



## Automatic compatibility tests of HSQC NMR spectra with proposed structures of chemical compounds

Lorant Bodis<sup>a,b</sup>, Alfred Ross<sup>c</sup>, Jenö Bodis<sup>d</sup>, Ernő Pretsch<sup>a,\*</sup>

<sup>a</sup> Institute of Biogeochemistry and Pollutant Dynamics, ETH Zurich, CH-8092 Zurich, Switzerland

<sup>b</sup> Cambridge Technology Partners, Leutschenbachstr. 41, CH-8050 Zurich, Switzerland

<sup>c</sup> F. Hoffmann-La Roche Ltd., Pharmaceuticals Division, CH-4070 Basel, Switzerland

<sup>d</sup> Department of Organic Chemistry, Babes-Bolyai University, str. Arany János 11, 400028 Cluj, Romania

### ARTICLE INFO

#### Article history:

Received 12 January 2009

Received in revised form 25 May 2009

Accepted 2 June 2009

Available online 12 June 2009

#### Keywords:

Spectra comparison

HSQC spectra

Automatic NMR spectra interpretation

Quality control

### ABSTRACT

A recently introduced similarity measure is extended here for comparing two-dimensional spectra. Its applicability is demonstrated with heteronuclear single-quantum correlation (HSQC) NMR spectra. For testing the compatibility of a spectrum with the proposed chemical structure, first, the spectrum is predicted on the basis of that structure and then, the proposed comparison algorithm is applied. In this context, the topics of optimization are peak picking, signal intensity measures, and optimizing the parameters of the two-dimensional comparison method. The performance is analyzed with a test set of 289 structures of organic compounds and their HSQC and <sup>1</sup>H NMR spectra. The results obtained with HSQC spectra are better than those achieved using the previously described one-dimensional similarity test with <sup>1</sup>H NMR spectra alone.

© 2009 Elsevier B.V. All rights reserved.

### 1. Introduction

The bottleneck of obtaining useful NMR spectroscopic information has successively shifted from registering to interpreting spectra of organic compounds. Consequently, there is an increasing need for computer programs performing automatic interpretation [1]. In many applications, NMR spectra are taken in order to confirm an expected structure of a compound. Thus, the confirmation or rejection of the expected structure is sufficient and no full interpretation of the spectra is required. Various approaches have been proposed on these lines. In a series of papers, Griffiths, first, automatically extracted the relevant parameters (i.e., chemical shifts [2], coupling constants [2], and integrals [3]) from <sup>1</sup>H NMR spectra. Then, the parameters were automatically compared with those predicted from the proposed structure [4]. The comparison was based on a full assignment of predicted and measured values and the calculation of the so-called mismatch matrix between the two sets. The test set consisted of 21 pairs of measured and predicted <sup>1</sup>H NMR spectra, for the latter using partly the correct and partly a similar structure. For this set, the mismatch methodology allowed unrelated compounds to be differentiated with very high levels of reliability (>99%) [4]. Later, the automatic evaluation of parameters was extended to flow-NMR spectra [5]. Corresponding compati-

bility tests with 27 samples (each one compared with all others) yielded an error of 10–15% (sum of false negatives and false positives) [6]. Another study was also based on the automatic evaluation of the above-mentioned three kinds of parameters, i.e., chemical shifts, multiplicities, and intensities, but used their weighted sum as similarity criterion [7]. Several test sets (with individually optimized parameters in each case) led to the conclusion “that a system is possible whereby 69% of the spectra are prepared and evaluated automatically, and never need to be seen or evaluated by a human” [7]. The reliability of the automatic comparison was increased by additionally using <sup>13</sup>C NMR chemical shifts [8,9]. The improvement was not only due to the additional parameters but also due to the increase in the reliability of the <sup>1</sup>H shift assignments obtained by 2D HSQC spectral information [10].

Especially with second-order spectra and overlapping signals, the automatic assignment of chemical shifts and multiplicities, obviously, is an error-prone task. Therefore, spectra comparison without previous assignment might be an advantage. By using the binning method, i.e., by evaluating the integrated intensities in fixed-length intervals [11], NMR spectra can easily be converted into vectors. Conventional comparison methods of vectors, such as distance of angle measures (correlation coefficient), do not adequately reflect spectral similarity because there is no information about the neighborhood of the individual bins, i.e., between small and large deviations of estimated and measured chemical shifts. Spectral similarity is, however, well reflected by several recently introduced measures [12–14]. The similarity of related <sup>1</sup>H NMR

\* Corresponding author.

E-mail address: [pretsche@ethz.ch](mailto:pretsche@ethz.ch) (E. Pretsch).

spectra was successfully detected by dividing them in varying numbers of bins and, for each division, calculating the signal intensities within each bin (for details, see below) [14]. For a test set of 1146  $^1\text{H}$  NMR spectra, each measured spectrum was compared with two predicted ones, one on the basis of the correct structure (normal assignment) and the other based on a randomly selected structure from the library (random assignment). The overlap between the distribution of correct and incorrect structures was 12%.

In this work, we present a generalization of the previously introduced method for 2D NMR spectra and implement it for comparing HSQC spectra. For the practical application, the peak-picking procedure had to be optimized. Peak intensities in HSQC spectra depend on the value of the C–H coupling constants. Since the similarity measure is based on peak intensities, which in HSQC spectra are biased, various intensity measures have been tested.

