Available online at www.sciencedirect.com
ScienceDirect
Chemometrics and Intelligent Laboratory Systems 85 (2007) 1 – 8
Chemometrics and intelligent laboratory systems
www.elsevier.com/locate/chemolab

A novel spectra similarity measure

Lorant Bodis [a], Alfred Ross [b], Ernö Pretsch [a,*]

[a] *Laboratorium für Organische Chemie, ETH Hönggerberg, CH-8093 Zürich, Switzerland*
[b] *F. Hoffmann-La Roche Ltd., Pharmaceuticals Division, CH-4070 Basel, Switzerland*

Received 5 September 2005; received in revised form 29 September 2005; accepted 12 October 2005
Available online 21 November 2005

Abstract

A new method is described for calculating the similarity degree of two spectra. Its performance is optimized with similar, computer-generated ^{1}H NMR spectra. The method is compared with a recently proposed local cross-correlation method. Using a test set, its power to discriminate between related and unrelated ^{1}H NMR spectra is better than with the cross-correlation method. Better results are also obtained when comparing measured spectra of a database with the corresponding estimated ones or with estimated spectra of randomly assigned structures. Although, so far, it has only been tested with ^{1}H NMR spectra, due to the generality of the approach, the novel procedure can be applied to comparing other spectra or patterns as well.

Keywords: Spectra comparison; NMR spectra; Quality control

1. Introduction

Due to the recent advent of high throughput instruments, there is an increasing need for the automatic interpretation of molecular spectra. For example, it is possible today to automatically register ^{1}H NMR spectra of submicrogram samples in the order of minutes [1–3]. A key step in automatic interpretation is to establish the degree of similarity between the measured and a reference spectrum, which may originate from a database or computer prediction. Numerous measures have been proposed for describing the similarity of chemical structures [4] and some of them have been used to detect similarities in spectra. The similarity measures used so far with UV and IR spectra involve the correlation coefficient [5–7] or the dot product [7] of the two compared spectra, the Euclidean distance [6,7], and the match probability [8]. However, these approaches only work with spectra showing relatively broad signals since it fails to detect similarities if their positions in the two spectra differ by more than their widths. In other words, no information about the neighborhood of a signal is detected by these measures. Such similarity measures would not do when comparing NMR spectra because changes in signal positions of up to ca. 100 times the half widths are to be expected even in closely related spectra.

One possible solution to this problem is to artificially increase the line widths before further processing the NMR data. For example, Kalelkar et al. [9], by means of a moving average filter, reduce the number of data points from 16,384 to 820 (i.e., lowering the digital resolution from 0.46 to ca. 5 Hz). In another approach often applied in metabonomic studies, the NMR data is first compressed using a binning method, in which the normalized relative integrals are taken in each segment having a typical width of 0.04 ppm [10,11]. Then, the principal component or partial least squares analysis is used as a second compression step [11]. However, since the goal of the present study is to pair-wise compare spectra and not to find similarities or dissimilarities in a series of spectra, these methods are not applicable to spectra interpretation.

Also in the case of X-ray powder spectra, the relative line widths are much smaller than the tolerable differences in their position. Not surprisingly, this is the field where various similarity measures have been developed, which are able to cope with differences in signal positions that are much larger than the line widths. The first method in this direction,

* Corresponding author.
E-mail address: pretsch@org.chem.ethz.ch (E. Pretsch).

0169-7439/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2005.10.002

proposed by Karfunkel et al. [12], calculates a weighted cross product of the spectra with a weighting matrix having 1 as diagonal elements and values continuously decreasing with increasing distance from the diagonal (cf. [13]).

More recently, de Gelder et al. [14] have shown that various similarity criteria of two functions, $f(x)$ and $g(x)$, including the sum of squared differences [15], the correlation coefficient, and the overlap integral [16], are related to the cross-correlation function, $c_{fg}(r)$ at $r=0$:

$$c_{fg}(r) = \int f(x)g(x+r)dx. \quad (1)$$

Thus, they cannot provide any information about patterns that are shifted relative to each other. While the integral over r of the cross-correlation function is always equal to the product of the integrated intensities of the two spectra, i.e., in itself is not a similarity measure, its shape contains the information on the degree of similarity. The authors proposed a generalized expression for similarity, S_{fg}, which is based on a weighted cross-correlation function (weighting function, $w(r)$) normalized with the product of the two weighted autocorrelation functions, $c_{ff}(r)$ and $c_{gg}(r)$ defined in analogy to Eq. (1):

$$S_{fg} = \frac{\int w(r)c_{fg}(r)dr}{\sqrt{\int w(r)c_{ff}(r)dr \int w(r)c_{gg}(r)dr}}. \quad (2)$$

In their studies, de Gelder et al. used a triangular weighting function of width l defined as $w(r)=1-|r|/l$ if $|r|<l$, and $w(r)=0$ if $|r|\geq l$.

