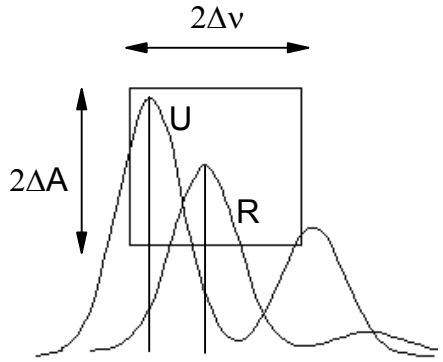

Fig. 1. Matching two peaks, one from the unknown spectrum (U), another from the reference one (R), using tolerances along the abscissa and the ordinate.

Mixture Analysis Procedure

The main requirement for the application of this procedure is that all mixture components are members of the library. At the beginning, the mixture spectrum is searched in the library. The first hit is assumed as one of the mixture components and its spectrum is subtracted from the mixture spectrum, see Eqn. (1). Further, the negative values of the remainder spectrum are truncated, normalized and then it is searched in the library. The first hit in this newly obtained hitlist is supposed to be the second component [5, 6].

$$\text{Remainder} = \text{Mixture} - \text{Coefficient} \times \text{“Hit \#1”} \quad (1)$$

Described in this way, the procedure looks pretty straightforward, but even for a mixture of components with quite different spectra it could fail and give erroneous results. There are no recommenddations in the literature to what extent the subtraction is performed, except that one or more selected spectral bands of the mixture spectrum have to be nullified. Another complication can arise if the mixture components have similar spectra with overlapping bands (because of their similar structures) thus leading to an over-subtraction – and as a result of it – the second component might not be the first hit in the hitlist that results from the second library search.

To test the procedure and to make it robust and reliable, ten mathematically composed mixture spectra and ten recorded mixture spectra are analyzed: all spectra are registered in our laboratory and searched in the SpecInfo library. Since our spectra are not identical with those of SpecInfo, the mathe-

matically composed spectra have all features of the spectra of real samples: they differ from the reference ones in the level of baseline and the band widths, and they are also recorded by quite a different sample path and spectral resolution. Five of them are composed of samples recorded in KBr pellet and this fact has additionally complicated the subtraction.

RESULTS AND DISCUSSION

Optimization of Tolerances

The full spectrum search was studied earlier [11] and proved to be better than the peak search if applied to spectra of pure compounds. Since a mixture spectrum is nearly an algebraic sum of components spectra, the comparison of full spectra gives inadequate result and components spectra are not at the first two positions in the hitlist. On the other side, the tolerance values optimized for identification of pure compounds [9] are expected to be unsuitable for identification of mixture components because the relative intensity of spectral bands decreases in the mixture spectrum and appeared as shifted alongside the wavenumber by the overlapping bands of the components. Because of this, new values of the tolerances have to be established when mixture spectra are searched in the library.

To find the optimal tolerances, ten mixture spectra recorded in our laboratory were searched in the SpecInfo library. They are five mixtures of butanol and 2-methyl-1-propanol (*iso*-butanol) with volume ratio of 1:9, 1:4, 1:1, 4:1 and 9:1 v/v, and five mixtures of pyridine and benzene with volume ratio of 1:9, 1:4, 1:1 v, 4:1 and 9:1 v/v. There IR spectra were searched with $\Delta\nu$ varying from 3 to 20 cm^{-1}; the threshold of peak-picking applied to mixture spectrum was 0.01 a.u. in order to find most of the peaks of the component with smaller concentration. The usage of ΔA less than 1.0 a.u. proved to be unfavorable when there was a component with smaller concentration. In this way, the found optimal tolerances are ΔA = 1.0 a.u. and $\Delta\nu$ in the range of 8 to 11 cm^{-1}: the latter is a little wider than the optimal $\Delta\nu$ found by the pure compound identification (7 cm^{-1}). The HQI_R was better suited than HQI_F for mixture components identification.

The four full spectrum HQIs were also tested for mixture identification. As expected, they performed worse than peak search HQIs.

Identification of Mixture Components

One of the authors (P.N.P) composed five mixture spectra of 2-nitrobenzaldehyde and indole-3-carboxaldehyde with volume ratio of 1:9, 1:4, 1:1, 4:1 and 9:1 v/v, and five mixture spectra of 3,3-dimethyl-2-butanone and propiophenone with volume ratio of 1:9, 1:4, 1:1, 4:1 and 9:1 v/v. For the first five mixtures the components spectra were recorded in KBr pellet and for the next five as thin films. These ten spectra were searched in random order by another author (V.L.M.) who knew neither the components nor the composition of the mixtures. The threshold by pick-peaking from mixture spectra and remainder spectrum is 0.01 a.u., and the reverse peak search was used with $\Delta\nu$ = 9 cm^{-1} and ΔA = 1.0 a.u. The routine procedure described in the Mixture Analysis Procedure section was applied. As expected, the component identification was not a straightforward one, and in eight cases the second component was not identified. The two mixtures with correctly identified components have a volume ratio of 1:1 v/v.