## 2. Theory

For the one-dimensional bin method introduced earlier, the spectra are divided in ranges (bins) of variable length, and the integrated intensities of the individual ranges are calculated [14]. In the present work, two-dimensional spectra are successively divided into  $n^2$  rectangle bins ( $n$  divisions along rows and columns) with  $n = 1, \bar{N}$ ,  $\bar{N}$  corresponding to the maximal number of divisions in both directions ( $^1\text{H}$  NMR and  $^{13}\text{C}$  NMR), which is obtained through dividing the spectral ranges by the minimal bin widths. For a total spectral range of 190 ppm, minimal bin widths of 10, 5, and 2.5 ppm were tested for  $^{13}\text{C}$  NMR spectra. In the case of  $^1\text{H}$  NMR, the total range was 12 ppm and the corresponding minimal bin widths were 0.64, 0.32, and 0.16 ppm. First, the total integral of the spectrum is normalized to the total number of protons. Then, for each division, the similarity index,  $SI_n$ , between the two spectra  $x$  and  $y$  is calculated [14] as:

$$SI_n = \frac{I_{xy}(n)}{I_x + I_y - I_{xy}(n)} \quad (1)$$

$I_x$  and  $I_y$  being the total integrals of the spectra  $x$  and  $y$ , respectively, and

$$I_{xy}(n) = \sum_{i=1}^n \min(I_x(i), I_y(i)) \quad (2)$$

with  $I_x(i)$  and  $I_y(i)$  as the integrated intensities for the respective spectra within bin  $i$ .

The overall similarity,  $S$ , is defined as the normalized integral of the function,  $SI_n^*$ , which is obtained by connecting the global maxima of the  $SI_n$  values (cf. [14]):

$$S = \frac{1}{\bar{N}} \sum_{n=1}^{\bar{N}} SI_n^* \quad (3)$$

where:

$$SI_n^* = \max \left( SI_n, \frac{SI_a(n-b) - SI_b(n-a)}{a-b} \right) \quad (4)$$

with

$$SI_a = SI_{n-1}^* \quad \text{and} \quad SI_1^* = 1 \quad (5)$$

$$SI_b = \{ \max SI_i \mid i = \bar{n}, \bar{N} \} \quad (6)$$

As shown in Fig. 4 of ref. [14], these equations interpolate between the local maxima;  $a$  and  $b$  correspond to the running index number of the previous and the next maximum, respectively. The method is demonstrated further below using a practical example.

## 3. Experimental

The algorithms (input/output modules, similarity criteria) were implemented in Borland® Delphi™ 5.0 [15]. The tests were performed on a Windows® PC with Intel® Pentium® 4 2.8 GHz CPU and 512 MB RAM. To estimate the HSQC NMR spectra, the program NMR-Prediction 3.0 program [16–18] was used for predicting the  $^1\text{H}$  and  $^{13}\text{C}$  chemical shifts.

The test set consisted of 289 organic compounds having an average molecular mass of 300 Da (range, 80–1200 Da) and their HSQC and  $^1\text{H}$  NMR spectra. The HSQC spectra were recorded in the range of –5 to 185 ppm with 512 digital points in the  $^{13}\text{C}$  NMR direction, and in the range of –0.5 to 11.5 ppm with 2048 digital points in the  $^1\text{H}$  NMR direction. Note that the library consisted of routine spectra without manual adjustments, i.e., impurities and spectral imperfections contributed to the errors. The signals of the not fully deuterated dimethyl sulfoxide (DMSO- $d_5$ ) and  $\text{CHCl}_3$  were automatically removed prior to comparison. The  $^1\text{H}$  NMR spectra were recorded in the range of –1 to 21 ppm with 32 K digital points. During the tests, the original full sizes of the spectra were used.

Each measured HSQC spectrum was compared with two predicted ones: one on the basis of the correct structure (normal assignment) and the other based on a randomly selected structure from the library (random assignment). Note that the same random assignment was used for each experiment. Ideally, all normal comparisons should lead to a high and the random ones to a low similarity value. The similarities calculated for each spectra pair, both for normal and random assignments, were analyzed with histograms having 100 clusters/bin. The overlap between two histograms (in the case of normal and random assignment) is used as a measure of performance, i.e., the lower the overlap, the better and more selective is the similarity method and/or the HSQC signal selection algorithm. In the following, several tests were conducted with different parameter settings of the bin method and using the two HSQC signal selection algorithms with various parameter values.

Computing times for comparing one pair of spectra, including the spectra prediction and elimination of noise and solvent signals, were of the order of 2–4 s.

## 4. Results and discussion

The goal of this work was the development of a computer program capable of predicting, in a fully automatic mode, the compatibility of an HSQC NMR spectrum with a proposed chemical structure. To implement the 2D spectral comparison method, first, a peak-picking algorithm had to be developed. Signal intensities in HSQC spectra strongly vary because they depend on the coupling constants,  $^1J_{\text{CH}}$ , and instrumental parameter settings [19,20]. Therefore, in this work, different heuristic intensity measures were tested. Then, the parameters of the spectral comparison as presented in Section 2 were optimized.

### 4.1. Peak picking

One of the major advantages of HSQC spectra is their short recording time, which is shorter than for one-dimensional  $^{13}\text{C}$  NMR spectra [19]. Therefore, they are increasingly used as an alternative of 1D  $^{13}\text{C}$  NMR spectra. In view of automatic spectra interpretation, their further advantages are that the  $^1\text{H}/^{13}\text{C}$  chemical shift information is available simultaneously and that X–H signals, for which a reliable shift prediction is difficult, are absent from HSQC spectra. Nevertheless, they have also some disadvantages. Obviously, there is no coupling information, and no signals for quaternary carbon atoms are present. One major problem with HSQC spectra is that it

- read the HSQC spectrum
- normalize the integrated intensities to the range of 0–1
- find the average intensity value in the upper left corner and bottom right corner of the HSQC spectrum; this will be the starting value of the threshold
- increase the threshold level in small steps while there are isolated signals
- use the threshold to define the signals
- keep only the most intense signal along the same  $^1\text{H}$  NMR value
- keep only the strongest signal within a range of 0.06 ppm (10 digital points) in the  $^1\text{H}$  NMR direction and 1.2 ppm (3 digital points) in the  $^{13}\text{C}$  NMR direction
- keep only those signals whose summed intensity is greater than the  $1/18^{\text{th}}$  of the average sum of intensities of all remaining signals
- assign an intensity value of 1 to the remaining signals

