Analytica Chimica Acta, 122 (1980) 357—361
Computer Techniques and Optimization

SYSTEMATIC COMPUTER-AIDED INTERPRETATION OF VIBRATIONAL SPECTRA

T. VISSER and J. H. van der MAAS*

Laboratory for Analytical Chemistry, University of Utrecht, Croesestraat 77A, 3522 AD Utrecht (The Netherlands)

(Received 7th November 1979)

SUMMARY

The interpretation process for vibrational spectra is considered in detail. The concepts of spectral data, structural elements and basic file are described. Two parameters proved to be useful for expressing the degree of correlation of wavenumber regions and structural elements. Three types of intervals are distinguished. A computer program has been developed to obtain these intervals. Factors influencing the results of an interpretation system are discussed and possible criteria are reported.

The interpretation of vibrational spectra requires ample knowledge about correlations between structural elements and spectral data. Frequency or wavenumber, intensity (weak, medium, strong), half bandwidth (broad, sharp, α/β value) and band shape (wedge in carboxylic acids) are used in conjunction with structural elements that may vary from well-defined functional groups like OH, C=C, C=O to functionalities such as cyclohexyl, *para* substitution, etc.

A huge amount of correlations, especially for infrared spectroscopy, is available in the literature. Most of it has been included in Bellamy's books [1, 2] on infrared spectroscopy and the book of Dollish et al. [3] on Raman spectroscopy. Many Colthup tables are also available.

Probably because of the physical impossibility of having that amount of information ready at hand in one mind, interpretation is still an art that has to be learned from experience and that cannot be transferred simply from one person to another. Systematic interpretation would not only exclude accidental and irrelevant factors but would also afford maximum information yield [4, 5]. With the introduction of digitized infrared and Raman spectrophotometers and the decreasing cost of minicomputers, automatic fast interpretation is within reach [6] provided that the software is available. Experience with the development of interpretation systems, including programming, has revealed an unmistakable need for clear-cut rules, which in turn require greater insight into the different relevant factors. A detailed study of the interpretation process seemed to be a logical step.

GENERAL CONCEPTS

The essential feature of systematic interpretation of spectra is to make efficient use of correlations. As the existence of correlations depends on the different types of structural elements and the nature of the spectral data, a study of the development of an interpretation system must be preceded by clear arrangements in these respects.

Spectral data

In practice, the wavenumber of an absorption band is the most conspicuous piece of spectral data and is easy to determine. The intensity of a peak is also useful, but conditions concerning the minimum detection limit, the use of absolute or relative intensity, etc., are required. Other spectral parameters are more difficult to obtain and for that reason the data used are generally limited to wavenumber (regions) and transmittance (thresholds).

In order to be able to use spectral data optimally, recording has to be done under identical scanning conditions (e.g., a wavenumber accuracy of ± 2 cm^{-1}). Although a large amount of spectral data is available in the literature (e.g., Sadtler [7] and DMS [8]), it cannot be used as the scanning conditions are not fulfilled. Therefore it is necessary to compose a basic file of spectra.

Basic file

As the aim of an interpretation system is the detection of structural elements (and not retrieval), a basic file can be composed of a limited number of spectra. For a deliberate selection of compounds, chemical and spectroscopic knowledge, and thus experience, is essential. In order to be able to correlate a structural element with a spectral region, two spectra representing the extreme frequencies of that element will do. Thus the number of structural elements defines the number of spectra in the file. At first glance, further reduction of the amount of data seems possible by combining several elements within one compound but this is not really viable for reasons set out in a later section. The file may, of course, be large, the only objection being that more computer time and space is required. Compounds containing other elements than those to be investigated may be present. In fact, the spectral assembly has to reflect the compounds for which the system will be used.

