Joint Spectral Database of Infrared and Raman Spectra

Slava Tsoneva, Stefka Nachkova, Plamen Penchev

Abstract: *A spectral database of 185 artificially made spectra of organic compounds is composed by averaging the corresponding IR and Raman spectra. The database is maintained with the original software for library search, IRSS. The similarity search is proved to be better with these average spectra. The other advantages are the usage of the same software, reducing the number of search sessions performed by the user and hereof the twice-increased speed of search if applied to substantially bigger databases.*

Key words: *Library search; IR spectra; Raman spectra.*

INTRODUCTION

The infrared (IR) and Raman spectra reflect in a great extent the compound's structure, therefore both techniques are very relevant for the process of structure elucidation via library search routines [1-3]. The library search procedure consists of comparison of a spectrum of the unknown with a collection of reference spectra of known compounds. The collection of spectra together with a chemical information and structure of the compounds is called a spectral library. The result obtained by a search, called a hitlist, is a list of spectra that are most similar to the unknown spectrum. The hits in the hitlist are sorted according to a real number called hit quality index (HQI) which reflects the spectral similarity between unknown and reference spectrum. If the unknown is among the library entries, then the correct answer often appears among the first several hits and a visual inspection of them makes it possible to identify the unknown: this is called identity search [4]. However, if the unknown compound is not in the spectral library, a more sophisticated interpretation of the hitlist is necessary. Assuming that similar spectra indicate similar structures, the hitlist structures characterize the unknown structure: this is the quintessence of the similarity search [4].

Important aspects for both types of searches are the representation of the spectra and the applied spectral similarity measures [2]; for the similarity search another key characteristic is the method of analysis of the hitlist structures. Other characteristics of a library search system include the speed and versatility of the implemented search algorithms, the size, contents, and reliability of the database, the options for an update of spectral libraries, the availability of modules for analysis of the hitlist entries and the possibility to derive spectrum-structure correlations [5].

In the Sadtler software KnowItAll a system for searching multiple spectral techniques simultaneously has been introduced [6]: several spectra from complementary analytical techniques are used as unknowns to search multiple databases containing spectra from the complementary analytical techniques. The hit quality indices are plotted in 2D graph that allows a better separation of the correct hit from the rest of the results.

The present paper reports on a preparation of joint database of Fourier transform infrared (IR) and Raman spectra in a quite unusual way. The "spectrum" of a library compound is an average of the corresponding IR and Raman spectra thus preserving features of both types of spectra. The user loads the vibrational spectra of the unknown, calculates their average and searches the latter in the library of average IR and Raman spectra.

METHODS

The method for library search of IR and Raman spectra is implemented into a Windows-based, user-friendly, program called IRSS [7]. The program contains

software tools for an import of IR spectra in JCAMP-DX format, peak picking, and an interactive analysis of IR spectra of mixtures based on multiple linear regression techniques.

The database spectra are in the range from 500 to 3700 cm^{-1} with a sampling interval of 4 cm^{-1}. This corresponds to 801 data points of IR absorbance or Raman scattering intensity, both recorded as byte (values from 0 to 255). Seven different algorithms for the comparison of IR and Raman spectra are implemented: three methods for matching peaks [8] and four methods for comparing full spectral curves [9]. In this study, the used spectral similarity measure, HQI_4, is the correlation coefficient of spectral curves [9], Eq. 1.

$$HQI_4 = 999(S_4 + 1)/2; \qquad S_4 = \frac{\sum_k (A_k^U - \overline{A^U})(A_k^R - \overline{A^R})}{\sqrt{\sum_k (A_k^U - \overline{A^U})^2 * \sum_k (A_k^R - \overline{A^R})^2}}, \qquad \text{(Eq. 1)}$$

where A_k are the absorbance values in the spectra (801 such values); U = unknown spectrum, R = reference one; the bar over a variable indicates its average value.