**Scheme 1.** Algorithm for peak picking.

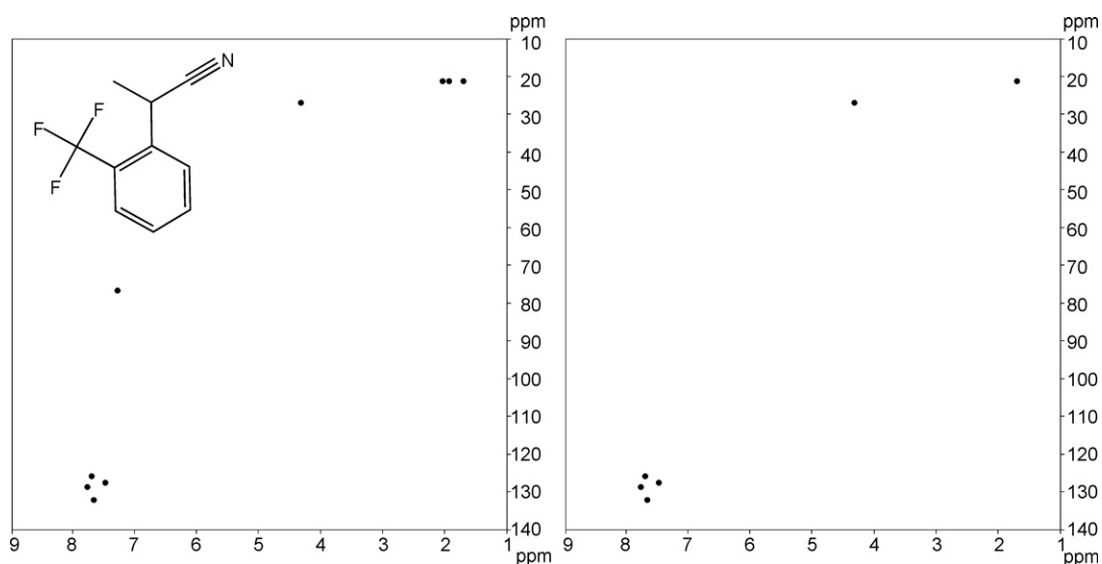
might be difficult to find all relevant signals because, as mentioned above, their intensities strongly depend on the respective one-bond coupling constants,  $^1J_{\text{CH}}$ , for which one single value is assumed in an experiment. An exact value of the pulse lengths is essential but even with optimal parameter settings, various signals may occur as artifacts [19,20].

Initially, the standard Bruker TopSpin 1.3 processing software [21] was tested for peak picking. In several cases, methods to increase the signal-to-noise ratio had to be used such as the so-called window function and linear prediction. It was possible to find all signals but the parameters had to be customized by trial and error and their optimal values were different for individual spectra. Hence, the method is not adequate for automatic processing so that another peak-picking algorithm had to be developed (see Scheme 1).

First, the intensities of the spectrum were normalized to the range of 0–1 by dividing them by the maximal intensity value. In the next step, the intensity of the noise was estimated by calculating the average intensity values in those ranges of the HSQC spectrum where no signal is expected, i.e., in the upper left and bottom right corners in a range of 0.05 ppm (11 digital points) in the  $^1\text{H}$  NMR direction and of 4 ppm (9 digital points) in the  $^{13}\text{C}$  NMR direction. This calculated average intensity of the noise was the starting value

of the threshold that defines the signals in the spectrum. The level of the threshold was then increased in small steps (0.001) until all so-called isolated signals disappeared, i.e., till there were no more points above the threshold value being surrounded by points with intensities below the threshold level. In other words, it is assumed that a true signal consists of at least two neighboring points whose intensities are above the threshold value. These points represent the initial guess of possible signals, which usually also contain a series of artifacts.

One kind of artifacts often encountered in HSQC spectra is maxima appearing around strong signals along a vertical line at the same  $^1\text{H}$  NMR chemical shift ( $t_1$  noise) [20]. To eliminate such artifacts, only the most intense signal is kept along this coordinate, using a tolerance of  $\pm 0.016$  ppm (3 digital points) for the  $^1\text{H}$  NMR shift. The minimal  $^{13}\text{C}$  NMR chemical shift difference between two signals with the same  $^1\text{H}$  NMR shift (within this tolerance) has to be  $>3.7$  ppm (10 digital points). Note that this is a heuristic approach, which may lead to a loss of signals. Another observed artifact was that, due to spectral noise, large signals were divided in two or more “islands” of signals. This was avoided by keeping only the most intense signal within a range of 0.06 ppm (10 digital points) for  $^1\text{H}$  NMR and 1.2 ppm (3 digital points) for  $^{13}\text{C}$  NMR. Even after these steps, in a number of cases, too many peaks



**Fig. 1.** A sample HSQC NMR spectrum of a target structure from the test set before (left) and after (right) automatic removal of artifact and solvent signals.

