ETH zürich Research Collection

Doctoral Thesis

Quantification of spectral similarity towards automatic spectra verification

Author(s):
Bodis, Lorant

Publication Date:
2007

Permanent Link:
https://doi.org/10.3929/ethz-a-005479205 →

Rights / License:
In Copyright - Non-Commercial Use Permitted →

This page was generated automatically upon download from the ETH Zurich Research Collection. For more information please consult the Terms of use.

ETH Library

Diss. ETH No. 17361

Quantification of Spectral Similarity:
Towards Automatic Spectra Verification

A dissertation submitted to
ETH ZURICH

for the degree of
DOCTOR OF SCIENCES

presented by
LORANT BODIS
MSc Computer Science, Babes-Bolyai University
born 23.03.1980
citizen of Romania

accepted on the recommendation of
Prof. Dr. Ernö Pretsch, examiner
Prof. Dr. Gabor Székely, co-examiner

2007

Acknowledgments

My warm thanks go to:

Prof. Dr. Ernö Pretsch for the opportunity to work in his group, for his scientific and personal guidance, for the numerous creative and productive discussions, and for his readiness to help; in the last three years he gave me the chance to develop myself not only professionally but also personally;

Prof. Dr. Gabor Székely for accepting the co-examination of this doctoral thesis;

Dr. Alfred Ross for his valuable collaboration and promptitude and for the new ideas exchanged during our conversations in Zurich and Basel;

Dr. Dorothée Wegmann for the very conscientious and precise correction of the manuscript and for her kindness and support during my entire PhD studies;

Finally, F. Hoffmann-La Roche AG, Basel, for the financial support.

Motto:

Per aspera ad astra

Latin saying

Szüleimnek

Table of contents

1 Summary 9

2 Zusammenfassung 13

3 Similarity of patterns 17

3.1 Classical methods for vector-based similarity 17
  3.1.1 Euclidean distance 17
  3.1.2 City-block distance 19
  3.1.3 Mahalanobis distance 20
  3.1.4 Correlation coefficient 21
  3.1.5 Dot product 26
  3.1.6 Tanimoto coefficient 27

3.2 Pattern recognition and image analysis 29
  3.2.1 Clustering 29
  3.2.2 Classification of clustering techniques 30
  3.2.3 Application of clustering 34

3.3 Spectra similarity measures: Mass spectra and IR spectra 36

3.4 Match probability 37

3.5 Chromatograms 39

3.6 Metabolomics and metabonomics: ^{1}H NMR of biological fluids 40

3.7 New similarity methods using shifts 42
  3.7.1 Binning 42
  3.7.2 The spectral fold 42
  3.7.3 Weighted cross-correlation method 44

3.8 Comparison of similarity criteria 45

3.9 Performance metrics 46
  3.9.1 Contingency tables 46
  3.9.2 Histograms 49

4 A novel spectra similarity measure: the bin method 51

4.1 Introduction 51

4.2 Test sets 52

4.3 Limitations of weighted cross-correlation method 54

4.4 Novel similarity criterion 55

4.5 Overlapping bins 58

4.6 Tests with artificial spectra 60

4.7 Tests with measured spectra 62

4.8 Conclusions 63

5 Similarity of ^{1}H NMR spectra 65

5.1 Elimination of noise 65

5.2 Elimination of solvent signals 66

5.3 Test set ... 68

5.4 Results and discussion ... 69

5.4.1 Influence of parameters on the performance of compatibility tests with ^{1}H NMR spectra ... 69
5.4.2 Eliminating X–H signals from ^{1}H NMR spectra ... 73

5.5 Conclusions ... 77

6 Automatic spectra verification of a combinatorial library ... 79

6.1 Introduction ... 79

6.2 Combinatorial library ... 80

6.3 Generating reference spectra ... 82

6.3.1 Comparison with predicted spectra ... 82
6.3.2 Comparison with other measured product spectra ... 82
6.3.3 Comparison with the sum of the measured spectra of the educts ... 83
6.3.4 Comparison with the individual measured spectra of the educts ... 83

6.4 Results and discussion ... 83

6.5 Conclusions ... 94

7 Similarity of HSQC spectra ... 95

7.1 Introduction ... 95

7.2 Finding the signals in the measured HSQC spectra ... 96

7.2.1 Using uniform intensity values of 1 ... 98
7.2.2 Counting the number of protons for the measured HSQC signals ... 101

7.3 Elimination of solvent signals ... 108

7.4 Prediction of the HSQC spectra ... 108

7.4.1 Using intensity values of 1 ... 108
7.4.2 Counting the number of protons for the predicted HSQC signals ... 110

7.5 Bin method applied to HSQC spectra ... 111

7.5.1 Overlapping bins ... 113

7.6 Rotating the HSQC spectra ... 115

7.7 Test set ... 116

7.8 Results and discussion ... 116

7.9 Conclusions ... 127

8 Similarity of IR spectra ... 129

8.1 Data set ... 129

8.2 Clustering ... 130

8.3 Results and discussion ... 131

8.4 Conclusions ... 134

9 Conclusions ... 135

10 Appendix ... 137

References ... 141

Curriculum vitae ... 149

1 Summary

Nowadays, especially in the process of drug discovery, new compounds are easily generated by parallel syntheses. High-throughput instruments are capable of automatically registering ^{1}H NMR spectra on the order of minutes. Thus, the bottleneck of structure verification is the interpretation of molecular spectra. So far, efficient tools for a reliable automatic interpretation are not yet available.

The aim of the present project was to develop an automatic spectra verification software tool that is capable of classifying spectra in terms of their quality and correctness of the proposed chemical structures. 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. Various methods have been proposed for describing the spectral and structural similarity. However, most spectral similarity measures only work with spectra showing relatively broad signals since they fail to detect similarities if the signal positions in the two spectra differ by more than the signal widths. In other words, they are unable to detect information about the neighborhood of a signal. Such similarity measures would not be helpful when comparing NMR spectra because changes in signal positions of up to ca. 100 times the signal half-widths are to be expected even for closely related spectra. Thus, the goal was to develop a new similarity measure that is selective and able to discriminate between related and unrelated ^{1}H NMR spectra.

In Chapter 3, methods for quantifying the similarity of vectors are reviewed that have been used for mass, IR, and UV spectra, but not for ^{1}H NMR spectra.

A novel spectral similarity measure, called bin method, is introduced in Chapter 4. Its performance is characterized by means of contingency tables and histograms, first, with artificial and then, with real-life ^{1}H NMR spectra. Using a computer-generated test set, its power to discriminate between related and unrelated ^{1}H NMR spectra is found to be better than with the recently proposed cross-correlation method. Superior results are also obtained when comparing measured ^{1}H NMR spectra of a database with the corresponding estimated ones or with estimated spectra of randomly assigned structures.

In a next step, the bin method is applied to the comparison of ^{1}H NMR spectra using two further test sets. The first test set (Chapter 5) consists of 289 chemical structures and the corresponding ^{1}H NMR spectra. Automatic solvent signal and noise elimination were

developed as preprocessing tools, which are applied prior to the comparison of the spectra. Further tests are conducted with spectra from which the X–H signals are excluded both from the measured and predicted ^{1}H NMR spectra. Since the chemical shifts of such signals highly depend on experimental conditions, they are difficult to be predicted accurately. As expected, the performance of the bin method is, thus, improved. In order to remove X–H signals from the measured ^{1}H NMR spectrum, the corresponding measured HSQC spectrum is used because it only shows signals of hydrogens attached to carbons.

In Chapter 6, a second test set with 96 ^{1}H NMR spectra is used, which are obtained by parallel syntheses (combinatorial library). For this library, different methods of automatic spectra verification are developed. On the one hand, the problem is more challenging than in the previous tests because the structures and, therefore, the spectra are highly similar. On the other hand, since the data set is a combinatorial library, the measured spectra can be compared not only with the predicted ones but also with other spectra of the library, e.g., with product spectra, or combined or individual spectra of the educts (reagents). The performance of the methods is characterized with contingency diagrams and the optimal threshold values are defined. Finally, for each spectrum of the combinatorial library, the results of the different approaches are represented in traffic light style. The outcome of the automatic spectra verification is compared with the results of the manual analysis. Gross errors in the test set are readily detected with all the methods described here.