The detail analysis showed several reasons for that misidentification. First, the difference spectrum is too noisy and has some kind of "wings" that result from the different band widths of the component spectrum and the reference one. Second, the spectroscopist selects usually one band from the mixture spectrum and tries to nullify it by subtracting the first hit spectrum. There is no warranty that the selected "nullified" band is not present in the second component spectrum or severely overlaps with one of its bands: in this case, substantial spectral information is lost in the remainder. Third, an unambiguous criterion when to stop the subtraction was not used. The smaller the coefficient from Eqn. (1), the more spectral features remain from the first hit; the bigger the coefficient, the less spectral features remain from the spectrum of the second component. The presence of some "wings", positive or negative – see Fig. 2, is unavoidable and they harm the search. That is why it is better to minimize their magnitude.

To overcome these problems three heuristics were formulated and applied by the next test searches of other ten mixture spectra. They are:

(1) By the second search, i.e. by the search of the remainder, a higher threshold is applied by the peak-picking procedure. Our experiments proved that instead of 0.01 a.u., the threshold has to be between 0.03 and 0.05 a.u.

(2) Several bands (not only one) are supervised by the subtraction. These are the bands which have close relative intensities in the mixture spectrum and in the reference one.

(3) The subtraction is done till the bands selected by heuristic 2 give equal positive and negative "wings" as is shown in Fig. 2.

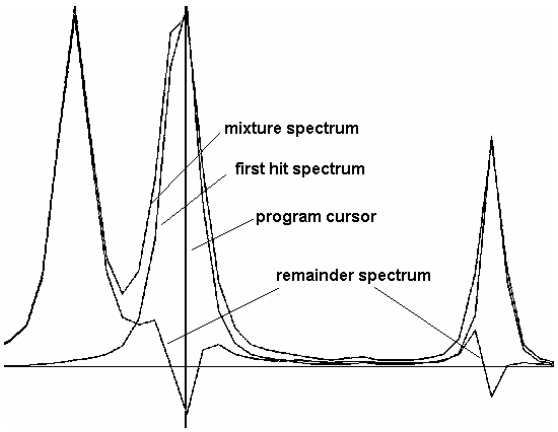

Fig. 2. Illustration of the third heuristic.

To test the significance of these heuristics ten mixture spectra recorded in our laboratory were searched in random order by one of the authors (V.L.M.) who knew neither the components nor the composition of the mixtures. The thresholds by pick-peaking from mixture spectra and remainder spectrum were, respectively, 0.01 a.u. and 0.05 a.u., and the reverse peak search was used with $\Delta\nu = 9$ cm^{-1} and $\Delta A = 1.0$ a.u. The routine procedure described in the Mixture Analysis Procedure section was applied. The results are presented in Table 1.

For seven out of ten mixtures, both components were correctly identified. Not surprisingly, the failed identification is for mixtures 1 and 4. Mixture 1 gives nearly "null" (flat) spectrum after the subtraction, and that illustrates the limits of the routine to the range of concentration of the components. The second component, *o*-xylene, of mixture 8 was also not identified: it was positioned at the 43 place in the second hitlist. One possible explanation might be that the spectrum of *o*-xylene is a subset of the spectrum of the *iso*-propylbenzene (first identified component). The most intensive band of *iso*-propylbenzene spectrum is at 700 cm^{-1} and it is not overlapping with any *o*-xylene bands. This band was used by subtraction and the other bands of *o*-xylene disappeared obviously in the remainder. Mixture 2 was successfully analyzed: it has a coefficient of 0.99 despite the predominant component has volume percentage of 80%. The normalization of a spectrum (unknown or reference) in IRIS is between 0.0 a.u. (for the background) and 1.00 a.u. (for the most intensive band). In a mixture spectrum, the most intensive band can solely be the band of one of the components. That is why by an idealized spectrum measurement, when the path lengths of mixture sample and reference spectrum sample are equal, the coefficient from Table 1 has to be close to 1.00. This could be regarded as another heuristic by performing the spectra subtraction.

As the remainder spectrum (see Eqn. (1)) is very noisy, one could expect that the full spectral HQIs will give better results when applied only to the second search. This did not happen, and one possible explanation is the partially remaining spectrum of the first identified component.