- read the  $^1\text{H}$  NMR spectrum
- normalize the integrated intensities to the range of 0–1
- remove solvent signals
- eliminate noise
- normalize the integrated intensities to the number of protons
- read the HSQC spectrum
- find the signals (cf. Scheme 1)
- remove solvent signals
- get the  $^1\text{H}$  NMR shifts of HSQC signals (project signals) in ppm
- order signals by descending ppm position
- for each signal do the following:
  - get the position of the signal in the  $^1\text{H}$  NMR spectrum
  - in a given range ( $\pm 0.15$  ppm), find the peak position (maximum intensity value); if the current signal is closer than 0.15 ppm to the next signal, then, use half the distance to the next signal position as upper range
  - in this range, find the edges of the signal: go to the left and right until the ratio of the current intensity and maximum intensity (globally in the spectrum) is very small (0.0001 ppm)
  - calculate the integral of the signal between the found edges and store it
- for each calculated integral do the following:
  - divide by the following number:  $(\text{sum} + X\text{-HNo})/\text{ProtonNo}$ , where *sum* is the sum of the integrals of the identified signals, *X-HNo* is the number of X-H and *ProtonNo* is the total number of protons in the structure
  - replace the intensity of the corresponding HSQC signal (currently 1) by the normalized value of the corresponding integral

**Scheme 2.** Assigning intensities to the HSQC NMR signals.

were retained. After analyzing the signals in terms of their maximum intensity, the number of digital points of each of them, and the sum of intensities within each signal, the algorithm was fine-tuned by keeping only those signals whose summed intensity was  $>1/18$  of the average sum of intensities of all remaining signals. Finally, the signals originating from the not fully deuterated solvents, dimethyl sulfoxide- $d_5$  (DMSO- $d_5$ ) and  $\text{CHCl}_3$  were removed. A tolerance range of  $\pm 0.02$  ppm ( $^1\text{H}$  NMR) and  $\pm 1.5$  ppm ( $^{13}\text{C}$  NMR) was applied to the expected signal positions of 2.50/39.50 ppm (DMSO- $d_5$ ) and 7.26/77.0 ppm ( $\text{CHCl}_3$ ). As an example, a HSQC NMR spectrum from the test set is shown in Fig. 1 before (left) and after (right) automatic removal of artifact signals.

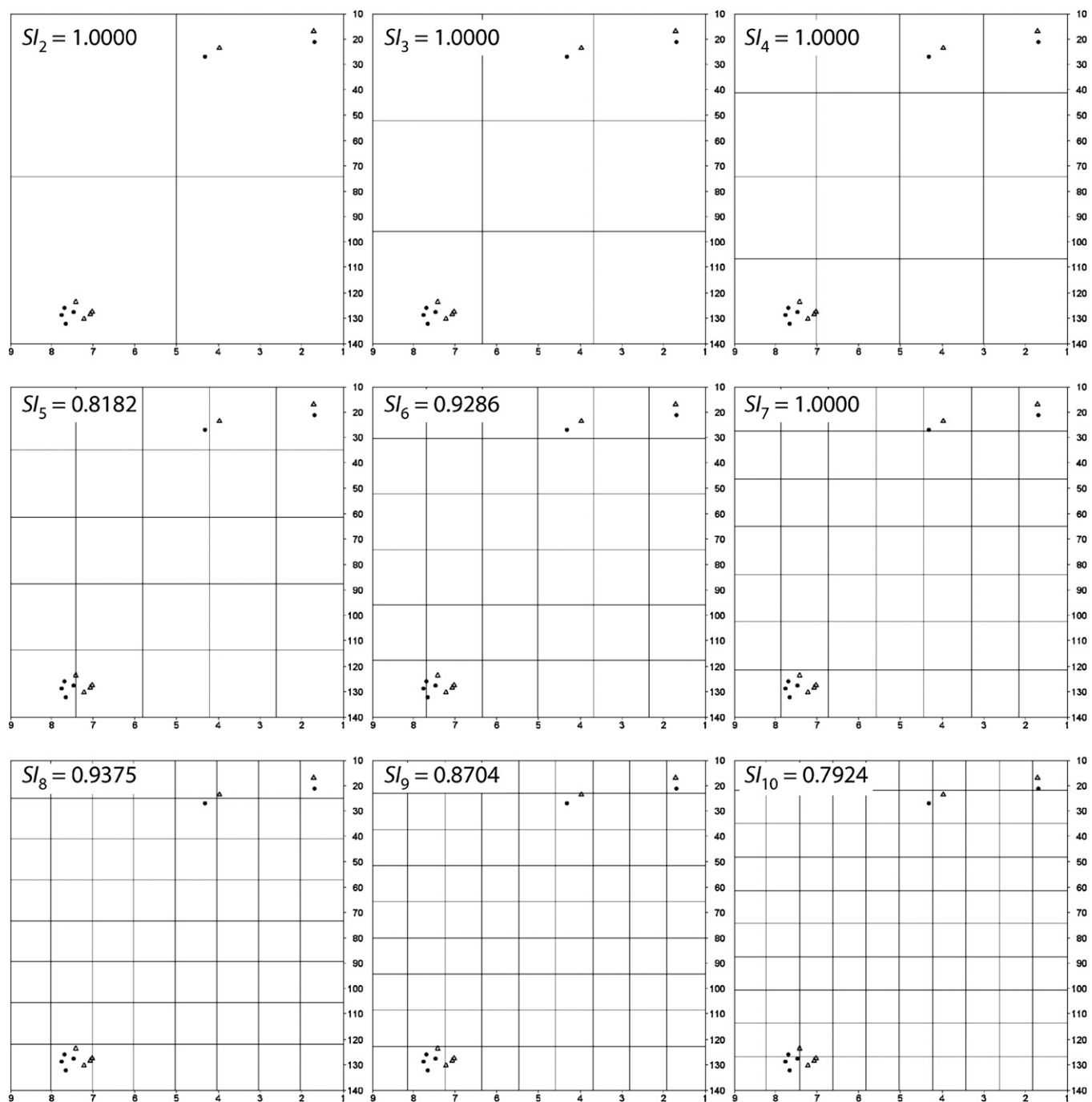
#### 4.2. Assignment of intensities

As presented in Section 2, the basis of comparing 2D NMR spectra is to divide them in varying numbers of squares, in which the signal intensities are added up. Therefore, an intensity must be assigned to each signal. Since measured intensities are biased by various possible artifacts, heuristic intensity measures were used in this work in that two different ways of representing the intensities of identified signals were applied: (1) the uniform intensity of 1 for all signals and (2) the integrated intensity of the corresponding  $^1\text{H}$  NMR signal. The algorithm for the second method is shown in Scheme 2. Here, we also used the  $^1\text{H}$  NMR spectrum belonging to the same compound as the HSQC spectrum in question. Thus, the  $^1\text{H}$  NMR spectrum was read from file and preprocessed: First, the integrated intensities were normalized to the 0–1 range; then, the solvent and noise signals were eliminated and, finally, the integrated intensities were normalized to the number of protons. Next, the signals in the HSQC spectrum were assigned (Scheme 1) and the signals of the solvent impurities were removed. Then, the  $^1\text{H}$