Due to the generality of the approach, the bin method can easily be extended and applied to comparing other types of spectra or patterns as well. In Chapter 7, the novel similarity measure is adapted to two-dimensional spectra. Because of their high practical relevance, the HSQC spectra were chosen. To test the compatibility of such a spectrum with the proposed structure, first, the spectrum is predicted on the basis of the proposed structure. Then, it is compared with the measured one. In this context, the automatic peak picking, the use of the various signal intensity measures, and the two-dimensional bin method had to be optimized first. Several procedures of finding the signals in the HSQC spectra and defining their intensities are presented. In a preprocessing step, solvent signals and noise are eliminated from the raw HSQC spectra. Tests using various parameter settings and rotated spectra are also performed. The results are even better with two- than with one-dimensional NMR spectra. Moreover, this is the first similarity measure applied to HSQC spectra.

Finally, in Chapter 8, the bin method is applied to the comparison of IR spectra. Since IR spectra have relatively broad signals, they can be efficiently compared using the correlation coefficient, which has been extensively described in the literature. Here, IR spectra from a large available database are clustered and analyzed using the similarities calculated with the bin method. As expected, the improvements in this case are not as remarkable as with NMR spectra. Nevertheless, based on the outcome, it can be stated that the performance of the bin method is competitive with that of a commercial software.

2 Zusammenfassung

Heutzutage werden neue Verbindungen besonders bei der Suche nach Medikamenten sehr leicht durch Parallelsynthesen hergestellt. Da Hochleistungsinstrumente die automatische Aufnahme von ^{1}H-NMR-Spektren in wenigen Minuten ermöglichen, liegt der Engpass bei der Strukturaufklärung in der Interpretation der Molekülspektren. Bis jetzt sind jedoch noch keine effizienten Hilfsprogramme für die zuverlässige automatische Spektreninterpretation erhältlich.

Ziel der vorliegenden Arbeit war die Entwicklung eines Computerprogramms für die automatische Spektrenverifizierung, das Spektren auf ihre Qualität und die Richtigkeit der vorgeschlagenen chemischen Strukturen prüft. Bei der automatischen Interpretation muss vor allem die Ähnlichkeit zwischen dem gemessenen Spektrum und einem Referenzspektrum definiert werden, das von einer Datenbank oder Computervorhersage stammt. Zwar gibt es bereits verschiedene Methoden zur Beschreibung der Ähnlichkeit zwischen Spektren und Strukturen. Sie funktionieren aber meistens nur mit Spektren, die relativ breite Signale zeigen, da sie Ähnlichkeiten nur finden können, wenn die Peaklage in den zu vergleichenden Spektren weniger als eine Signalbreite auseinanderliegt. Das heisst, dass sie keine Information über die Umgebung eines Signals erfassen können. Solche Methoden zur Ähnlichkeitsbestimmung würden beim Vergleich von NMR-Spektren nichts nützen, da die Peaklagen auch bei sehr ähnlichen Verbindungen bis zur 100-fachen Signalhalbwertsbreite ändern können. Das Ziel bestand deshalb darin, eine neue Methode zu finden, die selektiv ist und zwischen verwandten und fremden ^{1}H-NMR-Spektren unterscheiden kann.

Kapitel 3 gibt einen Überblick über Methoden zur Quantifizierung der Ähnlichkeit von Vektoren, die bis jetzt für Massen-, IR- und UV-, nicht aber für ^{1}H-NMR-Spektren verwendet werden.

In Kapitel 4 wird die Bin-Methode als neuartiges Kriterium für die spektrale Ähnlichkeit eingeführt, deren Leistung anhand von Kontingenztabellen und Histogrammen zunächst mit künstlichen und dann mit realen ^{1}H-NMR-Spektren charakterisiert wird. Mit Hilfe eines rechnergenerierten Testsets wurde gefunden, dass ihre Fähigkeit, zwischen verwandten und fremden ^{1}H-NMR-Spektren zu unterscheiden, besser ist als die kürzlich vorgeschlagene Kreuz-Korrelations-Methode. Bessere Resultate erhielt man auch beim Vergleich von gemessenen ^{1}H-NMR-Spektren einer Datenbank mit den entsprechenden geschätzten

Spektren oder mit ebensolchen von zufällig zugeteilten Strukturen.

Als nächstes wird die Bin-Methode anhand von zwei weiteren Testsets zum Vergleich von ^{1}H-NMR-Spektren eingesetzt. Das erste Testset (Kapitel 5) besteht aus 289 chemischen Strukturen mit ihren ^{1}H-NMR-Spektren. Mit einem zur Entfernung der Lösungsmittelsignale und des Rauschens entwickelten Computerprogramm werden die Spektren vor dem Vergleich bereinigt. In weiteren Tests werden die X–H-Signale aus den gemessenen und vorhergesagten Spektren entfernt, was, wie erwartet, die Leistungsfähigkeit der Bin-Methode verbessert. Da die chemischen Verschiebungen solcher Gruppen stark von den experimentellen Bedingungen abhängen, ist es schwierig, sie genau vorauszusagen. Zur Eliminierung der X–H-Signale aus den gemessenen ^{1}H-NMR-Spektren stützt man sich auf die zugehörigen HSQC-Spektren, die ja nur Signale von an Kohlenstoff gebundenen Wasserstoffatomen aufweisen.

In Kapitel 6 wird das zweite Testset mit 96 ^{1}H-NMR-Spektren verwendet, die aus einer kombinatorischen Bibliothek mit durch Parallelsynthesen erhaltenen Verbindungen stammen. Dazu werden verschiedene Methoden der Spektrenverifizierung entwickelt. Einerseits ist die Problemstellung anspruchsvoller als bei den vorhergehenden Tests, da die untersuchten chemischen Strukturen und ihre Spektren sehr ähnlich sind. Da anderseits der Datensatz eine kombinatorische Bibliothek ist, können die gemessenen Spektren nicht nur mit den vorhergesagten verglichen werden, sondern auch mit anderen in der Bibliothek, z. B. mit jenen von Produkten und kombinierten oder Einzelspektren der Edukte (Reagenzien). Die Leistungen der Methoden werden mittels Kontingenzdiagrammen charakterisiert und die optimalen Schwellenwerte definiert. Schliesslich werden die Resultate der verschiedenen Versuche für jedes Spektrum der kombinatorischen Bibliothek im Verkehrsampelstil dargestellt. Das Ergebnis der automatischen Spektrenverifizierung wird verglichen mit jenem, das man manuell erhält. Mit allen hier beschriebenen Methoden werden grobe Fehler im Testset leicht gefunden.

Dank ihres allgemein gültigen Ansatzes kann die Bin-Methode einfach erweitert und auch beim Vergleich von anderen Arten von Spektren oder Mustern eingesetzt werden. Im Kapitel 7 wird das neue Ähnlichkeitskriterium auf zweidimensionale Spektren angewendet. Dazu wählte man HSQC-Spektren wegen ihrer hohen praktischen Relevanz. Um die Kompatibilität eines solchen Spektrums mit einer vorgeschlagenen Struktur zu testen, wird das HSQC-Spektrum zuerst aufgrund dieser Struktur berechnet und dann mit dem gemessenen verglichen. Dafür musste man vorher die automatische Auswahl der Signale, den Gebrauch der verschiedenen Methoden für die Signalintensität und die zweidimensionale Bin-Methode optimieren. Es werden verschiedene Verfahren vorgestellt, um die HSQC-Signale zu finden und ihre Intensitäten zu bestimmen. Vorgängig werden Lösungsmittelsignale und das Rauschen aus den Rohdaten der HSQC-Spektren eliminiert. Unterschiedliche Parametersätze sowie der Einfluss der Spektrenrotation werden getestet. Die Resultate sind sogar noch besser mit zwei- als mit eindimensionalen NMR-Spektren. Zudem ist dies das erste Mal, dass ein Ähnlichkeitskriterium auf HSQC-Spektren angewendet wird.

