Bulgarian Chemical Communications, Volume 49 Special issue G (pp.61 –64) 2017

Spectral similarity versus structural similarity: Raman spectra

P. N. Penchev*, S. H. Tsoneva, S. R. Nachkova

Department of Analytical Chemistry and Computer Chemistry,

University of Plovdiv "Paisii Hilendarski", 24 Tzar Asen Str., 4000 Plovdiv, Bulgaria,

Received November 13, 2016; Revised November 28, 2016

The study of the relation of spectral similarity and structural similarity shows that the search in Raman spectral libraries gives hitlist compounds structurally similar to the unknown and that the peak search performs equal with the full spectrum search algorithm based on the correlation coefficient. As a result of the study the optimal values for peak search tolerances were found and they are higher than those suited for identity search.

Key words: *Raman spectra, spectral similarity, structural similarity*

INTRODUCTION

Despite the advance of various 1D and 2D NMR experiments [1] and hyphenated techniques in chromatography [2] the Fourier transform (FT) mid-infrared (MIR) absorption spectroscopy is still the most used routine method in organic labs for the confirmation of structures or identification of known organic compounds (either pure or in mixtures). Nowadays, as an analytical technique, FT-Raman spectroscopy offers many pluses over FT-MIR spectroscopy [3]. The most important of them are the following: (1) little or no sample preparation is required; (2) water as a liquid is a weak scatterer – no special accessories are needed for measuring aqueous solutions; (3) water vapors and carbon dioxide are very weak scatterers – sample compartment purging is unnecessary; (4) Raman bands are narrower and both the overtones and combination bands are generally weak; (5) the spectral range reaches well below 400 cm^{-1}, making the technique applicable to organic compounds containing heavier elements; (6) The symmetric molecular vibrations which appear as low-intensity bands in the IR spectrum exhibit very strong Raman bands. Also here has to be mentioned that the construction of Raman spectral libraries is widely in progress [3].

The current study of ours cast light on the capability of so called *similarity search* in Raman spectral libraries to obtain a list of compounds (called *hitlist*) whose structures are most similar to that of the unknown. The similarity search capability of IR and Raman databases has been extensively explored since late 1970s [4] and has been proven to exist. Later, a practically oriented approach of extracting large and frequent substructures contained in the IR hitlist structures has been developed [5]. A successful attempt has been made by Varmuza and co-workers to formalize and systematize the study of structure/spectrum relation for IR [6] and low resolution mass-spectra [7]. In the present paper we apply the method of Varmuza *et al*. [6] and extend it to Raman spectra.

SPECTRA AND METHODS

The Raman database used consists of 330 Raman spectra of organic compounds. The Raman spectra are measured in our lab on RAM II (Bruker Optics) with a focused laser beam of Nd:YAG laser (1064 nm) from 4000 cm^{-1} to 51 cm^{-1} at resolution 2 cm^{-1} with 25 scans. The compound structures are represented as connection tables and binary substructure descriptors (0/1) are calculated by software *SubMat* [6] (supplied by Prof. Varmuza) using a set of 500 substructures. The similarity of chemical structures is calculated with these descriptors by the commonly used Tanimoto index [8] as it is done in [6-7].

The comparison of Raman spectra is done by two different spectral similarity measures, SpSim$_1$ and SpSim$_2$, given by Eqns. 1 and 2, respectively.

$$\mathrm{SpSim}_1 = 2K / (M + N) \tag{1}$$

$$\mathrm{SpSim}_2 = \frac{\sum_k A_k^U A_k^R}{\left\|A^U\right\| \cdot \left\|A^R\right\|}, \tag{2}$$

where K is the number of matched bands in the two spectra (U and R) each of them consisting of M and N spectral bands. The two expressions in the denominator of Eqn. 2 are the Euclidean norm of the vector composed of the corresponding peak intensities [9]. In our library search software IRSS [5, 9] the spectral bands are represented with their intensity and wavenumber only. Two peaks are

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

regarded as matched if one falls within the two intervals centered around the other peak. First interval has a width equal to twice the wave number tolerance, Δν, and the second twice the intensity tolerance, ΔI. The peak matching is illustrated in detail in our previous paper [10]. For comparison purpose, a full spectrum search algorithm based on the correlation coefficient is used [5].