NMR chemical shifts of the HSQC signals (i.e., the signals projected on the *x* axis) were ordered by descending ppm positions. For each of these shifts, the corresponding signal was searched in the  $^1\text{H}$  NMR spectrum: In a given range, first, the signal peak was identified and, then, the edges of the signal were defined. The intensities between these edges were added up (equal to the integral of the  $^1\text{H}$  NMR signal) and stored. In the next step, each calculated integral was divided by  $(\text{sum} + X\text{-HNo})/\text{ProtonNo}$ , where *sum* is the sum of the integrals of all identified signals, *X-HNo* is the number of X-H, i.e., hydrogen atoms not bonded to carbon, and *ProtonNo* is the total number of protons in the structure. Finally, the normalized integral was assigned to the intensity of the corresponding HSQC signal. In the case of estimated HSQC spectra, the appropriate number of protons was attributed to each predicted HSQC signal.

#### 4.3. Comparing spectra

The 2D spectral comparison approach described in Section 2 is demonstrated in Fig. 2. The similarities,  $SI_n$  (Eq. (1)), between the measured and predicted HSQC spectra from Fig. 1 are calculated using different numbers,  $n^2$ , of bins. The signals in both spectra are represented here with the intensity value of 1. Both spectra are successively divided into  $n^2$  bins ( $n = \overline{2, 10}$ ), yielding similarity values of  $SI_n = 1.0000, 1.0000, 1.0000, 0.8182, 0.9286, 1.0000, 0.9375, 0.8704, \text{ and } 0.7924$ , respectively. A division of 10 bins in both directions (i.e., a partition of totally 100 bins) results in a minimal bin width of 13 ppm in the direction of  $^{13}\text{C}$  NMR (range: 130 ppm) and 0.8 ppm in that of  $^1\text{H}$  NMR (range: 8 ppm). Usually, for testing purposes, a larger number of bins is used: Optimal results have been achieved with  $38^2$  or  $76^2$  bins (see below), corresponding to minimal bin widths of 3.42/0.21 or 1.71/0.11 ppm, respectively.



**Fig. 2.** Illustration of the similarity calculation based on uniform intensities between measured and predicted HSQC NMR spectra (scales in ppm). The overlaid spectra are successively divided into  $n^2$  bins ( $n = 2, 10$ ), yielding a similarity index,  $SI_n$  (Eq. (1)), of 1.0000, 1.0000, 1.0000, 0.8182, 0.9286, 1.0000, 0.9375, 0.8704, and 0.7924, respectively (cf. Eq. (1)).

It is observed that, as in the one-dimensional case [14], the similarity values do not decrease monotonously, i.e., a finer division with more bins may provide a higher value of similarity than a coarser one. For example, the similarity with 7 divisions in both directions (i.e., a total of 49 bins, second row, right in Fig. 2) is equal to 1.0000, while with 5 divisions, it is only 0.8182. This behavior is a basic limitation of the bin method in that signals of two spectra (e.g., measured and calculated ones) may fall into different ranges (bins) even if the difference in their position is small. A finer division may shift the borders so that they again fall into the same bin. To reduce this adverse effect, tests were made with overlapping bins. With this approach, some regions (a given percentage) of the spectra are

shared by several bins. Thus, tests were conducted using  $38^2$  and  $76^2$  bins and various degrees of overlaps, i.e., 10%, 30%, 50%, 70%, and 90% (results not shown). In most of the cases, the results were better with nonoverlapping bins. Since the use of overlapping bins, which is rather time-consuming, did not significantly improve the selectivity of the bin method, it was abandoned.

Since  $^1\text{H}$  and  $^{13}\text{C}$  chemical shifts of  $\text{CH}_n$  groups have a certain degree of correlation, HSQC signals tend to be close to the diagonal of the individual bins. A better similarity measure is expected if the spectra are rotated so that the bins are perpendicular to the signals' main trend. Moreover, the probability that measured and predicted signals close to each other are not separated in differ-