In Kapitel 8 wird die Bin-Methode auch noch zum Vergleich von IR-Spektren eingesetzt. Da diese relativ breite Signale aufweisen, können sie effizient mit Hilfe des Korrelationskoeffizienten verglichen werden, was in der Literatur sehr häufig beschrieben wird. Hier nun werden IR-Spektren aus einer großen verfügbaren Datenbank in Cluster geordnet und anhand der mit der Bin-Methode berechneten Ähnlichkeitsmasse analysiert. Wie erwartet, sind die Verbesserungen in diesem Fall nicht so beachtlich wie mit NMR-Spektren. Trotzdem kann aufgrund der Ergebnisse festgestellt werden, dass die Leistung der Bin-Methode durchaus mit jener der kommerziellen Software konkurrenzfähig ist.

3 Similarity of patterns

In the first part of the present chapter, methods for vector-based similarities are reviewed. Most of them have been extensively used in the literature for comparing spectra since these can be represented as vectors. Several spectral similarity measures used for mass and IR spectra are presented here. However, they do not include any information about the signal neighborhood and are, therefore, commonly called pointwise similarity criteria. While in the case of mass and IR spectra, the classical distance metrics are capable of accurately calculating the similarities, this does not hold for NMR spectra. In a further section, several methods are presented that detect similarities between NMR spectra taking into account the chemical shifts.

Metabonomic studies of biological fluids are currently very frequent experiments and mostly imply the comparison of ^{1}H NMR spectra. Several applications are outlined in this chapter.

With the development of high-throughput analytical instruments, an enormous amount of data is generated so that chemometric processing of chromatographic data becomes more and more important and widespread. These processing methods are listed in an additional section.

Finally, two types of performance metrics are presented that were used in the experimental part (Chapter 4–7) of the thesis.

3.1 Classical methods for vector-based similarity

Since similarity is fundamental to the definition of a cluster, a measure of similarity between two patterns is essential. There exist numerous distance measures, which have to be chosen carefully for each task. They can be used for problems for which the data are represented as vectors. In the following, the most important distance measures are presented, at the same time pointing out some of their recent applications.

3.1.1 Euclidean distance

The Euclidean distance[1-3] is the most popular distance measure. The Euclidean metric is the function $d_2 : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ that assigns to any two vectors, $x_i = \left(x_{i1},...,x_{in}\right)$ and $x_j = \left(x_{j1},...,x_{jn}\right)$, in the Euclidean n-dimensional space (number of variables) the following real number:

$$d_2(x_i,x_j)=\left(\sum_{k=1}^{n}\left(x_{i,k}-x_{j,k}\right)^2\right)^{\frac{1}{2}}=\left\|x_i-x_j\right\|_2 \qquad 3.1$$

and so gives the "standard" distance between any two vectors in $\mathbb{R}^n$.

The Euclidean distance is the special case of the Minkowski metric:[1]

$$d_p(x_i,x_j)=\left(\sum_{k=1}^{p}\left(x_{i,k}-x_{j,k}\right)^p\right)^{\frac{1}{p}}=\left\|x_i-x_j\right\|_p \qquad 3.2$$

The Euclidean distance has an intuitive appeal as it is commonly used to evaluate the proximity of objects in the two- or three-dimensional space. It works well when a data set has compact or isolated clusters.

The Euclidean distance is widely used in science and engineering of today. It is the easiest and most straightforward method to calculate the distance between objects, i.e., their similarity, or to quantify the performance of different techniques and procedures. Recently, for example, it has been used as a quantification tool in a principal components space for computer screen photo-assisted classification (CSPT).[4] This technique is a method for the evaluation of colorimetric assays using regular computer sets and web cameras as the only instruments. The method is able to distinguish between different color substances, and its performance depends on the color of the evaluated substances and on the particular illuminating sequences used for the measurements. The classification capability of the technique allows to identify specific substance transmittances, which is relevant in the design of assays customized for CSPT.[4]

The Euclidean distance is often used in hierarchical clustering for calculating the distance between clusters. As an example, the cluster analysis of the data is done with the Euclidean distance and Ward’s linkage.[5]

In order to compute microstructural gradients at interfaces in composite materials, an image analysis procedure was introduced using Euclidean distance mapping (EDM).[6] This method is a basic operation applied in computer vision, pattern recognition, and robotics, where high computation speed is essential. It uses distance transformation to convert a binary image consisting of foreground and background pixels into a grayscale image where each pixel has a brightness value equal to its Euclidean (linear) distance to the nearest background pixel. A very similar method, the gray-weighted distance transform (GDT) is used in a further study.[7] For more explanations, see Section 3.2.3.

Fluorescence spectra and Fourier transform (FT)-IR spectra of diesel oils have also been compared.[8] For matching the spectra, the Euclidean distance, the correlation coefficient, and a derivative of these parameters (squared difference correlation coefficients) are used. The algorithm has a very practical application in detecting environmental spills (e.g., diesel oil) and gives a clear indication of whether two samples did come from the same source.

A recent application of the Euclidean distance is the point scattering, where the scattering of a point ensemble is defined in terms of the Euclidean distance matrix.[9] Point scattering is a new geometric invariant denoting a quantity that remains unchanged during certain transformations such as translation or rotation. It is practical for comparing objects as it usually denotes intrinsic properties of the objects. Point scattering is useful for studying clusters, molecules, crystals, and biomolecules. It was tested with natural clusters of hard spheres, such as colloidal particles and fullerenes, as well as of protein-peptide complexes. The results are promising as the new method is capable of differentiating point ensembles with different structures, which are not distinguished by other geometric invariants.[9]

3.1.2 City-block distance

The city-block distance[10, 11] (also called Hamming distance, Manhattan distance, or Taxicab metric) between two items (vectors) is the sum of the differences of their corresponding components (variables). The city-block distance (defined in $\mathbb{R}^n$) of vectors $x_i = \left(x_{i1},\ldots,x_{in}\right)$ and $x_j = \left(x_{j1},\ldots,x_{jn}\right)$ is given by the following expression:

$$d(x_i,x_j) = \sum_{k=1}^{n} \left| x_{i,k} - x_{j,k} \right| \qquad 3.3$$

In two dimensions and with vectors of discrete value represented as a grid, it simply is the number of edges between points that must be traversed to get from *x* to *y* within the grid. This is the same problem as getting from corner x_i to x_j in a rectilinear downtown area, hence the name city-block metric (for a graphical visualization, see Figure 3.1).

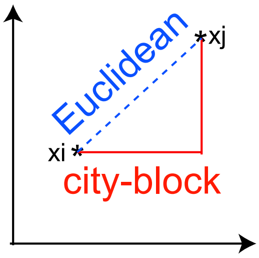


Figure 3.1. City-block distance versus Euclidean distance.

The city-block metric is not as popular as the Euclidean one, but has the big advantage of being computationally cheap. It has recently been used in an algorithm that recognizes trace elemental homeostasis in serum samples of patients with Parkinson disease.[10]

In an earlier publication, the city-block metric was applied to implement a narrow-band pattern-matching model of vowel perception.[11] The narrow-band model classified vowels in a large database (12 vowels spoken by 45 men, 48 women, and 46 children) with an accuracy degree of 91.4%, approaching that of human listeners.

3.1.3 Mahalanobis distance

The Mahalanobis distance[1-3, 12] is a widely used multivariate distance measure that calculates the distance between an object and the centroid (mean) of a cluster. Unlike the Euclidean distance, the Mahalanobis distance takes into account the covariance (size and shape) of a cluster.

The squared Mahalanobis distance, $d_M(x_i, x_j)$, of the row vectors $x_i = \left(x_{i1}, ..., x_{in}\right)^T$ and $x_j = \left(x_{j1}, ..., x_{jn}\right)^T$ follows the equation:

$$d_M(x_i, x_j) = (x_i - x_j)\sum{}^{-1}\left(x_i - x_j\right)^T \tag{3.4}$$

where $\sum$ is the sample covariance matrix of the patterns. If the covariance matrix corresponds to the identity matrix, the Mahalanobis distance reduces to the Euclidean distance. If the covariance matrix is diagonal, then, the resulting distance measure is called the normalized Euclidean distance:

$$d_M(x_i, x_j) = \sqrt{\sum_{k=1}^{n} \frac{(x_{i,k} - y_{j,k})^2}{\sigma_i^2}} \tag{3.5}$$

where σ_i is the standard deviation of x_i over the sample set.