To evaluate the performance of similarity search in spectral libraries, 500 binary structural descriptors are calculated by software SubMat [10]. For each library structure a set of substructures is searched in the structure and the output text file represent a so called descriptor matrix. It is composed of zeros and ones, with a size of 500 columns and the number of rows is equal to the number of library spectra. A row of this matrix characterizes the structure of the corresponding library entry and is called a fingerprint. If a descriptor matrix element is designated as $d_{k,m}$, the similarity between two structures of entries with numbers k and n is evaluated by the Tanimoto index [11], Eq. 2.

$$Tan_{k,n} = \frac{\sum_m d_{k,m} \text{ and } d_{n,m}}{\sum_m d_{k,m} \text{ or } d_{n,m}}, \qquad \text{(Eq. 2)}$$

EXPERIMENTAL

More than 270 Raman spectra are recorded with a Ram II spectrometer (Bruker Optics); the spectra are measured from 4000 cm^{-1} to 50 cm^{-1} at resolution 2 cm^{-1} with 25 scans. For solid samples the stirred crystals of compound are placed in aluminium disc and the liquids are measured in NMR tubes. Another 116 Raman spectra are used with the kind permission of Dr. Gennady Gudy (Julius Kühn-Institute, Beriln).These spectra are recorded on an RFS-100 Bruker FT-spectrometer. After removing the bad spectra, a library of 330 Raman spectra is composed. Structure search among our other spectral libraries [12] has revealed that there are 185 pairs of IR and Raman spectra of the same compound. Thus an IR and a Raman library are composed for these 185 organic compounds. All used IR libraries and their compilation have been described in detail in [12]. The software IRSS and the spectral libraries are available on request from one of the authors (P.P.).

RESULTS AND DISCUSSION

There exist many ways of combining two different types of spectra if each spectrum is represented as a vector with the spectrum ordinate values as vector components. The simplest one is to augment one spectrum with the other [13], i.e. composing a vector with the leftmost components from one spectrum and rightmost from the other. The result vector preserved the information from each individual

spectrum; it proved very useful when IR and mass spectra together with melting and boiling points had been used for pattern recognition [13].

The augmentation of IR with Raman spectra from our databases will results in 1602-component vectors but that data format is not supported by our software IRSS. One solution to that problem is a reduction of sampling interval of both types of spectra to 8 cm^{-1} but then this will complicate the peak-search routines. The last could be avoided by artificially "shifting" the spectral range of one of the spectrum (e.g. Raman) but another complication would result for the user when he/she is manually entering peaks in the peak-search routine.

As in our software IRSS the IR and Raman spectra are represented with the same sampling interval and in the same spectral range, the ordinate value at a given abscissa value (wavenumber) of the new spectral representation can be calculated from the corresponding ordinate values in IR and Raman spectra. If ordinate value is designated as A_k, k = 1, 2 … 801 (resp. at 500, 504 … 3700 cm^{-1}) the reasonable combinations between IR and Raman spectra are:

averaging: $A_k^{new} = (A_k^{IR} + A_k^{Ra})/2$ (Eq. 3)

multiplication: $A_k^{new} = (A_k^{IR} \times A_k^{Ra})$ (Eq. 4)

subtraction: $A_k^{new} = (A_k^{IR} - A_k^{Ra})$ (Eq. 5)

combination of addition and subtraction: $A_k^{new} = (A_k^{IR} + A_k^{Ra}) \times (A_k^{IR} - A_k^{Ra})$ (Eq. 6)

As the library spectra are range-scaled in 0-1 interval, the spectra resulted from Eq. 3 - Eq. 6 need to be range-scaled in 0-1 interval.

Four spectral libraries are composed with the help of Eq. 3 – Eq. 6: IRRaAv, IRRaMu, IRRaSu and IRRaAS. The size of each one is 185 "spectra" and it is too small. The last hinders the check of the performance of identity search and comparison between usage of the original IR and Raman libraries and the new four ones. On the other side, the similarity search in these six libraries can be compared.