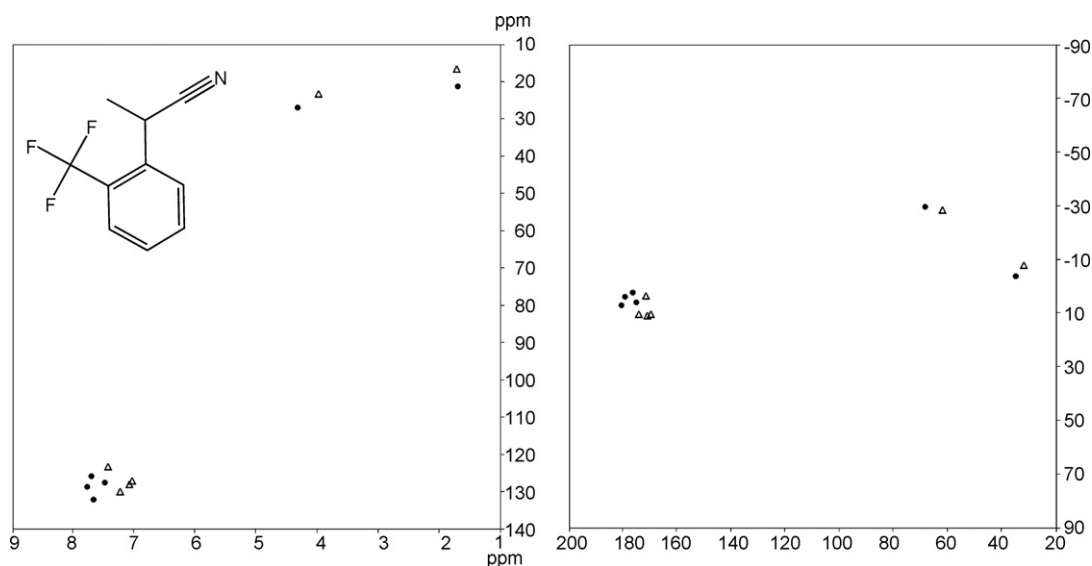


Fig. 3. Measured (dots) and predicted (triangles) HSQC NMR spectra before (left) and after (right) rotation.

ent bins is higher. Therefore, further tests were made with rotated spectra. Since rotation is done using the ppm coordinates, first, the spectrum is stretched along the  $x$  axis, i.e., along the  $^1\text{H}$  NMR coordinates, in order to obtain a square. Then, it is rotated clockwise by  $45^\circ$ . The measured and estimated HSQC spectra of the compound in Fig. 1 is shown in Fig. 3 before (left) and after (right) rotation. With a maximal number of  $N=26$  (i.e.,  $26^2$  bins), the calculated overall similarity,  $S$  (Eq. (3)), between measured and predicted HSQC spectra is 0.4425 without and 0.6727 with rotation. The results for the whole test set are given below.

The influence of the various parameters, i.e., the intensity measure, the maximal number of bins, and the rotation of the spectra, is investigated in the following. Each HSQC spectrum of the test set was compared with the spectrum estimated on the basis of the correct structure (normal assignment) and with a structure randomly selected from the test set (random assignment). In all cases, the same randomly selected target compound was used (cf. Fig. 3). The similarity values,  $S$ , are depicted as histograms for both cases (Fig. 4) and the overlap between the two distributions, i.e., between normal and random assignment,

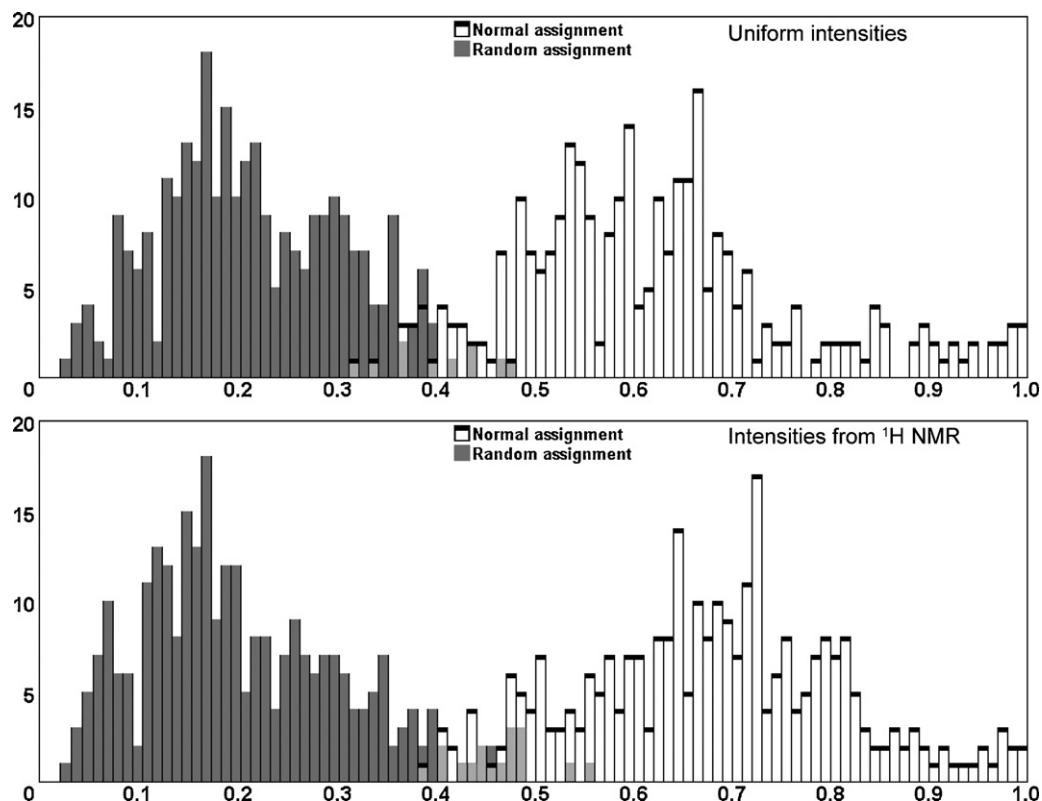
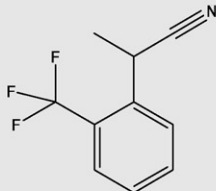
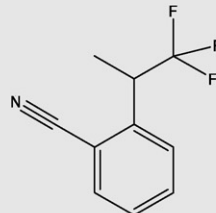
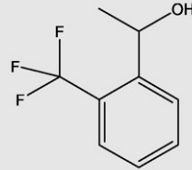
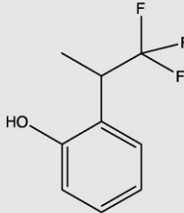
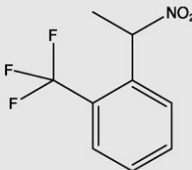
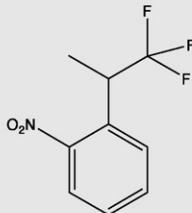
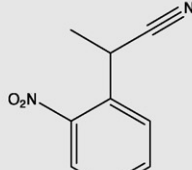
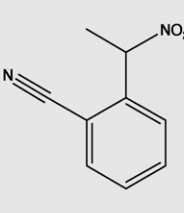
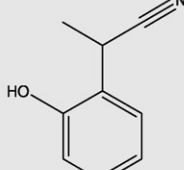
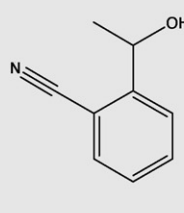
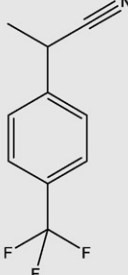
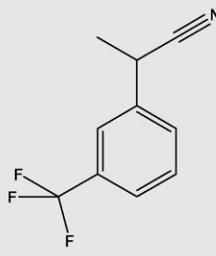