Among other methods, the Mahalanobis distance is used for distinguishing pathologically normal, premalignant, and malignant oral tissues.[13] Then, a principal component analysis[14] is performed for the spectral analysis and classification of these three cases.

The Mahalanobis distance is used with the Mahalanobis-Taguchi system (MTS) in order to classify hot rolled steel products based on their chemical composition.[15] The MTS is a powerful process for recognizing patterns and forecasting results.[16] It is a pattern information technology, which has been used for making quantitative decisions by constructing a multivariate measurement. In the MTS approach, Mahalanobis distance is applied to measure the degree of abnormality of patterns. The principles of the Taguchi methods are used to evaluate the accuracy of predictions based on the constructed scale. Very recently, the different stages of the MTS method have been summarized.[15]

3.1.4 Correlation coefficient

The Pearson product-moment correlation coefficient (commonly called correlation coefficient) is a quantity that gives the degree of association between two random variables.[17] At the same time, it indicates the strength and direction of a linear relationship between these two random variables. The squared correlation coefficient, r^2, is obtained by dividing the covariance, ss_{xy}^2, of the two variables by the product of their standard deviations:[18]

$$r^2 = \frac{ss_{xy}^2}{ss_{xx} ss_{yy}} \qquad 3.6$$

where ss_{xx} and ss_{yy} are the sums of squared values of a set of n data points (x_i, y_i) about their respective means:

$$ss_{xx} = \sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2 \qquad 3.7$$

$$ss_{yy} = \sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2 \qquad 3.8$$

$$ss_{xy} = \sum_{i=1}^{n} \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right) \qquad 3.9$$

with:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{and} \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \qquad 3.10$$

These quantities are simply unnormalized forms of the variances and covariance of X and Y given by:

$$ss_{xx} = N\,\mathrm{var}(X) \qquad 3.11$$

$$ss_{yy} = N\,\mathrm{var}(Y) \qquad 3.12$$

$$ss_{xy} = N\,\mathrm{cov}(X, Y) \qquad 3.13$$

where N is the degree of freedom.

As geometric interpretation, the correlation coefficient can also be viewed as the cosine of the angle between the two mean-centered vectors of samples drawn from the two random variables:

$$\cos\theta = \frac{x \cdot y}{\|x\| \|y\|} \tag{3.14}$$

This is closely interconnected with the dot product. The dot product of vectors x and y is defined as follows (cf. Section 3.1.5):

$$x \cdot y = \|x\| \|y\| \cos\theta \tag{3.15}$$

where $\|x\|$ and $\|y\|$ denote the length (magnitude) of vectors x and y, and θ is the angle between them. Thus, it is seen that the covariance is the mean-corrected dot product, and the squared correlation coefficient is the covariance corrected by the standard deviation.

The Pearson product-moment correlation coefficient as defined above was developed by Pearson at the beginning of the 20th century[19, 20] and is based on the work of Galton,[21] who was the first to introduce the concept of correlation. Actually, correlation charts, also known as scatter diagrams, are one of the basic tools of statistical quality control.

The correlation coefficient is much more complex than it seems and has a wide range of applications. Special cases and extended versions of the correlation coefficient have been developed for different problems. For example, the multiple correlation coefficient, R^2, is the percent of the variance in the dependent variable explained uniquely or jointly by the independent ones. It can also be interpreted as the proportionate reduction in error in estimating the dependent when knowing the independent variables. It generalizes the standard coefficient of correlation and is used in multiple regression analysis to assess the quality of prediction of the dependent variable.[22, 23] That is, R^2 reflects the number of errors made when using the regression model to guess the value of the dependent, in ratio to the total error made when using only the mean of the dependent as the basis for estimating all cases. More precisely, if $(x_1, \ldots, x_k)$ is a random vector with values in R^k, then, the multiple-correlation coefficient between x_1 and $x_2, \ldots, x_k$ is defined as the usual correlation coefficient between x_1 and its best linear approximation, $E(x_1 \mid x_2, \ldots, x_k)$, relative to $x_2, \ldots, x_k$, i.e., as its regression relative to $x_2, \ldots, x_k$. If more than two independent variables are present, the use of matrices is recommended. More generally, it is written in the following form:

$$R^2 = \frac{SS_{regression}}{SS_{regression} + SS_{error}} = \frac{SS_{regression}}{SS_{total}} \tag{3.16}$$

with $SS_{regression}$ as the regression sum of squares, SS_{error} as the residual (or error) sum of squares, and SS_{total} as the total sum of squares. This shows that R^2 can also be interpreted as the proportion of variance of the dependent variable explained by the independent ones.

Another special case, the partial correlation coefficient, ρ, is a measure of the linear dependence of a pair of random variables from a collection of random variables, i.e., "a correlation when all other variables are kept at fixed values". The partial correlation coefficient, $\rho_{12;3}$, can be expressed in terms of the Pearson correlation coefficients, r:[24, 25]

$$\rho_{12;3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1-r_{13}^2)(1-r_{23}^2)}} \qquad 3.17$$

where $\rho_{12;3}$ is the partial correlation between x_1 and x_2 holding x_3 constant, while r_{12}, r_{13}, and r_{23} are the product-moment correlation coefficients between x_1, x_2, and x_3, as indicated by the subindex. The partial correlation coefficient has recently been applied to predict protein-protein interactions. For more information, see the literature.[26]

The correlation coefficient, r, is one of the most popular measures of spectral similarity. For comparing IR spectra, it has been used already in the dawn of the computer era[27] and is still popular nowadays. Baumann and Clerc[28] compared the performance of the Pearson product-moment correlation coefficient with Spearman's rank correlation coefficient and Kendall's τ rank correlation coefficient. Spearman's rank correlation coefficient is simply a special case of the Pearson product-moment coefficient in which the data are converted to a ranking list before calculating the coefficient. It is a non-parametric measure of correlation and measures how well an arbitrary monotonic function can describe the relationship between two variables without making any assumptions about the frequency distribution of the variables.[29] Kendall's τ rank correlation coefficient is used to measure the degree of correspondence between two rankings and assessing the significance of this correspondence.[30]

In a recent contribution, it is shown that among four investigated spectra similarity measures for searches in an IR spectra database, the correlation coefficient performs best.[31] The other three similarity criteria, the dot product, the mean of the absolute differences, and the mean of the squared differences, proved to be less optimal. The same authors have also successfully applied the correlation coefficient to the comparison of mass spectra.[32]

A comprehensive overview of the correlation coefficient presents numerous graphical illustrations and examples of recent chemical applications.[17] It is shown that, although the correlation is interrelated with the regression, both are quite different. The correlation gives the degree of association of two independent variables, while the regression provides the functional relationship between given values of the first variable (the independent one) and the means of all corresponding values of the second variable (dependent or response variable). In a recent thesis,[33] the correlation coefficient is applied to the comparison of IR spectra in order to perform automatic structure-spectra compatibility tests.

The correlation coefficient has been successfully used also for the comparison of UV spectra;[34] moreover, the book section gives a comparison of various similarity measures applied to UV spectra, including besides the correlation coefficient, the cosine of the angle and a few matching procedures.[34]

Unfortunately, the correlation coefficient is more and more misused and confused with regression. In several works, it is incorrectly used for expressing the quality in linear regression analysis.[35] The correlation coefficient is unable to evaluate linearity and it is useless as a quality measure for estimating parameters or constants. The paper presents pitfalls in statistical regression analysis and examples of misuse and misinterpretation of the correlation coefficient. Several alternatives of correlation are described and more efficient and informative quality methodologies and calibration methods are presented.

An extended version of a dynamic time warping algorithm was introduced recently and used for the alignment of chromatography mass spectrometry proteomics data sets.[36] It is shown that the Pearson product-moment correlation coefficient outperforms the Euclidean distance, the dot product, and the covariance method.