The criterion for capability of the similarity search, *grand mean of Tanimoto indices*, T(h), is borrowed from Varmuza *et al*. [6] (Eqns. 11 and 12 in the cited paper). With 'h' is designated the number of hits which are used for finding the corresponding Tanimoto index average values.

RESULTS AND DISCUSSION

A general opinion among the spectroscopists is that the peak search in IR libraries gives lesser results than full spectrum search: even such was our experience when the maximum common substructure (MCS) concept had been applied to hitlist structures [5]. It has to be mentioned, however, that these findings were produced with both tolerances–in absorbance, ΔA, and wavenumber, Δν–being small and thus suited for identity search. Furthermore, that was a result of numerous search sessions and intuitive assessment of search performance by a spectroscopist.

The found optimal tolerances for identity IR library search are in the range ΔA = 0.4 - 1.0 a.u. and Δν = 4 - 7 cm^{-1} [11] when SpSim_2 was used. For identity Raman library search these appeared to be ΔI = 0.2 - 1.0 arbitrary intensity units (a.i.u.) and Δν = 4 - 11 cm^{-1}. For both spectral methods as measure of the identity search effectivity the average hit position of 'the unknown' was used as optimum criterion: both tolerances were varied in the intervals 0.1 - 1.0 a.u./a.i.u. and Δν = 3 – 50 cm-1, respectively, and test search sessions with 50 spectra were used.

The proposed by Varmuza et al. method allows to quantitatively estimate similarity search results and find the optimal tolerance values. To find the optimal tolerances for similarity peak search (PS) each Raman spectrum of the spectral library was searched in the same library and the first hit (identical spectrum) was removed thus calculating the grand mean T(1) of Tanimoto indices upon 330 search sessions. Fig. 1 shows T(1) as a function of ΔI and Δν when SpSim_1 is used, and Fig. 2 when SpSim_2 is applied.

As can be seen from both figures, the highest T(1) values are around 0.60 which compared to the average best structural similarity of 0.80 is neither so low nor so high as one can wish. The average structural similarity of Raman library is 0.30 which is quite low than the highest T(1) values.

When the full spectrum search algorithm based on the correlation coefficient (CC) is used, the value of T(1) is nearly the same as the above mentioned one: the situation different from the corresponding comparison made for IR spectra where $T_{PS}(1) = 0.70 < T_{CC}(1) = 0.74$. It has to be noted here that the statistical distribution of these obviously random variables is unknown and thus an interval estimate could not be made but the curves 'h / T(h)' for IR spectra are quite distinctive and different from one another which is not the case for Raman spectra (see Fig. 3).

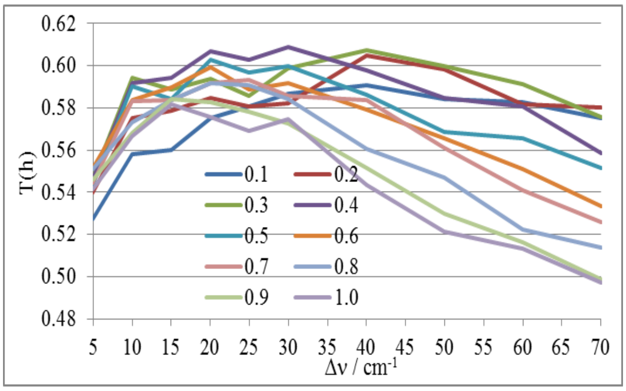

Figure 1. Grand mean of Tanimoto indices as a function of the wavenumber tolerance, Δν, varied by different values of the second tolerance, ΔI. The used spectral similarity measure is SpSim_1.