In this contribution, we introduce a novel method to quantify spectra similarity and compare its performance with the generalized similarity measure of de Gelder et al. [14].

2. Experimental

The algorithms (input/output modules, similarity criteria) were implemented in Borland® Delphi™ 5.0 [17]. The tests were performed on a Windows® PC with Intel® Pentium® 4 2.8 GHz CPU and 512 MB RAM. For estimating the spectra, the program NMRPrediction 3.0 [18] was used.

The first test set consisted of estimated ^{1}H NMR spectra of ten arbitrarily chosen compounds (Fig. 1). For each spectrum, two other spectra were calculated by randomly shifting signal groups using a normal distribution with a standard deviation (SD) of 0.2 or 0.4 ppm (an example is shown in Fig. 3).

The second test set consisted of 1146 ^{1}H NMR spectra derived from a library of Chemical Concepts [19]. From the 5003 compounds, those database entries were selected for which NMRPrediction 3.0 is capable of predicting all chemical shifts with optimal accuracy. Since this is not the case for –OH and –NH protons, the corresponding entries were omitted. Spectra recorded in D_2O or CD_3CN were also excluded. Additionally, 108 spectra containing obvious errors were removed. For spectra recorded in dimethyl sulfoxide-d_6, the solvent signals (including that of water) were eliminated using an automatic procedure. The noise and negative intensity values were removed by first analyzing the standard deviation of the noise in the signal-free region at both ends of the spectra and then zeroing all data points that were smaller than three times the standard deviation of the noise. Finally, the integrated intensities were normalized to the total number of protons.

The ^{1}H NMR spectra applied in these studies have 8–32 K data points corresponding to a range of 10–20 ppm. They were successively divided into $n=1$, N bins, with N being up to 25. Since the number of data points was, in general, not an exact multiple of the number of bins, after the division there usually was a remainder of <50 points, which were included in the neighboring (last) bin. As the finest division corresponded to 0.4 ppm or >150 data points, the reminder corresponded to a spectral range <0.1 ppm on the right side of the spectrum.

Computing times for comparing one pair of spectra, including the spectra prediction and elimination of noise and solvent signals, were of the order of 0.5 s.

3. Results and discussion

3.1. Introduction

In order to check the compatibility of a ^{1}H NMR spectrum with a proposed structure, the spectrum is estimated with the computer program NMRPrediction [18] and compared with the measured one. Since no exact match of the signals in the two spectra can be expected, the spectral comparison method must recognize the similarity of patterns having slightly shifted signals. The order of magnitude of the expected shifts can be estimated from the mean deviation of the predicted and measured chemical shifts, which is in the order of 0.2 ppm [20,21]. Similar variations in the signal positions are also expected between spectra measured in different solvents, e.g., in $CDCl_3$ and dimethyl sulfoxide-d_6 [22].

Preliminary comparisons of measured and estimated spectra were performed with the cross-correlation method by de Gelder et al. [14] using triangle and rectangle weighting functions $w(r)$. Surprisingly, in a series of cases, the rectangle but not the triangle as weighting function gave similarity values, $S_{fg}>1$ (cf. Eq. (2)). Indeed, there is no mathematical reason why the weighted integral of the cross-correlation function should always be smaller or equal than the geometric mean of the corresponding weighted integrals of the autocorrelation functions. This is illustrated by the two simple vectors in Fig. 2. As can easily be verified, a rectangle of width 4 as weighting function leads to a similarity value of $S_{fg}=1.00223$. Another drawback of the weighted cross-correlation method is the insufficient discrimination of spectra assigned to incorrect structures (see below). For these reasons, different other similarity measures were tested. They included artificial line broadening of up to 20 Hz [9] before calculating the correlation coefficient or using the weighted cross-correlation method and other types of normalization within the weighted cross-correlation method. The so far best-performing one is described in the following section.