In another work,[37] the so-called sample-sample (SS) correlation spectroscopy is presented, which denotes the correlation between samples, while that showing the relation between intensities of a signal is called wavenumber-wavenumber (WW) correlation. In order to carry out the analysis, the correlation matrices, $\mathbf{X}\mathbf{X}^T$, are generated, which gives the SS correlation maps, and $\mathbf{X}^T\mathbf{X}$, which yields the WW correlation maps, $\mathbf{X}$ being the matrix of experimental data with the spectra aligned in rows. The contour plots are relatively easy to interpret and can precede a principal component analysis.[14] In the paper, it is shown that there is no need for any special algorithm or programming routine to classify near infrared (NIR) spectra.

In a novel application of the two-dimensional correlation coefficient,[38] NIR spectra are analyzed and the spectral variations are classified. Two-dimensional (2D) correlation spectroscopy is able to detect tricky spectral variations and overlaps in various types of spectra, the results usually being visualized in the form of a contour plot.

Recently, the statistical correlation coefficient mapping has been successfully applied to the tissue classification for cancer diagnosis.[39] Diagnosis of cancer at an early stage is a key to timely treatment and, thus, improved survival rate of the patients. Tissue fluorescence spectra (which allow to achieve an early cancer diagnosis) have been analyzed with the statistical correlation coefficient mapping method. A two-dimensional contour plot of the correlation coefficients with respect to two sample axes is called the sample-sample correlation coefficient map. This is very similar to the method described previously.[38] The classification of spectra is done by quantitative analysis of the correlation coefficients within the data set. Since correlation coefficients are not normally distributed, they are transformed to Fisher's z values:[40]

$$z = \frac{1}{2}\ln\frac{1+r}{1-r} \tag{3.18}$$

where r is the correlation coefficient as defined above. As a next step, the Bayesian decision theory is applied to the correlation coefficients between tissue samples. The improved two-dimensional correlation coefficient was successfully tested on simulated spectra, yielding high sensitivity and selectivity.

The two-dimensional correlation coefficient mapping has also been applied to jet fuel classification for environmental analysis, comparing the gas chromatograms.[41] The method has proved to be a promising approach to sample classification in environmental analysis.

In analogy to the previous work, Bayes' theorem is applied to the squared difference correlation coefficients (i.e., a combination of correlation coefficient and Euclidean distance) between fluorescence spectra and FT-IR spectra.[8] Bayes' theorem relates the conditional and marginal probabilities of stochastic events A and B:

$$\Pr(A \mid B) = \frac{\Pr(B \mid A)\Pr(A)}{\Pr(B)} \tag{3.19}$$

where $\Pr(A)$ is the prior probability or marginal probability of A. It is "prior" in the sense that it does not take into account any information about B. $\Pr(B)$ is the prior or marginal probability of B, and acts as a normalizing constant. For a given B, $\Pr(A \mid B)$ is the conditional probability of A. It is also called the posterior probability because it is derived from, or depends upon, the specified value of B. Analogously, $\Pr(B \mid A)$ is the conditional probability of B for a given A.

The weighted Pearson product-moment correlation coefficient is used to estimate the similarity of chromatographic fingerprints of traditional Chinese medicine (TCM).[42] In order to control the quality of the injections (containing several extractions) of TCM, a chromatographic fingerprinting technique is used. The content of TCM components is reflected by the peak areas of the chromatographic fingerprint. The similarity measure calculates the difference between the peak areas of two chromatographic fingerprints, since a stable quality of TCM also means a stable area of the peaks. The new similarity measure allows relatively larger differences for large values and smaller differences for small values and is able to discriminate between the relative differences caused by the same absolute differences. The weighted correlation coefficient is expressed as follows:

$$r_w = \frac{\sum_{i=1}^{n} w_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} w_i (x_i - \bar{x})^2 \sum_{i=1}^{n} w_i (y_i - \bar{y})^2}} \qquad 3.20$$

where w_i is the weight, depending on the specific problem. In this case, the optimal weights have the following forms:

$$w_1 = \frac{1}{2}\left(\frac{1}{x_i} + \frac{1}{y_i}\right) = \frac{x_i + y_i}{2x_i y_i} \qquad 3.21$$

$$w_2 = \frac{1}{2}\left(\frac{1}{x_i^2} + \frac{1}{y_i^2}\right) = \frac{x_i^2 + y_i^2}{2x_i^2 y_i^2} \qquad 3.22$$

The results obtained both with real and simulated data sets are more reliable than those using the standard correlation coefficient.

3.1.5 Dot product

The dot product of vectors x and y is defined as follows:

$$x \cdot y = \|x\| \|y\| \cos\theta \qquad 3.15$$

where $\|x\|$ and $\|y\|$ denote the length (magnitude) of vectors x and y, and θ is the angle between them.

The dot product comparison builds a vector in a multidimensional space for each of two vectors being compared and determines the angle between them. Higher angles imply greater differences between the vectors, and angles approaching zero indicate considerable similarity between them.

The dot product (also referred to as “the spectral contrast angle”[43]) has been traditionally used mostly for the comparison of mass spectra, but application to IR spectra is also known.[44]

The dot product is used to calculate similarities among spectra from one- and two-dimensional liquid chromatography experiments.[45] In order to identify peptides from complex protein mixtures, a technique based on liquid chromatography paired with tandem mass spectrometry is applied. Peptide fragment ion spectra contain more peaks than those typically used with the dot product comparison; thus, with a large number of peaks, the algorithm will have a lower level of discrimination. This requires a peak selection process in order to reduce the complexity of spectra prior to comparison. For this purpose, a dot product algorithm was implemented to group uninterpreted peptide tandem mass spectra by similarity. The algorithm infers clusters of similar spectra and can retain a representative from each group while removing duplicates, which improves the speed of peptide identification, at the same time lowering the number of false positives.[45]

Recently, a very similar application of the dot product was presented.[46] Also here, the dot product metric is used to define the similarity between tandem mass spectra. As peptides are often detected in multiple experiments, it is proposed to store their spectra in a reference library once they have been identified. Any new spectra can be looked up in the library using the comparison method. Thus, library search offers a reliable method for the identification of peptide spectra.

The dot product has been used in a so-called “ChromAlign” algorithm developed very recently.[47] It is an efficient two-step procedure for the time alignment of three-dimensional LC-MS chromatographic surfaces. In the first step, the chromatographic profiles (two-dimensional projections of chromatographic surfaces) are pre-aligned. Then, a temporal offset is defined, which maximizes the overlap using the dot product between two chromatographic profiles. In the second step, the algorithm generates correlation matrix elements between full mass scans (the range being indicated by the temporal offset, which speeds up the algorithm) of the reference and sample chromatographic surfaces. Similarity is defined as a path in the correlation matrix that maximizes the sum of the correlation matrix elements. ChromAlign accurately aligns the LC-MS chromatographic surfaces of real samples, e.g., a mixture of known proteins, samples from digests of surface proteins of T-cells, or samples prepared from digests of cerebrospinal fluid.

Stein and Scott[43] reviewed and compared five of the most popular library search identification algorithms proposed for mass spectra and tested by matching test spectra against reference spectra in the NIST/EPA/NIH Mass Spectral Library.[48] The best performance was achieved with the dot product, followed by Euclidean distance and absolute value distance. On the other hand, probability-based matching[49] and Hertz similarity index[50] were less optimal algorithms for mass spectra comparison.