Fig. 4. Histograms of similarity values,  $S$  (Eq. (3)), of measured and calculated HSQC NMR spectra of the test set using normal (black) and random (gray) structure assignments. The overlaps between the similarity values of correctly and randomly assigned spectra are: 5.9% (uniform intensities, top) and 4.8% (intensities from  $^1\text{H}$  NMR spectra, bottom). In both cases, the maximal number of bins was  $38^2$  and the spectra were rotated clockwise by  $45^\circ$ .

**Table 1**

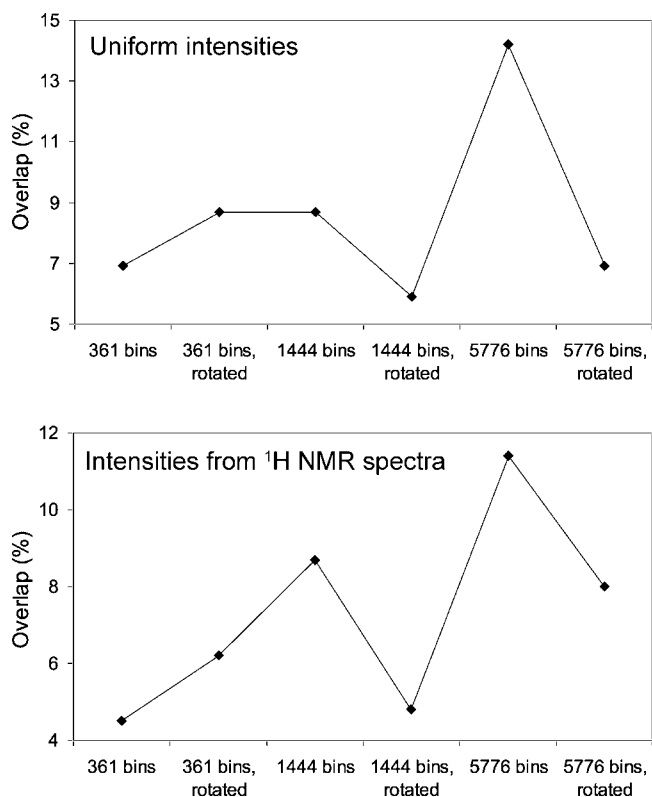
Similarity,  $S$  (Eq. (3)), obtained with  $N=26$  (first value) and  $N=38$  (second value) for the HSQC spectrum of the target molecule (first left structure, spectrum in Fig. 1, left) when compared with 11 related structures.

	0.6759 0.5483		0.4127 0.3290
	0.5937 0.4408		0.3522 0.2502
	0.6057 0.5184		0.4552 0.3584
	0.7648 0.5876		0.6065 0.4976
	0.5342 0.4281		0.6066 0.5746
	0.7233 0.5443		0.6730 0.5020

is used as a quality measure. The test results are shown in Fig. 5.

The overlap between the similarity distributions for normal and random assignments is higher throughout when using the uniform intensity of all identified HSQC signals (Fig. 5, top) than with intensi-

ties of the corresponding  $^1\text{H}$  NMR signals (Fig. 5, bottom). A detailed analysis of the results shows that this is mainly due to remaining errors in the signal assignment. Theoretically, 2281 signals should be found in the 289 spectra. The algorithm used made altogether 344 erroneous assignments (196 signals missing and 148 signals



**Fig. 5.** Influence of the number of bins ( $19^2$ ,  $38^2$ , or  $72^2$ ), the rotation of the spectra, and the intensity measure on the performance of the 2D NMR spectral comparison. For the different parameters, the overlap is shown between the histograms of overall similarity values,  $S$  (Eq. (3)), of measured and calculated HSQC NMR spectra using correct and random structure assignments. The signal intensities of the HSQC spectra were either uniform (top) or calculated from the integrals of the corresponding  $^1\text{H}$  NMR spectra (bottom).

in excess). On the other hand, using the intensities of the  $^1\text{H}$  NMR spectra, the sum of measured signal intensities is 4719 as compared with the predicted value of 4730. This small remaining difference indicates that errors in peak picking are partly compensated by the assignment of intensities from the  $^1\text{H}$  NMR spectra. With this intensity measure, the overlap between the similarities of correct and random structure assignments is 4.5% with  $19^2$  bins without rotation and 4.8% with  $38^2$  bins and rotated spectra. Since the latter corresponds to bin widths of 3.42 and 0.21 ppm ( $^{13}\text{C}$  and  $^1\text{H}$  NMR, respectively), which is close to the precision errors of the program used [18], this seems to be the best selection. Table 1 further illustrates the capabilities of the system. Of the eleven related structures shown, seven have similarity values,  $S$ , rather close to those of the correct (target) structure (top left). Obviously, the present system is mainly appropriate to detect gross errors, but further improvements are required for distinguishing closely related structures. Note, however, that with the recently improved quality of prediction (1.40 and 0.18 ppm for  $^{13}\text{C}$  and  $^1\text{H}$  NMR, respectively [22]), smaller bin widths will lead to a better discrimination of similar structures.