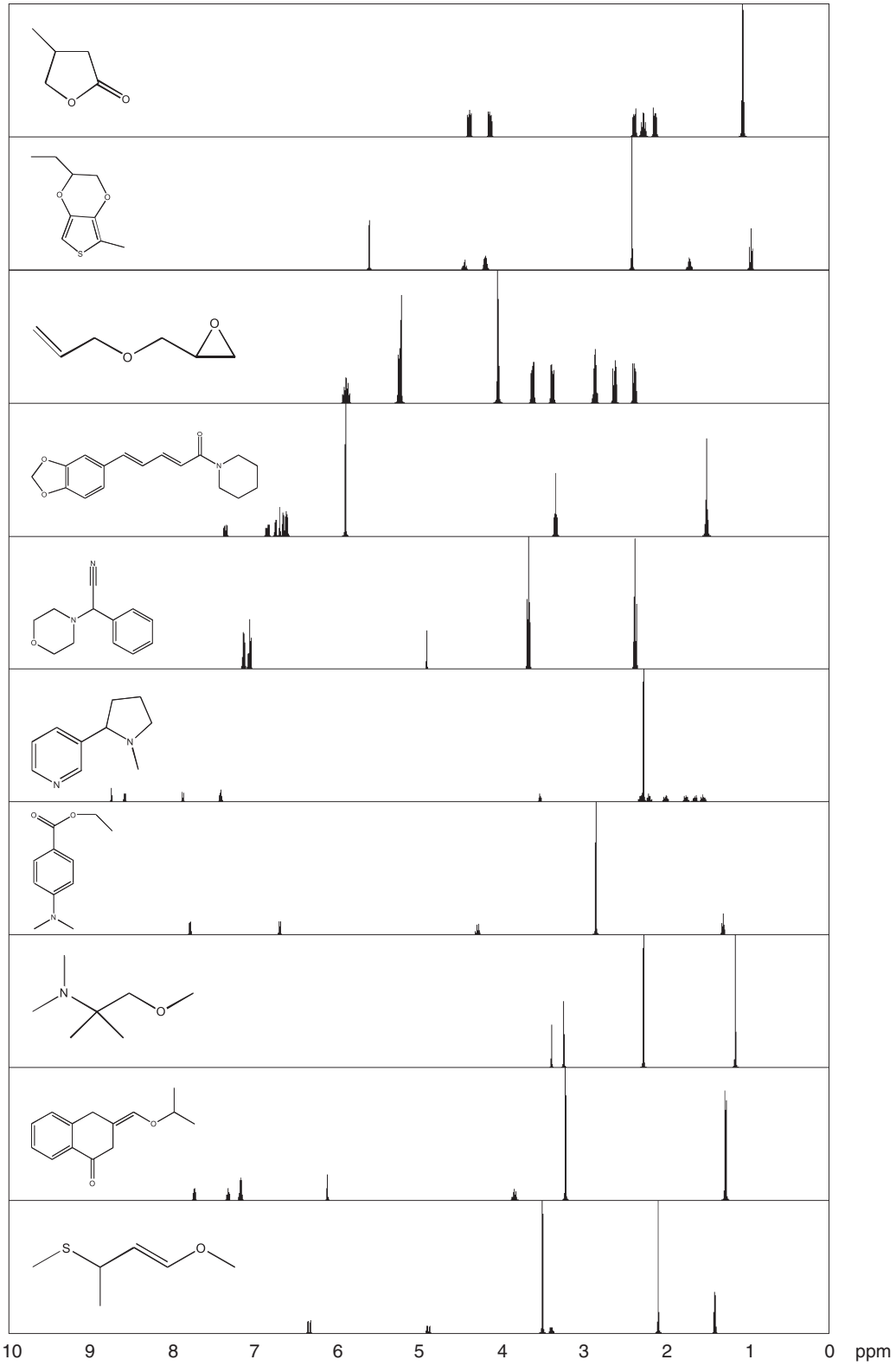

Fig. 1. Computer-generated ^{1}H NMR spectra of ten arbitrarily chosen structures of organic compounds used as a test set (16 K digital points, 10 ppm range, 500 MHz).

3.2. Novel similarity criterion

The novel similarity criterion of two spectra x and y is related to the binning method applied in metabonomic studies [10,11,23]. First, the total integral of each individual spectrum is normalized to the same value. In the case of ^{1}H NMR spectra, it is the most natural to normalize with respect to the number of H atoms in the corresponding molecule. In general,