3.1.6 Tanimoto coefficient

The Tanimoto coefficient[51, 52] (or Jaccard index or Jaccard similarity coefficient[53, 54]) is a statistic quantity used for comparing the similarity and diversity of sample sets. The Tanimoto coefficient, $T(A,B)$, is defined as the size of the intersection divided by the size of the union of the sample sets:

$$T(A,B)=\frac{|A\cap B|}{|A\cup B|} \tag{3.23}$$

The Tanimoto distance, $d(A,B)$, which measures dissimilarity between sample sets, is obtained by subtracting the Tanimoto coefficient from 1:

$$d(A,B)=1-T(A,B)=\frac{|A\cup B|-|A\cap B|}{|A\cup B|} \tag{3.24}$$

If objects *A* and *B* are two binary vectors, the Tanimoto coefficient is a useful measure of the overlap that *A* and *B* share with their attributes. In this case, the Tanimoto coefficient has the following, more specific formulation:

$$T(A,B)=\frac{N_{AB}}{N_A+N_B-N_{AB}} \tag{3.25}$$

where N_{AB} represents the number of 1 bits shared by both vectors, while N_A and N_B are the number of 1 bits in vectors *A* and *B*, respectively.

The Tanimoto coefficient is widely used for comparing chemical structures.[28, 31, 32] Already in early online chemical structure search systems, it has been applied to define the similarities between structures.[55] Various molecular descriptors[56] are used to convert the structures into a vector-based representation. Then, the similarity of vectors can easily be calculated with the Tanimoto coefficient. Similarity measures based on molecular fingerprints (binary vectors encoding the presence or absence of substructural fragments within a molecule) and the Tanimoto coefficient are highly effective and they are also extremely efficient. At the same time, they are the most important tools for virtual screening,[57] which is a technique for ranking the molecules in decreasing order of probability of activity in a bioassay of interest. Thus, it ensures that molecules with the highest associated probabilities are tested at the earliest possible stage in a lead-discovery program. However, many other types of similarity measures are possible, but they are more time-consuming.[55]

Willett et al.[58] reviewed and compared various similarity searching techniques utilized with chemical databases: substructure searching, pharmacophore searching, and similarity searching. A number of similarity and distance coefficients are presented, especially those that have gained widespread use in chemical information systems, i.e., Euclidean and Manhattan distances, cosine, Tanimoto and dice coefficients etc. In a very up-to-date research work, the Tanimoto coefficient is stated to be the most suitable method for computing fingerprint-based similarities.[59]

Early-phase virtual screening and compound library design usually involve similarity searching procedures for diversity analysis and the selection of compounds having higher activity.[60] Ligand-based similarity measures are frequently and successfully used for this purpose. In their contribution, Fechner et al.[60] compare the performance of three molecular descriptors and three vector-based similarity measures (Tanimoto coefficient, Manhattan and Euclidean distances) for ligand-based virtual screening. It is concluded that the proper choice of a distance measure critically depends on the descriptor that defines the chemical space.

3.2 Pattern recognition and image analysis

Pattern recognition is a popular chemometric technique used in many applications developed to solve the class-membership problem. It aims at classifying a set of objects (a series of measurements) into several categories (clusters) based on a similarity measure.[1, 61, 62] Classification can be supervised or unsupervised. In supervised pattern recognition, the classification is based on existing patterns that have already been classified. Unsupervised pattern recognition means that there is no a priori information about the classes: The groups are obtained on the fly from the data itself. Unsupervised classification of patterns is commonly called clustering.[1, 61, 62]

Clustering is one of the most important tools in the data mining process, which aims at discovering groups and identifying patterns and important relations in large data sets. After partitioning the given data set into groups, the objects in a cluster are more similar to each other than to those in different clusters.

3.2.1 Clustering

Clustering and classification[63] are the main subdivisions of pattern recognition techniques.[64] With these methods, samples can be classified on the basis of specific properties. Clustering is the classification of objects into different groups or, more exactly, the partitioning of a data set into subsets (clusters) so that the data in each subset share some common properties. Data clustering is a common technique in statistical data analysis, which is used in many fields, including pattern recognition, image analysis, data mining, and machine learning. Statistical classification is a procedure that places individual items into groups based on quantitative information on one or more characteristics and based on a training set of previously labeled items. Statistical classification algorithms are commonly used in pattern recognition systems.

In pattern recognition analysis, each sample is represented as a data vector, where the components correspond to measurements. Therefore, each sample is considered as a point in an n-dimensional measurement space. The distance between these points in the measurement space is inversely related to the similarity between the corresponding samples.

In the beginning of the 1980s, the development of instruments and analytical technologies such as gas chromatography, high-performance liquid chromatography (HPLC), and X-ray fluorescence spectroscopy has considerably increased the number of organic and inorganic compounds identified and quantified in the environment, even at trace level. Huge and complex data sets have been generated with a large number of samples and the corresponding properties and measurements. Extracting information and finding relationships in these multivariate data sets has been a real challenge and has required the development of sophisticated information handling tools. One of these techniques is the principal component analysis (PCA),[14] which was an already existing method, but not yet applied to this kind of huge data bases. Nowadays, PCA is the most widely used multivariate analysis procedure. The original measurement variables are compressed into new variables, called principal components, which are able to describe all the information in the data set. Thus, it is easier to identify key relationships in the data.

Another technique is the cluster analysis, which tries to find the structural characteristics of a data set by dividing the objects into clusters and hierarchies on the basis of their mutual distances.

In an extensive article, these techniques are described in detail, presenting numerous real-life examples and case studies from the field of analytical chemistry.[64]

3.2.2 Classification of clustering techniques

An optimal similarity measure is crucial for clustering, as clusters are formed on the basis of the calculated (dis)similarity between objects. It can be a distance in deterministic clustering or a likelikood in probabilistic clustering. These are called similarity functions. For this purpose, several distance measures are used (see Chapter 3.1).

In a comprehensive review of clustering methods,[65] the clustering algorithms are classified according to different criteria:

- the type of data set,
- the type of theoretical concept on which the clustering analysis techniques are based (e.g., statistics, fuzzy logic),
- the type of distance measure used to define the similarity between data points.

According to the distance measure implemented to define clusters, the clustering methods are usually assigned to one of the three main groups: partitional,[66-68] hierarchical,[69-71] and density-based clustering.[72]

3.2.2.1 Partitional clustering

Partitional clustering methods try to divide the data in a set of clusters by optimizing a given criterion (cost) function in an iterative manner. The most commonly used distance-based algorithm in this category is the so called K-means algorithm,[3, 12, 65] which minimizes the following objective function:

$$E = \sum_{k=1}^{K} \sum_{i \in k} d^2(x_i, m_k) \qquad 3.26$$

where m_k is the center of cluster k and $d(x_i, m_k)$ is the Euclidean distance between point x_i and m_k. Thus, the object function, E, depends on the distance of each point from the center of the cluster to which the point belongs. The K-means algorithm starts with K randomly selected cluster centers. Then, the objects of the dataset are assigned to the clusters whose center is the nearest, and the centers are updated. This is done by iteratively transferring data points between clusters. The process continues until a stop criterion is met: either the threshold for criterion E, or the number of iterations, or the stabilization of the clustering results.

The K-means algorithm is a “hard” clustering, as deterministic similarities are used and the object is assigned to only one cluster. Fuzzy clustering, instead, associates each object to every cluster, using a so-called membership function. The fuzzy c-means algorithm[66-68] is a variant of K-means and includes such a membership function in the object function, E. The membership function, u_{ik}, is given in the following equation:

$$u_{ik} = \frac{1}{\sum_{j=1}^{K} \left(\frac{d(x_i, m_k)}{d(x_i, m_j)} \right)^{\frac{1}{\gamma - 1}}} \qquad 3.27$$

where the fuzziness index, $\gamma > 1$, is 2 in most cases. If this value is close to 1, the membership of the closest cluster will become large.

Fuzzy-logic-based algorithms result in “soft” clustering, as probabilistic membership functions are used. A soft clustering is more general than a hard one, thus, it can be converted to a hard one. Fuzzy clustering algorithms are often used for multivariate image analysis.[3]

The major advantage of the distance-based partitional clustering is the short computation time. Thus, it is very practical for clustering large data sets. The disadvantage of this clustering method is the sensitivity to the initially chosen cluster centers, which may result in a local optimal solution instead of the global one. Another drawback is that the number of clusters has to be given in advance.

3.2.2.2 Hierarchical clustering

Hierarchical clustering methods generate a clustering structure that is additionally visualized with a dendrogram. It is a deterministic clustering method. Two implementation approaches are used, the agglomerative and the divisive one.