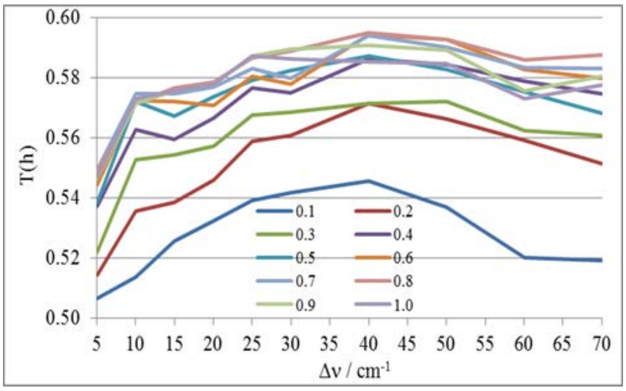

Figure 2. Grand mean of Tanimoto indices, T(h), as a function of the wavenumber tolerance, Δν, varied by different values of the second tolerance, ΔI. The used spectral similarity measure is SpSim_2.

The last could be attributed to higher variability of the intensity of the characteristic group vibrations in Raman spectra than in IR spectra [12]. As a whole, the discussed results proved that Raman spectra can be used for similarity search and one can derive some conclusions about the unknown structure by a surveying the hitlist structures in some way or the other.

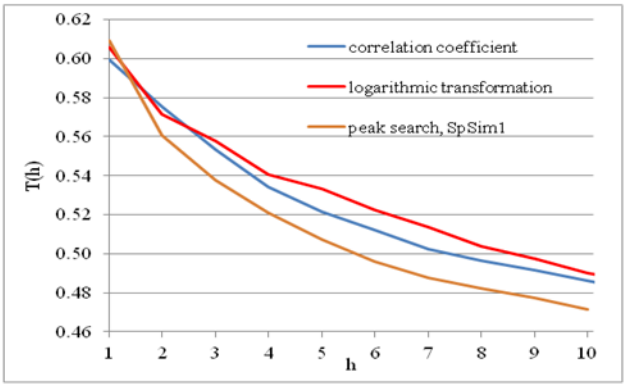

Figure 3. Grand mean of Tanimoto indices, T(h), as a function of the hit position, h, for search in the Raman library. The peak search is done with SpSim$_1$ and ΔI = 0.4 a.i.u. and Δν = 30 cm^{-1}. Logarithmic transformed spectra are compared by full spectrum search algorithm based on the correlation coefficient.

The found optimal values from Fig. 1 are ΔI = 0.3 - 0.4 a.i.u. and Δν = 20 - 40 cm^{-1} and from Fig. 2 ΔI = 0.7 - 0.9 a.i.u. and Δν = 30 - 55 cm^{-1}. As expected, the sensitivity (recall) is lost by small values of both tolerances and the selectivity is lost by their high values and obviously both tendencies lead to the T(1) decrease. The wider optimal tolerance intervals for the second spectral similarity measure are explained with its better selectivity which is achieved by the dot product (scalar product) of peak intensities in the numerator in Eqn. 2. The other facts that support the last assertion are the less steep of the curves in Fig. 2 when the wavenumber increases and that nearly all of the curves are close to each other. A detail view of both figures reveals that the curves are shifted upwards and downwards with the change of ΔI in a systematic way despite the small library size.

As mentioned above, for Raman spectra there is no difference between similarity search capabilities of the full spectrum search and the peak search. This can be clearly seen from Fig. 3 where an additional spectrum transformation–a logarithm of intensities–was tried to improve the similarity search. Since Raman intensities are saved as byte variable (from 0 to 255) in the libraries maintained by our software IRSS, it was hoped that the transformation given by Eqn. 3 will decrease the effect of variability of Raman band intensities, I_k. As seen, the gain in the similarity search capability is only for hits at positions 3 to 10 but the overall structural similarity of hitlist structures to that of the unknown is very important for such methods as MCS and k-nearest neighbors which use all hitlist structures.

$$I_k^{log} = \log_2 (I_k+1) / 8 \qquad (3)$$

The similarity search can be explored also with the distributions of structural similarity of three kinds of pairs as it is shown in Fig. 4.

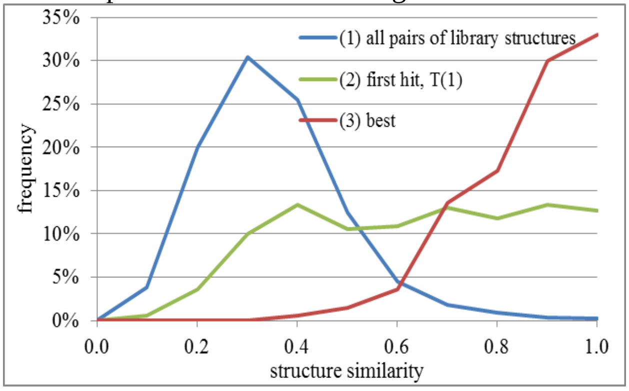

Figure 4. The distribution of structural similarity of three kinds of pairs (see text for explanation).