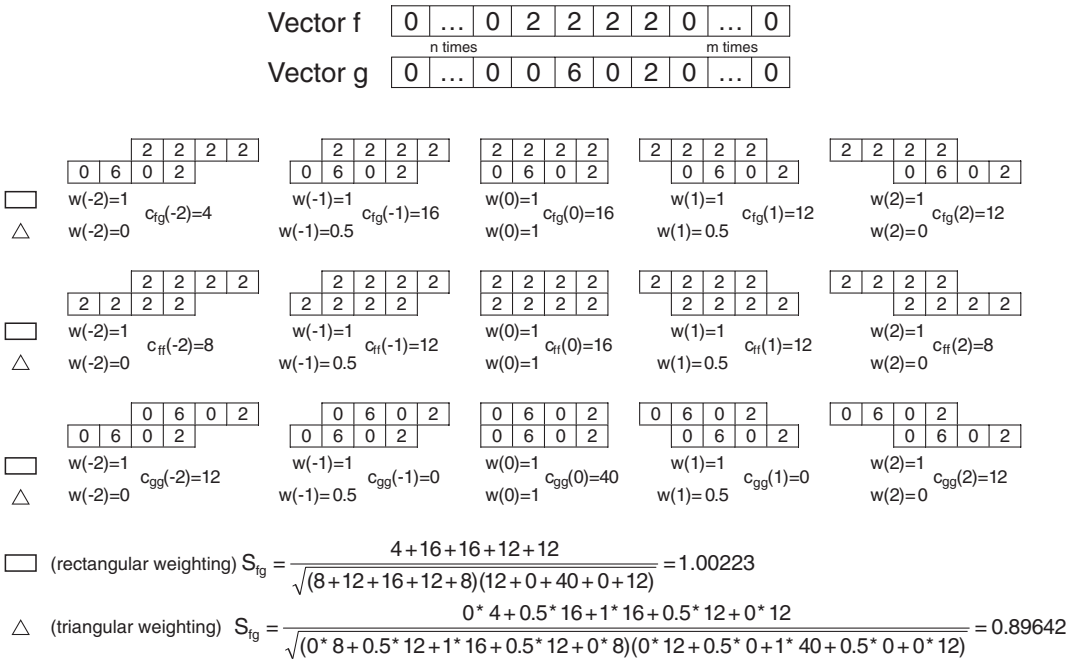

Fig. 2. Two simple vectors whose comparison by the cross-correlation method [14] with a rectangular weighting function and a cut-off range of 4 leads to a similarity value >1. With the triangular weighting function, the similarity value is 0.896.

any other value can be used as long as it is the same for the two patterns to be compared. Then, the spectra are successively divided into n bins with $n=1, N$, where N corresponds to the maximal number of bins. For each division, the similarity index, SI_n, is calculated according to Eq. (3):

$$SI_n = \frac{I_{xy}(n)}{I_x + I_y - I_{xy}(n)} \tag{3}$$

I_x and I_y being the total integrals (i.e., the number of H atoms) of the spectra x and y and:

$$I_{xy}(n) = \sum_{i=1}^{n} \min\left(I_x(i), I_y(i)\right) \tag{4}$$

with $I_x(i)$ and $I_y(i)$ as the integrated intensity of the respective spectra within bin i. The procedure is illustrated in Fig. 3 by two estimated spectra of piperine (cf. Fig. 1, entry no. 4) divided into 2 (central dashed line), 3 (dotted lines), and 4 (three dashed lines) bins. Since the molecule has 19 H atoms, $I_x=I_y=19$. For the cases shown in Fig. 3, $I_{xy}(2)=19.000$, $I_{xy}(3)=16.258$, and $I_{xy}(4)=17.992$, from which $SI_2=1.000$, $SI_3=0.747$, and $SI_4=0.899$, respectively. Using the above example, Fig. 4 gives the SI_n values for $n=1$ to 50 bins. Apparently, it may happen that a finer division (i.e., more bins) provides a higher value of the similarity index than a coarse one. The non-monotonous changes in the SI_n values (thin line in Fig. 4) occur if signals close to each other are partitioned into different bins for a given value of n, but belong into the same bin again if the number of bins is increased (see Fig. 3). To reduce the influence of such artifacts, the overall similarity, S, is defined according to Eq. (5) as the normalized

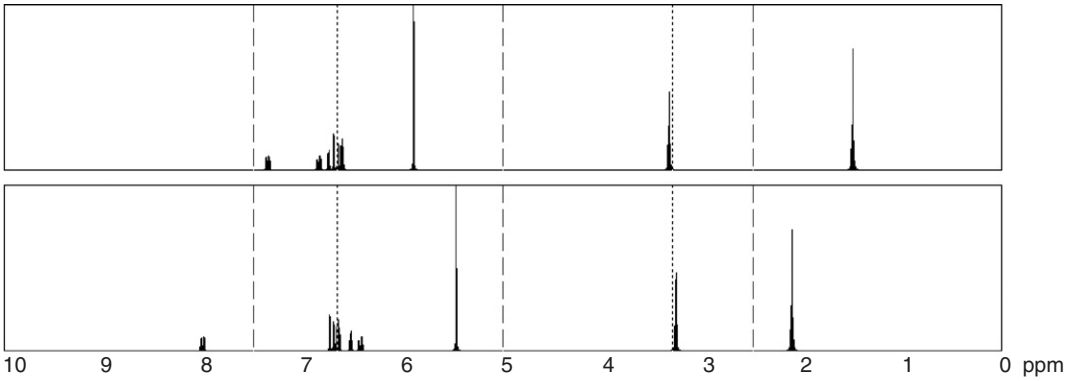

Fig. 3. Predicted spectrum (top) of piperine (cf. Fig. 1, entry no. 4) and the corresponding spectrum (bottom) with randomly shifted multiplets (SD=0.4 ppm) divided into 2 (central dashed line), 3 (dotted lines), and 4 (three dashed lines) bins. It can be observed that a finer division (i.e., more bins) may yield a higher similarity index than a coarse one.

