Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Lorant Bodis[1], Alfred Ross[2], Ernö Pretsch[1]

[1] Laboratory of Organic Chemistry, ETH Hönggerberg, CH-8093 Zürich, Switzerland
[2] Pharmaceuticals Division, F. Hoffmann-La Roche Ltd, CH-4070 Basel, Switzerland

Novel Similarity Measure for Comparing Spectra

Introduction

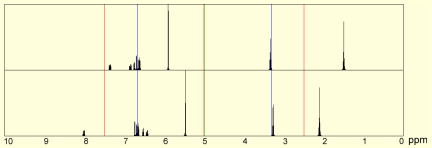
Most available vector comparison methods such as the correlation coefficient [1] and Tanimoto coefficient [2] are only able to find point-wise similarities. Similarity criteria for spectra comparison should include information about the neighborhood of the corresponding items in order to identify shifted signals as well. So far, only few such methods have been described. A recent method by de Gelder et al. [3] is based on a locally weighted cross-correlation function being normalized with the geometric mean of the individual autocorrelation functions. A much better performance has been achieved with our novel similarity criterion called bin method.

Bin Method

Similarity of two spectra *x* and *y*:

•The total integral of each individual spectrum is normalized to the number of H atoms in the corresponding molecule

•The spectra are successively divided into *n* bins ($n = 1,N$, N being the maximal number of bins):



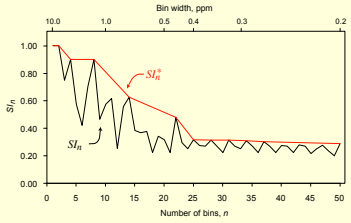
•For each division, the similarity index, SI_n, is calculated:

$$SI_n = \frac{I_{xy}(n)}{I_x + I_y - I_{xy}(n)} \qquad I_{xy}(n) = \sum_{i=1}^{n} \min\left(I_x(i), I_y(i)\right)$$

where I_x and I_y are the total integrals of the spectra *x* and *y*; $I_x(i)$ and $I_y(i)$ are the integrated intensities of the respective spectra within bin *i*

•**Similarity value**:

$$S = \frac{1}{N}\sum_{n=1}^{N} SI_n^*$$



Cross-correlation Method

Similarity of two functions $f(x)$ and $g(x)$ [3]:

$$S_{fg} = \frac{\int w(r)c_{fg}(r)\,dr}{\sqrt{\int w(r)c_{ff}(r)\,dr \int w(r)c_{gg}(r)\,dr}} \qquad c_{fg}(r) = \int f(x)g(x+r)dx$$

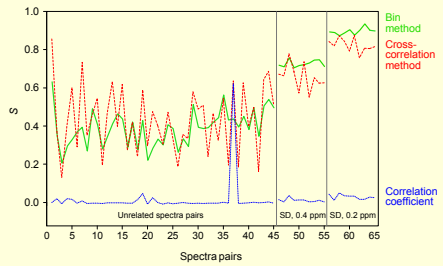
with $c_{fg}(r)$ as the cross-correlation function, $c_{ff}(r)$ and $c_{gg}(r)$ as the auto-correlation functions and $w(r)$ as the triangular weighting function

Tests with Artificial Spectra

•Ten arbitrarily chosen compounds and the corresponding predicted ^{1}H NMR spectra:

•Additionally, for each structure, two further spectra were calculated in which the multiplets were randomly shifted using a normal distribution with a standard deviation (SD) of 0.2 and 0.4 ppm.

•Performance of the bin method in comparison with other similarity criteria:

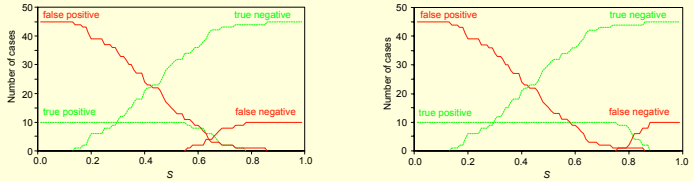


•The ten spectra are compared with those corresponding to other structures (entries 1–45) and with those having randomly shifted signal groups (entries 46–55: SD = 0.4 ppm and entries 56–65: SD = 0.2 ppm). The last two sets correspond to an average of the results obtained with 100 randomly shifted spectra.

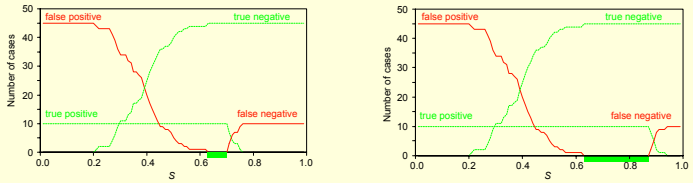
•Ideally, the comparison of spectra belonging to different structures should result in a low similarity, and of those with the randomly modified spectrum of the same structure, in a high one.

•Comparison with contingency diagrams: too low threshold values of *S* will consider incorrect pairs as correct ones, i.e., as false positives, while with too high threshold values of *S*, the number of false negatives will increase.

•**Cross-correlation method**: triangle weighting, 1.4 ppm cut-off range



•**Bin method**: minimal bin width of 0.4 ppm



•The ten true positive pairs result from comparing the original spectra with those having randomly shifted signal groups applying SD = 0.4 ppm (left) and SD = 0.2 ppm (right).

Tests with Measured Spectra

•1146 ^{1}H NMR spectra derived from a library of Chemical Concepts [4].

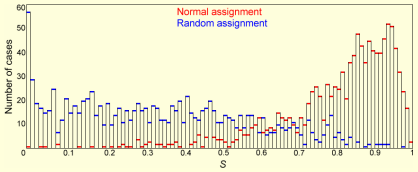
•Each measured spectrum was compared with two predicted spectra:

- •one on the basis of the correct structure (normal assignment) and
- •the other based on a randomly selected structure from the library (random assignment)

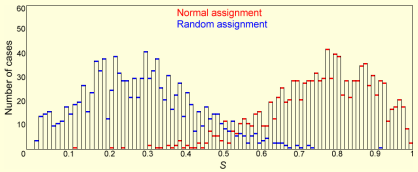
•Ideally, all normal comparisons should lead to a high, and the random ones to a low similarity value.

•Histogram of similarity values, *S*, of measured and calculated spectra using correct and random structure assignments:

•**Cross-correlation method**: triangle weighting, 1.4 ppm cut-off range; overlap: 306 (27%)



•**Bin method**: minimal bin width of 0.4 ppm; overlap: 138 (12%)



Conclusions

•Similarity of related ^{1}H NMR spectra has been successfully detected by a novel method based on dividing the spectra in bins.

•It has been shown that the correlation coefficient does not provide a useful similarity measure and that the recently introduced cross-correlation-based method performs less well than our novel similarity measure.

•Application of the new method with spectra of two or more dimensions including image analysis is straightforward.

References

[1] K. Varmuza, M. Karlovits, W. Demuth, *Anal. Chim. Acta* **2003**, *490*, 313.
[2] P. Willett, J. B. Barnard, G. M. Downs, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983.
[3] R. de Gelder, R. Wehrens, J. A. Hageman, *J. Comp. Chem.* **2001**, *22*, 273.
[4] Chemical Concepts GmbH, P.O. Box 100202, D-69442 Weinheim.