The resulting histograms of similarity values,  $S$ , obtained with  $38^2$  bins and rotated spectra for the correct and random assignments are presented in Fig. 4. For comparison, the corresponding histograms were calculated on the basis of the  $^1\text{H}$  NMR spectra alone using the previously described algorithm [14]. With the optimum of the minimal bin width of 0.4 ppm, the overlap between the correctly and randomly assigned spectra is 11.4%. Thus, as expected, the use of HSQC spectra with the 2D bin method is superior. Since the estimation of the chemical shifts of exchangeable protons is problematic, improved results can be expected if such protons are

excluded. This can be done by eliminating those protons, whose signals do not appear in the HSQC spectra [9]. Indeed, with this method, the results are significantly better in that the overlap between the correctly and randomly assigned spectra is reduced to 5.9% (minimal bin width, 0.4 ppm), i.e., it is nearly as low as with HSQC spectra.

Due to the different test sets and the rather low number of structures used in all studies so far, a quantitative comparison of the results of our study with those obtained earlier by Griffiths et al. [8] and Golotvin et al. [9] does not make sense. For a qualitative relationship, the number of false positives and false negatives can be read out from Fig. 4. With a similarity threshold of  $S = 0.39$ , 7.3% false positives and no false negatives occurred. With  $S = 0.55$ , there were 17.6% false negatives and no false positives. These results are comparable with those obtained earlier [8,9]. Since the approaches are different, their combination is expected to provide the best results.

## 5. Conclusions

The extension of the similarity measure, based on the comparison of signal intensities in variable numbers of bins, to two dimensions is straightforward. For applying the method to HSQC NMR spectra, automatic peak picking and assignment of signal intensities are crucial steps. The fully automatic heuristic methods developed here are not free of incorrect assignments. Corresponding errors as well as deviations between predicted and observed  $^1\text{H}$  and  $^{13}\text{C}$  NMR chemical shifts contribute to the errors of automatic compatibility tests. In spite of these errors, the results with routine spectra without any manual editing are encouraging. Since methods developed earlier are based on different approaches, their combination with that introduced here should further improve the reliability of automatic compatibility tests.

## Acknowledgements

This research was financially supported by a grant from F. Hoffmann-La Roche Ltd. We thank Dr. D. Wegmann for critical reading of the manuscript.

## References

- [1] M.E. Elyashberg, A.J. Williams, G.E. Martin, *Prog. Nucl. Magn. Reson. Spectrosc.* 53 (2008) 1–104.
- [2] L. Griffiths, *Magn. Reson. Chem.* 38 (2000) 444–451.
- [3] L. Griffiths, *Magn. Reson. Chem.* 39 (2001) 194–202.
- [4] L. Griffiths, J.D. Bright, *Magn. Reson. Chem.* 40 (2002) 623–634.
- [5] L. Griffiths, R. Horton, *Magn. Reson. Chem.* 42 (2004) 1012–1021.
- [6] L. Griffiths, *Magn. Reson. Chem.* 44 (2006) 44–58.
- [7] S.S. Golotvin, E. Vodopianov, B.A. Lefebvre, A.J. Williams, T.D. Spitzer, *Magn. Reson. Chem.* 44 (2006) 524–538.
- [8] L. Griffiths, R. Horton, *Magn. Reson. Chem.* 44 (2006) 139–145.
- [9] S.S. Golotvin, E. Vodopianov, R. Pol, B.A. Lefebvre, A.J. Williams, R.D. Rutkowske, T.D. Spitzer, *Magn. Reson. Chem.* 45 (2007) 803–813.
- [10] L. Griffiths, H.H. Beeley, R. Horton, *Magn. Reson. Chem.* 46 (2008) 818–827.
- [11] M.E. Bollard, E.G. Stanley, J.C. Lindon, J.K. Nicholson, E. Holmes, *NMR Biomed.* 18 (2005) 143–162.
- [12] H.R. Karfunkel, B. Rohde, F.J.J. Leusen, R.J. Gdanitz, G. Rihs, *J. Comput. Chem.* 14 (1993) 1125–1135.
- [13] R. de Gelder, R. Wehrens, J.A. Hageman, *J. Comput. Chem.* 22 (2001) 273–289.
- [14] L. Bodis, A. Ross, E. Pretsch, *Chemom. Intell. Lab. Syst.* 85 (2007) 1–8.
- [15] Embarcadero, San Francisco, CA 94111, U.S.A., <http://www.codegear.com/products/delphi>.
- [16] A. Fürst, E. Pretsch, *Anal. Chim. Acta* 229 (1990) 17–25.
- [17] E. Pretsch, A. Fürst, W. Robien, *Anal. Chim. Acta* 248 (1991) 415–428.
- [18] Porta Nova Software GmbH, <http://www.upstream.ch/products/nmr.html>.
- [19] J.K.M. Sanders, B.K. Hunter, *Modern NMR Spectroscopy, A Guide for Chemists*, 2nd ed., Oxford University Press, Oxford, 1993.
- [20] R. Freeman, *A Handbook of Nuclear Magnetic Resonance*, 2nd ed., Addison Wesley Longman, Harlow, 1997.
- [21] Bruker BioSpin GmbH, D-76287 Rheinstetten, Germany, <http://www.bruker-biospin.com/topspin.html>.
- [22] Modgraph Consultants Ltd., Herts AL6 0RJ, UK, <http://www.modgraph.co.uk/>.