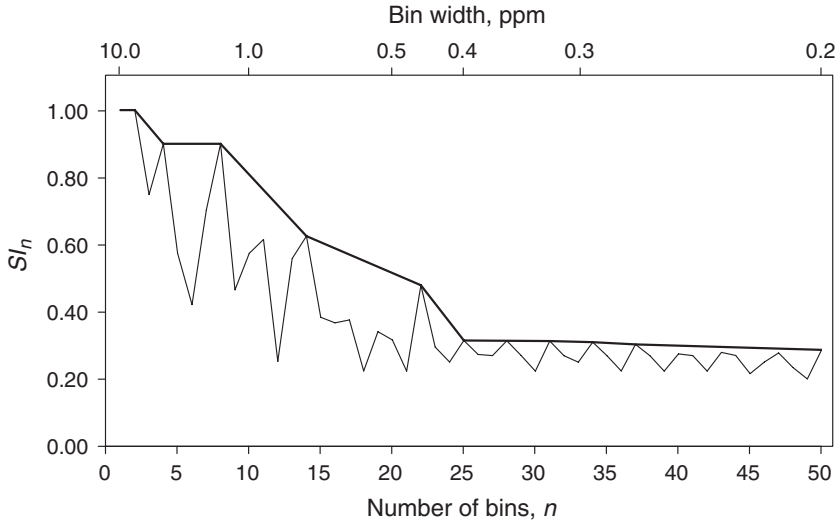

Fig. 4. Thin line: Similarity index (SI_n) calculated by Eq. (3) using divisions of up to 50 bins (minimal bin width, 0.2 ppm) for comparing the predicted spectrum of piperine (cf. Fig. 3, top) with the corresponding one having randomly shifted signal groups (normal distribution, SD=0.4 ppm). Thick line: The negative effect of the oscillating values of the similarity index is compensated by connecting the remaining maxima of SI_n.

integral of the function, SI_n^*, connecting the remaining global maxima (thick line in Fig. 4) rather than the average of the SI_n values:

$$S = \frac{1}{N}\sum_{n=1}^{N} SI_n^* \tag{5}$$

where

$$SI_n^* = \max\left(SI_n, \frac{SI_a(n-b) - SI_b(n-a)}{a-b}\right) \tag{6}$$

with

$$SI_a = SI_{n-1}^* \text{ and } SI_1^* = 1 \tag{7}$$

$$SI_b = \left\{\max\ SI_i | i = \overline{n, N}\right\}. \tag{8}$$

According to Eqs. (3) and (4), all values of SI_n and, therefore, also of S, can only lie between 0 and 1. The definition of SI_n is related to the Tanimoto coefficient [4] and does not award similarities due to the absence of signals (as, e.g., does the correlation coefficient). The only parameter to select in order to calculate S is the maximal number, N, of bins. For a given spectral width, this defines the highest division of the spectral range. As shown in Fig. 4, a too fine division, i.e., one that leads to windows smaller than the expected differences in signal positions (SD=0.4 ppm in the case shown), yields small SI_n values and, thus, decreases the overall similarity, S. Based on different tests with the databases used (see below), a minimal bin width of 0.4 ppm has proven to be optimal. This nicely fits the expectation of tolerable deviations of ± 0.2 ppm in the signal positions. In general, the maximum number of bins or the bin width is defined by the tolerable differences between signal positions. In the present case of applying the method to ^{1}H NMR spectra, the limit used here has the drawback that similarities or differences in the fine structure of the spectra do not influence the result.

3.3. Tests with artificial spectra

The performance of the two similarity criteria was first tested with the ten compounds of Fig. 1 whose spectra were predicted with the computer program NMRPrediction 3.0 [18]. Additionally, for each structure, two further spectra were calculated in which the multiplets were randomly shifted using a normal distribution with SD=0.2 or 0.4 ppm, one example with SD=0.4 ppm being shown in Fig. 3 (bottom). The spectral similarities calculated by the correlation coefficient, the cross-correlation method [14], and the new bin method proposed here are shown in Fig. 5 as dotted, dashed, and solid lines, respectively. The entries 1–45 correspond to the comparison of the calculated spectra of two different compounds. The next ten entries (46–55) are the results obtained by comparing the estimated spectrum of a compound with the corresponding one having signal groups randomly shifted by SD=0.4 ppm. Analogously, for the last ten entries (56–65) the random shifts correspond to SD=0.2 ppm. In order to eliminate fluctuations due to chance correlations, each entry of the last two sets was calculated 100 times and the mean of the similarities is shown. The type and width of the weighting function, $w(r)$, of the cross-correlation method and the minimal bin width for the bin method have been systematically varied. The parameters corresponding to the best discrimination between the first 45 and the last 20 comparisons are shown (triangle weighting function of width l=1.4 ppm for the cross-correlation and minimal bin width of 0.4 ppm for the bin method). It should be noted that the comparison is rugged with both methods, i.e., a slight variation of these parameters has only a minimal influence on the results.