Agglomerative clustering starts with assigning each object to an individual cluster. Then, the two closest clusters are merged and this process is repeated until there is one single cluster. The divisive method is the opposite of the agglomerative one.

Based on the definition of the distance between clusters, the agglomerative clustering methods can be of various types: single linkage, complete linkage, average linkage, and Ward's linkage.[73, 74] The distances can be the Euclidean, Manhattan, or, more generally, the Minkowski distances.

In the single linkage method (or the nearest neighbor method), the distance between two clusters is the minimal object-to-object distance, $d(x_i, y_j)$, where objects x_i belong to cluster X and objects y_j belong to cluster Y (for a visualization, see Figure 3.2). This often causes the chaining phenomenon, which is a direct consequence of this method tending to force clusters together due to single entities being close to each other regardless of the positions of other entities in that cluster. Mathematically, the distance, $D(X,Y)$, between two clusters is formulated as follows:[75]

$$D(X,Y) = \min_{x \in X, y \in Y} d(x,y) \qquad 3.28$$

where $d(x,y)$ is the distance between objects x and y.

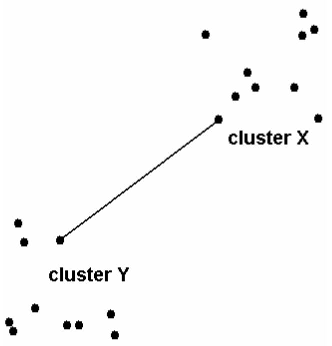


Figure 3.2. Single linkage: The distance between two clusters is the minimal object-to-object distance.

The complete linkage (Figure 3.3) is the opposite of the single linkage. It defines the distance between any two clusters as the maximum distance between the corresponding objects:

$$D(X,Y) = \max_{x \in X, y \in Y} d(x,y) \quad 3.29$$

where, again, $d(x,y)$ is the distance between objects x and y.

This method is not recommended for "noisy" data. It produces very compact clusters and is useful for detecting outliers, since more weight is given to outliers in the cluster decision.

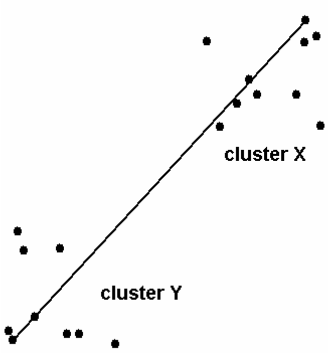


Figure 3.3. Complete linkage: The distance between any two clusters is the maximum distance between the corresponding objects.

The average linkage takes the average distance between all possible pairs (*x,y*) of objects, where *x* and *y* denote objects from clusters *X* and *Y*, respectively. Therefore, computationally it is more expensive than the above-mentioned methods. The effects caused by this method are somewhere between the single and complete linkages: The chaining problem is not observed, and outliers are not favored in the cluster decision.

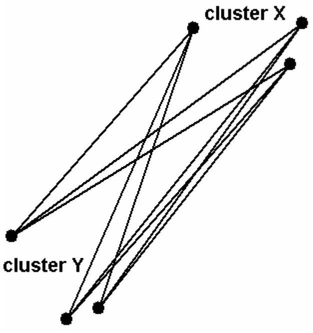


Figure 3.4. Average linkage: The distance is defined as the average distance between all possible pairs of objects.

Mathematically, the average linkage is described by the following expression:

$$D(X,Y) = \frac{1}{n_X n_Y} \sum_{i=1}^{n_X} \sum_{j=1}^{n_Y} d(x_i, y_j) \quad 3.30$$

where $d(x,y)$ is the distance between objects $x \in X$ and $y \in Y$; n_X and n_Y are the numbers of objects in clusters X and Y, respectively.

In Ward's clustering,[73, 74] the distance between two clusters is computed as the increase in the "error sum of squares" (ESS) after merging two clusters into a single one:

$$D(X,Y) = ESS(XY) - [ESS(X) + ESS(Y)] \tag{3.31}$$

where the ESS of a set X of n_X values is the sum of squares of the deviations from the mean value or the mean vector (centroid):

$$ESS(X) = \sum_{i=1}^{n_X} \left(x_i - \frac{1}{n_X} \sum_{j=1}^{n_X} x_j \right)^2 \tag{3.32}$$

Ward's method minimizes the increase in ESS at each step.

The big advantage of the hierarchical clustering is that the dendrogram can be cut at any level in order to obtain the desired number of clusters. Furthermore, the method is not sensitive to outliers, as outliers are clustered in separate groups and do not affect the other clusters. The major disadvantage of the hierarchical clustering is the time requirement. At each merging step, the distances between pairs of objects have to be updated. Thus, for multivariate image analysis, this would be unpractical.

3.2.2.3 Density-based clustering

Unlike the previous cluster categories, the density-based clustering methods are founded on a statistical approach to determine clusters and not on a distance measure. Density-based clustering estimates densities around individual objects. The clusters comprise dense regions of the objects in the data space that are separated by low-density regions. It is basically a hill-climbing procedure to a local density maximum.[72, 76] Density-based clustering was introduced as "mean-shift" method (a simple iterative procedure that shifts each data point to the average of data points in its neighborhood), based on the estimation of a gradient of local density functions.[77]

The major advantage of the method is its ability to detect clusters of arbitrary shapes and to isolate noise. As a drawback must be mentioned that the method needs plenty of time to calculate the density estimation function for each object.

3.2.3 Application of clustering

Clustering techniques have a wide range of applications in chemometrics including exploratory analysis of an environmental data set with different physical and chemical parameters,[69] gene expression profiling using microarrays in order to discover new classes of diseases and to understand their molecular pharmacology,[70] aerosol particle classification based on electron probe X-ray microanalysis,[71] or process monitoring.[67, 78] Another application is the analysis of chemical compounds for combinatorial chemistry,[2] where clustering was studied on a data set of alcohols. Clustering has also been combined with other optimization methods such as genetic algorithms for molecular descriptor selection.[66] In an extensive paper, further fields of applications are listed.[65]

An important and recent application of clustering is the multivariate image analysis.[3, 68, 76] This includes clustering methods applied to magnetic resonance images (MRI) and remote sensing images (of the Earth's surface). A multivariate image can be seen as a stack of images that contains multiple variables (in general, physical units: temperature, gravitational field, impedance, magnetic field, electrical field, mass, wavelength etc.) per pixel, thus providing detailed information in variable and image space.

By including spatial information during the different steps (initialization of clustering parameters, cluster iterations, and post-processing) of classification, the accuracy of clustering is shown to be higher in most cases.[3] Since most of the image data also contains region information (neighboring pixel information), the above-described improvement can easily be applied.

An image analysis method (Euclidean distance mapping; for details, see Section 3.1.1) is used even for investigating microstructural gradients at interfaces in composite materials.[6] The new method is faster and more flexible than conventional dilation-subtraction strip analysis and is not constrained by feature geometry and boundary conditions.

A very similar image analysis technique is applied to the calculation of relative path lengths as a direct measure of tortuosity in compacts.[7] In the imaged data (planes in cubic sodium chloride compacts using scanning electron microscopy), the average path length was calculated with the gray-weighted distance transform (GDT) method.[7] Tortuosity is used to determine the increase in the distance a diffusing molecule travels due to bending and branching of pores and is defined as the ratio of the actual path length through the pores to the Euclidean distance:

$$\tau = \frac{L_{actual}}{L_{Euclidean}} \qquad 3.33$$

in which τ stands for the tortuosity, L_{actual} is the actual path length through the pores, and $L_{Euclidean}$ is the shortest distance between the start and end points in the Euclidean space.

For the first time, a detailed study presents the application of qualitative volatile fingerprints for the identification of, and discrimination between, four *Trichophyton* species (fungal infection of the skin).[5] Principal component analysis[14] was used to find the sensor array system that is able to rapidly differentiate between the four species. Then, this result was double-checked with cluster analysis of the data using Euclidean distance and Ward's linkage.