There exist 17020 (= 185 x 184 / 2) entry pairs in a library of 185 entries. For each of these pairs there are calculated both the spectral similarity by Eq. 1 and the structural similarity by Eq. 2. It is expected that if a pair has a high spectral similarity it will also have a high structural similarity, and vice versa, a pair with a low spectral similarity will have a low structural similarity. Table 1 contains Pearson correlation coefficients between spectral-similarity pairs and structural similarity pairs for the six spectral libraries. The higher the correlation coefficient, the better the corresponding type of spectra reflects the structural features.

Spectral Library	Correlation coefficient
IR	0.513
Raman	0.479
IRRaAv	0.538
IRRaMu	0.444
IRRaSu	0.429
IRRaAS	0.413

Table 1. Pearson correlation coefficients between spectral-similarity pairs and structural-similarity pairs.

As can be seen from Table 1 the IR spectra are more structurally-informative than the Raman spectra. It is well known that some IR characteristic bands do not appear in the Raman spectrum as it is the case with ν(O-H), see Fig. 1. The average between an IR spectrum and Raman one contains spectral bands from both spectra and thus correlates with structure more than each of them alone; hence the highest coefficient in Table 1. The multiplication of both types of spectra has a drawback that if no band is present in either spectrum, there is no band in the product spectrum. This means that for an organic compound with a center of symmetry (such as

benzene) a very queer spectrum is produced with virtually no bands; obviously the reason for that is the rule of mutual exclusion [3]. The other two mathematical operations, Eq. 5 and Eq. 6, annihilate bands that appear in both spectra thus reducing information contents in the result spectrum. Surprisingly, the corresponding correlation coefficients are not so lower than others: one possible explanation is that the relative bands intensity is very different in IR and Raman spectra.

The IR and Raman spectra of 1-decanol and their average are shown in Fig.1. As can be seen, the hydroxyl stretching band is missing in the Raman spectrum: it is at 3328 cm^{-1} in the IR spectrum. Obviously, the average of both vibrational spectra has that band. As a rule, a Raman spectrum of an aliphatic alcohol resembles that of an alkane.

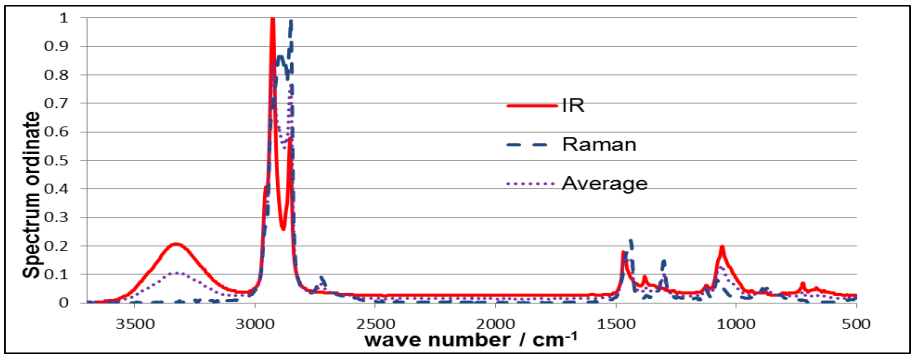

Figure 1. The IR and Raman spectra of 1-decanol and their average.

The search results produced by the three spectra from Fig.1 are given in Table 2. Each spectrum is searched in the library which corresponds to its type. As the library is small, only the first five hits are considered and the searched spectrum (actually the first hit) is removed from the hitlist.

Hit #	IR	Raman	Average
1	1-Nonanol	Dodecane	1-Nonanol
2	1-Dodecanol	1-Nonanol	1-Dodecanol
3	1-Hexadecanol	1-Dodecanol	Dodecane
4	2-Ethylhexane-1-ol	1-Heptanal	5-Nonanol
5	5-Nonanol	5-Nonanol	Hexane

Table 2. The search results with the IR and Raman spectra of 1-decanol and their average spectrum.