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. As shown in Fig. 5, the correlation coefficient is not an adequate measure but the other two methods discussed here

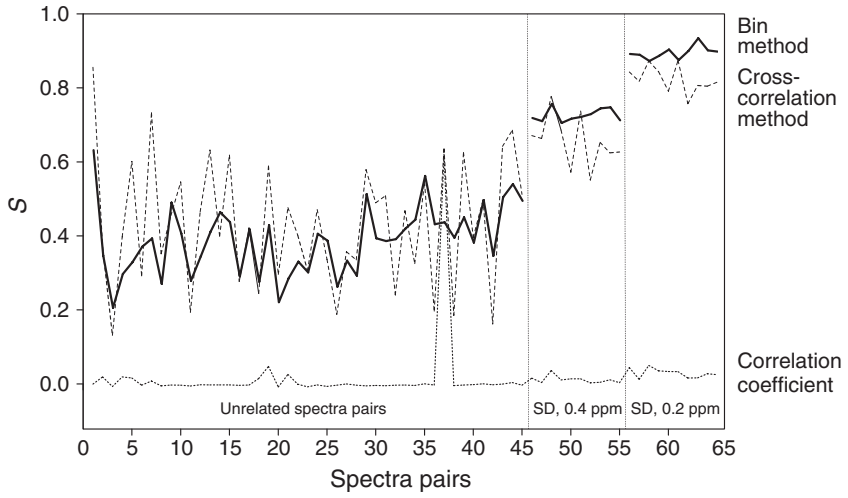

Fig. 5. Similarity achieved by the three measures investigated here: correlation coefficient (dotted line), cross-correlation method (dashed line), and bin method (thick line). The ten artificial spectra of Fig. 1 in different combinations. The ten spectra (Fig. 1) are compared with those corresponding 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.

fulfill the stated requirement. The overall discrimination is, however, better with the bin method. For example, for the spectra with SD=0.4 ppm (spectra pairs 46–55), even with a threshold of $S=0.7$, the cross-correlation method still gives two false positive entries but only two out of ten as true positives, whereas the bin method correctly assigns all structures.

A more detailed comparison is possible on the basis of the contingency diagrams shown in Fig. 6. If too low threshold values are assumed for S, a number of incorrect pairs will be considered as correct ones, i.e., as false positives. On the other hand, with too high threshold values of S, the number of false negatives increases. Ideally, there should exist a range in which the number of both is 0, i.e., the number of true positives and true negatives (dashed lines) are maximal. This is, indeed, possible using the bin method for both sets of spectra with shifted signal groups (SD=0.2 or 0.4 ppm). On the other hand, there is no similarity threshold that would fulfill this criterion with the cross-correlation method.