For (1) all pairs of library structures (altogether 330 x 329 /2 = 54,285 pairs), for (2) ‘unknown/first-hit’ pairs (330 T(1) values), and for (3) the ‘structure/its-most-similar-structure’ pairs (also 330 pairs). As can be seen, the distributions of the first two kinds of pairs are different and on average the library search does give a structurally similar first hit. Of course, there is a heavy overlap between them and the middle histogram is not closer the rightmost one which is a perfect case of T(1) distribution.

CONCLUSION

The performed study showed that the search in Raman spectral libraries gives hitlist compounds structurally similar to the unknown and the corresponding hit structures can be used in the structure elucidation process. The peak search performs equal with the full spectrum search algorithm based on the correlation coefficient due to higher variability of the intensity of the characteristic group vibrations in Raman spectra than in IR spectra as the high values of ΔI tolerance compensate the difference in band intensity in structurally similar compounds.

Acknowledgment *We thank Prof. Kurt Varmuza for providing the software SubMat and dedicate this paper to him on the occasion of his 75th anniversary (22 February 2017).*

REFERENCES

1. M. E. Elyashberg, A. Williams, K. Blinov, Contemporary Computer-Assisted Approaches to Molecular Structure Elucidation, Royal Society of Chemistry, London, 2012.
2. R. Freitag (ed.), Modern Advances in Chromatography. Springer-Verlag, Berlin, 2002.

3. E. Smith and G. Dent, Modern Raman Spectroscopy – A Practical Approach, John Wiley & Sons, New York, 2005.
4. H. J. Luinge, *Vib. Spectrosc.*, **1**, 3-18 (1990).
5. K. Varmuza, P. Penchev, H. Scsibrany, *J. Chem. Inf. Comput. Sci.*, **38** (3), 420-427 (1998).
6. K. Varmuza, M. Karlovits, W. Demuth, *Anal. Chim. Acta*, **490**, 313–324 (2003).
7. W. Demuth, M. Karlovits, K. Varmuza, *Anal. Chim. Acta*, **516**, 75-85 (2004).
8. K. Varmuza, P. Filzmoser, Introduction to Multivariate Statistical Analysis in Chemometrics, CRC Press, Boca Raton, 2009.
9. P. N. Penchev, N. T. Kochev and G. N. Andreev, *Compt. Rend. Acad. Bulg. Sci.*, **51**, 67-70 (1998).
10. P. N. Penchev, V. L. Miteva, A. N. Sohou, N. T. Kochev, G. N. Andreev, *Bulg. Chem. Commun.*, **40**, 556-560 (2008).
11. P. N. Penchev, Dr.Sc. Dissertation, University of Plovdiv, Plovdiv, 2016.
12. G. N. Andreev, private communication.

ВРЪЗКА МЕЖДУ СПЕКТРАЛНОТО ПОДОБИЕ И СТРУКТУРНОТО ПОДОБИЕ ЗА РАМАН СПЕКТРИ

П. Н. Пенчев*, С. Х. Цонева, С. Р. Начкова

Катедра „Аналитична химия и компютърна химия“,
Пловдивски Университет „Паисий Хилендарски”, ул. Цар Асен № 24, 4000 Пловдив,
e-mail: plamen@uni-plovdiv.net

Постъпила на 13 ноември 2016 г.; приета на 28 ноември 2016 г.

(Резюме)

Изследването на връзката между спектрално подобие и структурно подобие показва, че търсенето по подобие в Раман спектрални библиотеки е подходящо за процеса на разкриване структурата на неизвестни съединения. В резултат на изследването са намерени оптималните стойности на толерансите, използвани при пиково търсене и те са по-големи от тези, които са подходящи за търсене за идентификация.

Ключови думи: *Раман спектри, спектрално подобие, структурно подобие*