It is well known that structural similarity is a vague concept: any two organic chemists will stress on different structural features when asked to evaluate it. The “unknown” compound in this example is a primary aliphatic alcohol with a straight ten-carbon chain. The organic chemist usually regards deriving the chemical class – alcohol in this case – as very important when dealing with structure elucidation. That is why, in this example the Raman spectrum produces the worst results followed by the average spectrum. On the other side, the structural features, encoded in the descriptor matrix, are more objective criterion for the evaluation of the structural similarity than our subjective reasoning made above, and moreover, it is only one example.

CONCLUSIONS

The created spectral database by averaging IR and Raman spectra is a useful and helpful addition to the already created databases of vibrational spectra. It can be used with the software IRSS previously developed in our lab. It reduces the number of search sessions performed by the user and thus increases the search speed if applied to substantially bigger databases.

Acknowledgment

This work has been supported by the Plovdiv University Project NI15ChF001. We are grateful to Prof. Kurt Varmuza (Technical University, Vienna) for providing the software SubMat and Dr. Gennady Gudy (Julius Kühn-Institute, Beriln) for some of the Raman spectra.

REFERENCES

[1] Zupan J. (Ed.). Computer-supported Spectroscopic Data Bases, Chichester: Ellis Horwood, Inc, 1986.

[2] Luinge, H. Automated Interpretation of Vibrational Spectra. *Vib. Spectrosc*. 1990, 1, 3-18.

[3] Larkin P. Infrared and Raman Spectroscopy. Principles and Spectral Interpretation, Amsterdam: Elsevier, 2011.

[4] Clerc, J.T. 1987. Automated spectra interpretation and library search systems. pp. 145–162; In: Meuzelaar, H.L.C. Isenhour, T.L. (Eds.), Computer-Enhanced Analytical Spectroscopy. New York: Plenum, 1987.

[5] Debska, B., B. Guzowska-Swider, D. Cabrol-Bass. Automatic Generation of Knowledge Base from Infrared Spectral Database for Substructure Recognition. J. Chem. Inf. Comp. Sci., 2000, 40, 330-338.

[6] Banik, G., T. Abshear, K. Nedwed. Multi-Techniques Spectral Searchig in KnowItAll, Technical Note. Bio-Rad Laboratories, Inc., Informatics Division, 2005.

[7] Penchev, P., N. Kochev and G. Andreev. IRSS: A Programme System for Infrared Library Search. Comptes Rendus de l'Academie Bulgare des Sciences, 1998, 51, 67-70.

[8] Penchev, P., V. Miteva, A. Sohou, N. Kochev, G. Andreev. Implementation and Testing of Routine Procedure for Mixture Analysis by Search in Infrared Spectral Library. Bulg. Chem. Commun., 2008, 40, 556-560.

[9] Varmuza, K., P. Penchev, H. Scsibrany. Maximum Common Substructures of Organic compounds Exhibiting Similar Infrared Spectra. J. Chem. Inf. Comp. Sci., 1998, 38, 420-427.

[10] Scsibrany, H., M. Karlovits, W. Demuth, F. Muller, K. Varmuza. Clustering and similarity of chemical structures represented by binary substructure descriptors. Chemometr. Intell. Lab. Syst., 2003, 67, 95-108.

[11] Varmuza, K., P. Filzmoser. Introduction to Multivariate Statistical Analysis in Chemometrics. Boca Raton: CRC Press, 2009.

[12] Penchev, P., S. Tsoneva, Ts. Krusteva and S. Nachkova. Spectral Libraries of Viabrational Spectra. Scientific Researches of the Union of Scientist in Bulgaria – Plovdiv, Series B, Natural Sciences and the Humanities. 2014, 16, 79-84.

[13] Джурс, П., Т. Айзенауэр. Распознавание образов в химии. Москва: Мир, 1977.

About the authors:

Correspondence author: Plamen Nikolov Penchev, Assoc. Prof., PhD, Chemical Faculty, University of Plovdiv, E-mail: plamen@uni-plovdiv.net.

Slava Christova Tsoneva, Ph.D. student, Chemical Faculty, University of Plovdiv

Stefka Rumenova Nachkova, Assist. Prof., Chemical Faculty, University of Plovdiv.

This paper has been reviewed