Structural elements

A structural element is an entity of at least two bonded atoms. For carbon, hydrogen and oxygen, for example, the minimum number of different two-atomic entities is six: CC, CH, CO, HH, OH and OO. However, several types of bonding can be distinguished and so the combinations CC, CO and OO have to be increased to C—C, C=C, C≡C, C—O, C=O, O—O and O=O. Since the combination O=O represents the oxygen molecule and HH the hydrogen molecule these two can be omitted. In this way, the total

number of so-called "main" elements reduces to eight. Other possible combinations will consist of more than two atoms and can be regarded therefore as derivatives or sub-elements of those eight. For example, C≡C—H is a sub-element of C≡C. In this way all elements can be classified systematically.

Correlations

A structural element S_a is said to be correlated if its presence in a compound is generally associated with the appearance of at least one peak within a designated region of the spectrum of that sample. If a file is composed of n compounds, of which m have the element S_a, then if all compounds with S_a show a peak in a particular region, the correlation is at a maximum and the score percentage (*SP*) is $m \times 100/m = 100\%$. As each compound will have at least one peak in a complete spectrum, a 100% score can always be reached if the interval is broad enough. However, with the score percentage, the chance of having interfering peaks of elements different from S_a increases. With l interfering compounds, the interfering percentage (*IP*) is $(l/n—m) \times 100\%$. With *IP* and *SP*, it is possible to distinguish three kinds of regions: (a) $SP = 100\%$ and $IP = 0\%$, which is specific; (b) $SP = 100\%$ and $IP > 0\%$, which is selective; and (c) $SP < 100\%$ and $IP = 0\%$, which is pseudo-specific.

As the intensity threshold determines the presence or absence of a peak, the finding of a region, including *SP* and *IP* and thus its type, is related to that threshold. To find correlations for a structural element in relation to a certain intensity threshold, a program CRISE (Correlation of Raman and Infrared data with Structural Elements) has been developed. The program supplies for any element S_a, at a chosen intensity threshold, all possible (pseudo-)specific and selective regions, provided that the presence/absence of that particular S_a in each compound of the basic file is known. For the pseudo-specific and selective regions, the *SP* and *IP* values, respectively, are calculated.

RESULTS AND DISCUSSION

From preliminary results with CRISE, it appeared that more regions may be found than would be expected theoretically and also that a specific region is very rare; even an *SP* of 70% proved to be rather high. There are various reasons for this. First, regions, which could have been specific, are subdivided into some pseudo-specific regions because of the presence of a few interfering peaks ($\neq S_a$). Secondly, the simultaneous presence of an element S_b ($\neq S_a$) next to S_a is inevitable in a number of compounds and therefore regions may be found which have nothing to do with S_a but all with S_b; these intervals must be excluded. Thirdly, when no maximum is set for the interval width, a large number of very broad selective regions is found. These intervals are chemically and spectroscopically meaningless as functionalities rarely exceed a bandwidth of 300 cm^{-1}. Fourthly, regions may be found that are strongly correlated with vibrations of sub-elements of S_a. These will also be found

when the sub-element itself is investigated and may then even appear to be specific. From this last consideration, it follows that all regions which are (pseudo-)specific or selective for a certain sub-element also form part of the (pseudo-)specific and selective intervals for the corresponding main element. So correlations which are found for S_a may indicate the presence of useful regions for its sub-elements.

If no specific region is found for S_a, one can try to reach a 100% score by combining two or more pseudo-specific regions. With CRISE, it is possible to combine up to five regions (if present). They can be selected interactively. The score *SP* of each of the combinations is calculated. Thus the system has to be constructed from specific, selective and (a combination of) pseudo-specific regions.

The system

The program CRISE (which is available from the authors on request) supplies all data necessary for composing an interpretation system. In the system, regions are used in the form of questions, hence they will be referred to as *Q*'s in the following paragraphs.