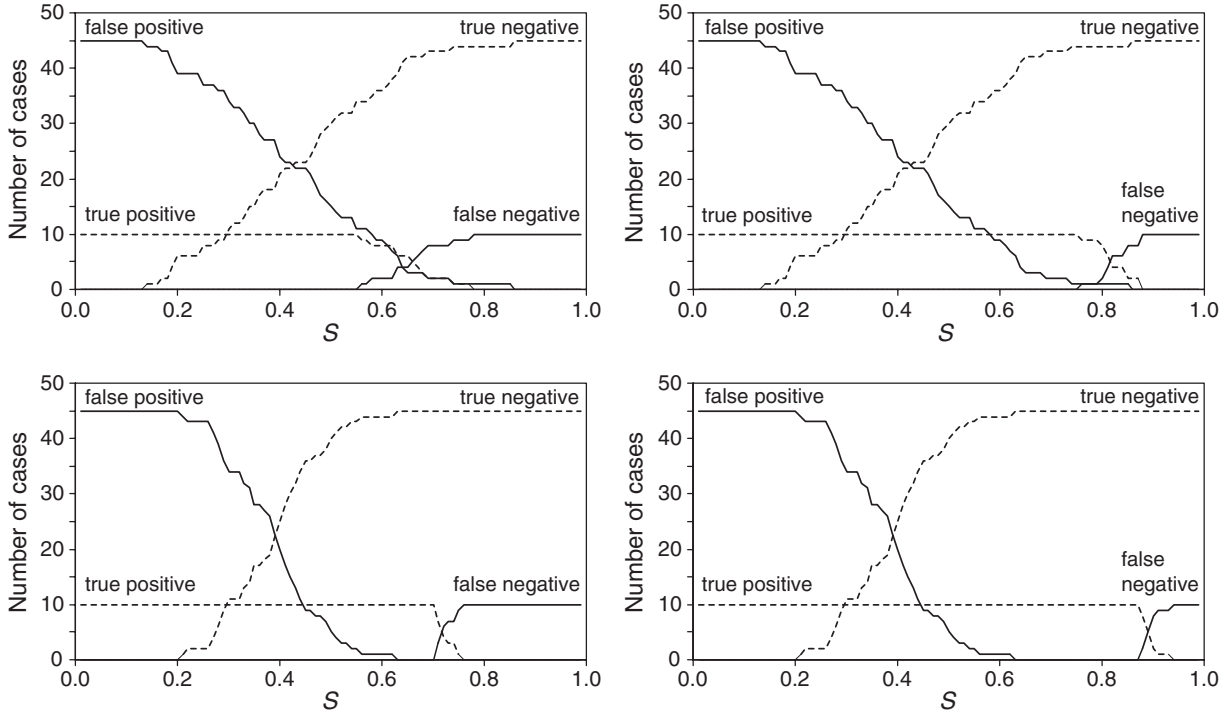

Fig. 6. Entries of the contingency tables as a function of the threshold value of the similarity, S. Top: weighted cross-correlation method (triangle weighting, 1.4 ppm cut-off range). Bottom: bin method (minimal bin width, 0.4 ppm). The ten true positive pairs result from comparing the original spectra with those having randomly shifted signal groups, corresponding to SD=0.4 ppm (left) and SD=0.2 ppm (right).

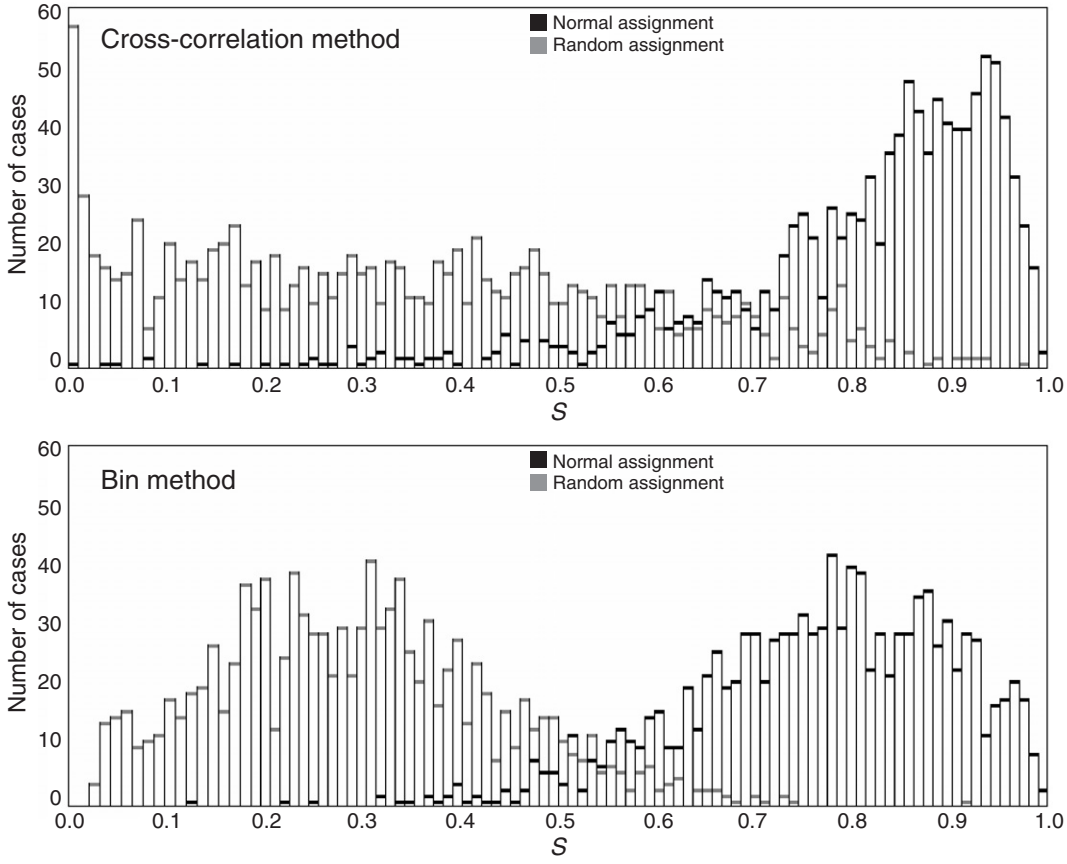

Fig. 7. Histogram of similarity values, S, of measured and calculated ^{1}H NMR spectra using correct and random structure assignments. Top: weighted cross-correlation method (triangle weighting, 1.4 ppm cut-off range). Bottom: bin method (minimal bin width, 0.4 ppm).