Table 1. Mixture component identification. The component is considered as identified if it was the first hit by the corresponding search. In parentheses the corresponding HQI is given.

#	Mixture	Volume ratio (v/v)	Found 1 component and (HQI)	Coefficient from Eqn. 1	Found 2 component and (HQI)
1	Butanol and *iso*-butanol	1:9	*iso*-butanol (979)	1.02	- [1)]
2	Butanol and *iso*-butanol	1:4	*iso*-butanol (978)	0.99	butanol (648)
3	butanol and *iso*-butanol	1:1	*iso*-butanol (877)	1.01	butanol (648)
4	butanol and *iso*-butanol	4:1	butanol (968)	0.96	-
5	*o*-xylene and *m*-xylene	1:1	*o*-xylene (949)	0.86	*m*-Xylene (888)
6	*o*-xylene and *p*-xylene	1:1	*p*-xylene (948)	1.06	*o*-Xylene (778)
7	*m*-xylene and *p*-xylene	1:1	*m*-xylene (938)	1.10	*p*-Xylene (879)
8	*iso*-propylbenzene and *o*-xylene	1:1	*iso*-propylbenzene (949)	1.15	-
9	*iso*-propylbenzene and *m*-xylene	1:1	*iso*-propylbenzene (958)	1.08	*m*-Xylene (848)
10	*iso*-propylbenzene and *p*-xylene	1:1	*iso*-propylbenzene (948)	1.13	*m*-Xylene (878)

[1)] the component was not found as the first hit by the corresponding search.

CONCLUSIONS

A routine procedure for mixture analysis by searching in infrared spectral library was implemented and tested. The peak search parameters were optimized to improve the mixture components identification. The procedure that uses spectra subtraction appeared to be not a straightforward one; that is why four new heuristics were devised that improved the identification of mixture components:

(1) A higher threshold has to be applied by the peak-picking procedure of the remainder;

(2) Several bands have to be supervised by the subtraction;

(3) The subtraction has to be performed until the bands selected by heuristic 2 give equal positive and negative "wings";

(4) Values near to 1.00 of the coefficient from Eqn. (1) have to be expected.

REFERENCES

1. H. J. Luinge, *Vib. Spectrosc*., **1**, 3 (1990).
2. W. A. Warr, *Anal. Chem*., **65**, A1087 (1993).
3. J. T. Clerc, in: Computer-Enhanced Analytical Spectroscopy, H. L. C. Meuzelaar, T. L. Isenhour (Eds.), Plenum Press, New York, 1987, pp. 145-162.
4. K. Varmuza, P. Penchev, H. Scsibrany, *J. Chem. Inf. Comp. Sci*., **38**, 420 (1998).
5. H. Somberg; Qualitative Mixture Analysis by Use of an Infrared Library Search System, Brucker Reports, 1988.
6. The Sadtler IR Search Software Manual, Sadtler Research Labs, Division of Bio-Rad Laboratories, Inc, 1988.
7. D. L. Eisner, C. S. Mayfield, IDRIS KERMIT, version 1.0, Perkin-Elmer Corp., July, 1985.
8. SpecInfo: Spectroscopic Information System, vers. 3.1, 1996; Available from: Chemical Concepts, P.O. Box 100202, D-69442 Weinheim, Germany.
9. P. N. Penchev, PhD. Thesis, University of Plovdiv, 1998.
10. IRIS can be downloaded together with a small demo library and all used mixture spectra from http://www.kosnos.com/spectroscopy/iris.
11. P. N. Penchev, A. N. Sohou, G. N. Andreev, *Spectrosc. Lett*., **29**, 1513 (1996).

РЕАЛИЗИРАНЕ И ПРОВЕРКА НА РУТИННА ПРОЦЕДУРА ЗА АНАЛИЗ НА СМЕСИ С ПОМОЩТА НА БИБЛИОТЕКА ОТ ИНФРАЧЕРВЕНИ СПЕКТРИ

П. Н. Пенчев, В. Л. Митева, А. Н. Сохоу, Н. Т. Кочев, Г. Н. Андреев*

Катедра „Аналитична химия", Химически факултет, Пловдивски университет, ул. „Цар Асен" № 24, Пловдив 4000

Посветена на акад. Иван Юхновски по повод на 70-та му годишнина

Постъпила на 4 февруари 2008 г.; Преработена на 14 февруари 2008 г.

(Резюме)

В статията е реализирана и проверена рутинна процедура за анализ на смеси с помощта на библиотека от инфрачервени спектри. Параметрите за търсене по пикове са оптимизирани за да подобрят идентификацията на компонентите на смеси. Формулирани са четири нови евристики, които подобряват идентификацията на компонентите на смеси.