A fundamental requirement of a system must be that there are no wrong answers. Another condition may be that for any structural element a certain score percentage (*SP*) has to be reached, which can be used as a criterion. It may be reached either with one or with a combination of pseudo-specific *Q*'s. The maximum number of *Q*'s one is prepared to combine (*mQ*) can be used as another criterion. Similarly to *SP* for pseudo-specific cases *IP* can be used for selective *Q*'s. These three criteria (*SP*, *IP* and *mQ*) and the variable intensity threshold determine the system. Accordingly, the following types of *Q*'s can be distinguished in a system: (a) specific: supply complete certainty about the presence or absence of S_a (Fig. 1a); (b) pseudo-specific: may supply certainty about the presence of S_a (Fig. 1b); (c) combined: as (b) (Fig. 1c); (d) selective: may only supply certainty about the absence of S_a (Fig. 1d).

All these questions can be put at any place in the system and they are therefore called unconditioned questions. Thus, results from CRISE that fulfil the mentioned criteria can be simply composed into a system.

Conclusions

Comparison of the system based on the results of CRISE and the earlier developed systems [4, 5] showed that the latter also contain combined answers such as "S_a and/or S_b present". Apart from the usefulness of this kind of information, the data necessary to find such types of question are present implicitly in the results of CRISE; to expose them a modification of the program is required. For the choice of variables and the criteria, and thus for the composition of the system, there is no difference in principle.

In summary, a viable system must be defined by several variables, viz. the basic file, scanning conditions, intensity threshold and the criteria for *SP*, *IP*

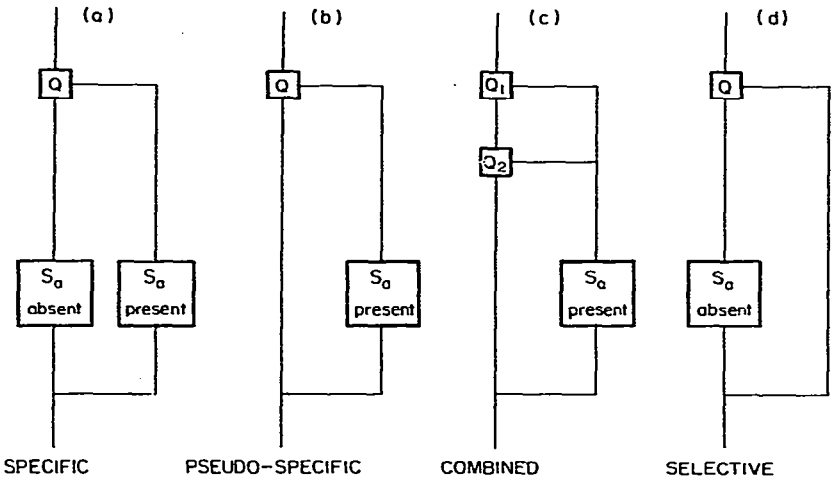

Fig. 1. The four types of questions in an interpretation system. The small squares are question elements (Q's) (positive to the right, negative downwards); the large squares are answers. S_a = structural element.

and mQ. It must be emphasized once more that without a solid knowledge of chemistry and vibrational spectroscopy a deliberate choice of the variables is impossible. An objective comparison of interpretation systems is possible on the basis of the criteria given in this paper.

REFERENCES

1 L. J. Bellamy, The Infrared Spectra of Complex Molecules, Chapman and Hall, London, 1975.
2 L. J. Bellamy, Advances in Infrared Group Frequencies, Methuen, London, 1968.
3 F. R. Dollish, W. G. Fateley and F. F. Bently, Characteristic Raman Frequencies of Organic Compounds, Wiley, New York, 1974.
4 T. Visser and J. H. van der Maas, J. Raman Spectrosc., 7 (1978) 278.
5 C. G. A. van Eijk and J. H. van der Maas, Fresenius Z. Anal. Chem., 291 (1978) 308.
6 J. P. Coates and S. Geary, Anal. Chim. Acta, 103 (1978) 303.
7 The Sadtler Standard Spectra, Sadtler Research Laboratories Inc., Philadelphia.
8 DMS Documentation of Molecular Spectroscopy, Verlag Chemie, Weinheim and Butterworth, London, 1975.