3.4. Tests with measured spectra

Further tests were conducted with 1146 entries of a ^{1}H NMR spectral library (see Experimental). 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. As indicated in Fig. 7, the similarity measure has an influence on the results. With the procedure by de Gelder et al. [14], using the same parameters as above, a stronger overlap of the two distributions (306 spectra pairs or 26.7% of the test cases) was found than with the bin method introduced here (138 spectra pairs or 12.0%). This example shows that our novel method is of advantage also when comparing measured and predicted spectra.

4. Summary

The similarity of related ^{1}H NMR spectra has been successfully detected by a novel method based on dividing the spectra in $n=1$ to N bins (with N being the maximal number of bins) and calculating the integrated signal intensities within each bin. It is shown that the correlation coefficient does not provide a useful similarity measure and that the recently introduced cross-correlation-based method performs somewhat less well than our novel similarity measure. Although, so far, it has only been tested with one-dimensional ^{1}H NMR spectra, the application of the new method with spectra of two or more dimensions including image analysis is straightforward.

Acknowledgements

This research was financially supported by a grant from F. Hoffmann-La Roche Ltd. We thank Dr. R. Neudert for the ^{1}H NMR database and Dr. D. Wegmann for critical reading of the manuscript.

References

[1] A. Ross, G. Schlotterbeck, H. Senn, M. von Kienlin, Angew. Chem., Int. Ed. Engl. 40 (2001) 3243–3245.
[2] M. Macnaughtan, T. Hou, J. Xu, D. Raftery, Anal. Chem. 75 (2003) 5116–5123.
[3] D. Raftery, Anal. Bioanal. Chem. 378 (2004) 1403–1404.
[4] P. Willett, J.B. Barnard, G.M. Downs, J. Chem. Inf. Comput. Sci. 38 (1998) 983–996.
[5] H.-J.P. Sievert, A.C.J.H. Drouen, in: L. Huber, S.A. George (Eds.), Diode Array Detection in HPLC, Marcel Dekker, Inc., New York, 1993, pp. 51–126.
[6] K. Baumann, J.T. Clerc, Anal. Chim. Acta 348 (1997) 327–343.
[7] K. Varmuza, M. Karlovits, W. Demuth, Anal. Chim. Acta 490 (2003) 313–324.
[8] S.L.R. Ellison, S.L. Gregory, Anal. Chim. Acta 370 (1998) 181–190.

[9] S. Kalelkar, E.R. Dow, J. Grimes, M. Clapham, H. Hu, J. Com. Chem. 4 (2002) 622–629.
[10] M.E. Bollard, E.G. Stanley, J.C. Lindon, J.K. Nicholson, E. Holmes, NMR Biomed. 18 (2005) 143–162.
[11] E. Holmes, J.K. Nicholson, A.W. Nicholls, J.C. Lindon, S.C. Connor, S. Polley, J. Connelly, Chemom. Intell. Lab. Syst., Lab. Inf. Manag. 44 (1998) 245–255.
[12] H.R. Karfunkel, B. Rohde, F.J.J. Leusen, R.J. Gdanitz, G. Rihs, J. Comput. Chem. 14 (1993) 1125–1135.
[13] D.S. Stephenson, G. Binsch, J. Magn. Reson. 37 (1980) 395–407.
[14] R. de Gelder, R. Wehrens, J.A. Hageman, J. Comput. Chem. 22 (2001) 273–289.
[15] J. Dods, D. Gruner, P. Brumer, Chem. Phys. Lett. 261 (1996) 612–619.
[16] S.L. Lawton, L.S. Bartell, Powder Diffr. 9 (1994) 124–135.
[17] http://www.borland.com/us/products/delphi/index.html.
[18] Porta Nova Software GmbH, CH-8037 Zürich, Switzerland.
[19] Chemical Concepts GmbH, P.O. Box 100202, D-69442 Weinheim.
[20] R. Bürgin Schaller, E. Pretsch, Anal. Chim. Acta 290 (1994) 295–302.
[21] R. Bürgin Schaller, M.E. Munk, E. Pretsch, J. Chem. Inf. Comput. Sci. 36 (1996) 239–243.
[22] E. Pretsch, P. Bühlmann, C. Affolter, Structure Determination of Organic Compounds, 3rd edition, Springer-Verlag, Berlin, 2000.
[23] J.C. Lindon, E. Holmes, J.K. Nicholson, Anal. Chem. 75 (2003) 385A–391A.