3.3 Spectra similarity measures: Mass spectra and IR spectra

Since the beginning of the 1970s, numerous methods have been proposed in the mass spectrometric literature for building and searching libraries. Searchable reference libraries of mass spectra are commonly used to identify a new measured spectrum of an unknown substance by doing a similarity search. In mass spectrometry (MS), two searching methods are typically applied, i.e., identity searches and similarity searches.[79-81] In the identity search, it is assumed that the reference library contains the spectrum of the unknown compound, only experimental variability preventing a perfect match between the reference and target spectra. In the more complex similarity search, the library does not contain the spectrum of the unknown compound. In this case, the similarity indices and the list of matches will provide information about the unknown structure.

Stein and Scott[43] reviewed, compared, and tested five of the most popular algorithms proposed for library searching: probability-based matching,[49, 81] Hertz similarity index,[50] Euclidean distance, dot product, and absolute value distance (cf. Section 3.1.5). The test spectra were matched against reference spectra in the NIST/EPA/NIH Mass Spectral Library,[48] which contained 12592 spectra of about 8000 compounds. Furthermore, most algorithms were optimized by tuning their mass weighting and intensity scaling factors. The best performance was achieved with the dot product, followed by Euclidean distance and absolute value distance. Probability-based matching[49] and Hertz similarity index[50] were less optimal algorithms for mass spectra comparison.

Library search and spectrum identification is also addressed.[82] The article is a very good review of the similarity algorithms applied to mass spectra and introduces a new matching algorithm for high resolution mass spectra. The aim of spectrum evaluation can be either the identification of a compound (assuming that a reference spectrum is already available), or the interpretation of spectral data in terms of the unknown chemical structure, or the comparison of spectra of complex mixtures.

Varmuza et al. have developed several library search methods applicable to the identification of organic compounds using infrared (IR) spectra[31] and mass spectra[32] and applied them to sets of thousands of hitlists obtained with different search methods of databases containing 13484 and 106955 compounds for IR and mass spectra, respectively. The similarities of chemical structures between a query compound and the found hitlist entry are characterized by an averaged Tanimoto coefficient,[51] and the similarity criteria for the spectra are based on the correlation coefficient.

3.4 Match probability

Match probability is a special similarity measure developed for identifying IR spectra within a database.[83] A common method of estimating chance match probabilities is match-binning. The target spectrum is associated with a random collection of n objects, usually wavelengths or other peak position measures, from a series of "bins", i.e., wavelength ranges, chosen according to the resolution of the technique. For example, an IR spectrum of 2000 cm^{-1} width can be seen as a set of peaks randomly and independently chosen from 1000 possible positions every 2 cm^{-1}; this corresponds to a peak measurement accuracy of ±1 cm^{-1}. The probability, $P(n,p)$, of obtaining a particular spectrum containing n peaks is estimated by the following formula:

$$P(n,p)=\frac{1}{C_n^p} \qquad 3.34$$

where p denotes the possible positions and C_n^p is the binomial coefficient[84] defined as follows:

$$C_n^p=\binom{p}{n}=\frac{p!}{n!(p-n)!} \qquad 3.35$$

A chance match of two spectra, each with n peaks, has the same probability (see Figure 3.5a). This is the so-called simple binomial model for estimating the match probability. However, in reality, we have the general situation where two spectra with m and q peaks share a subset of n peaks (exemplified in Figure 3.5b). This is easily illustrated by Venn diagrams[14] in Figure 3.6, where the set of N (with a set of size n) is the intersection between the set of M and that of Q chosen at random from P, representing the intersection of the two peak sets (Figure 3.6b). Thus, there is a higher probability of n peaks matching in the two spectra. The probability is given by the hypergeometric distribution:

$$P(n,m,q,p)=\frac{C_n^m\cdot C_{q-n}^{p-m}}{C_q^p} \qquad 3.36$$

It can easily be seen that Equation 3.36 converges to Equation 3.34 if $m=q=n$.

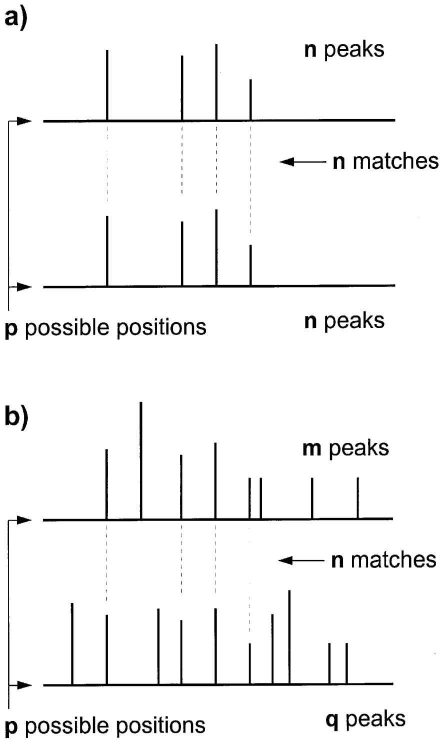


Figure 3.5. Example of peak matching: a) All *n* peaks in both spectra match, with *p* possible line positions; b) *n* common matches for two spectra, one with *m* and the other with *q* peaks, each from *p* possible line positions.

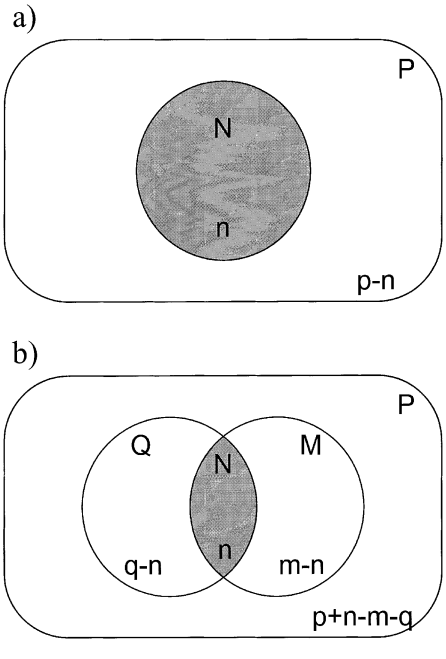


Figure 3.6. Venn diagrams visualizing a) a binomial distribution and b) a hypergeometric distribution. For a given M, there are many possible sets of Q, chosen at random from P, which contain a set of size *n* within M rather than just one set with $\mathrm{Q} = \mathrm{M} = \mathrm{N}$.

According to the tests, the number of hits recorded is far higher than predicted by the simple binomial model, and of broadly similar magnitude to the prediction from Equation 3.35. Thus, it is recommended that for the predictions of spectral matching probabilities, the hypergeometric probability distribution should be used.

3.5 Chromatograms

With the development of high-throughput analytical instruments, an enormous amount of data is generated so that chemometric processing of chromatographic data becomes more and more important and widespread. However, due to drifts in retention times from one run to another, from chromatogram to chromatogram, chromatographic data are not uniformly represented. In order to overcome this problem, two methods are typically used: analysis of the areas of the selected peaks detected in the chromatograms or a peak alignment (warping) of the chromatograms. In the latter case, two sequences of peaks have to be aligned allowing some basic edit operations on the peaks. Peak alignment is necessary in order to include all sampled points of the chromatograms in the data set. Several retention time alignment algorithms have also been developed.[85-89] A new fuzzy warping algorithm[90] uses only a few intense peaks in the individual chromatograms and aligns the signals to the corresponding detected peaks. That is why the method is fast, efficient (in comparison with existing signal alignment algorithms), and identification of the peaks is not required since the correspondence of the peaks is established using the fuzzy matching method. The procedure is iterative, alternating between fuzzy matching and calculation of transform parameters. It has been successfully applied to the peak alignment of urine NMR spectra.[91]

The chromatographic alignment problem very often occurs in proteomics studies. In research involving chromatography and mass spectrometry, two variables may need alignment: time and mass. In mass spectrometry proteomics, usually outcomes of a single analysis are investigated. Nevertheless, if the raw data were compared across multiple experiments, both peptide and protein identification and quantification would be much more accurate. If peptide quantities are compared between experiments, then, chromatographic alignment of MS signals is required in the absence of convincing tandem MS identifications.[36]

In the above publication, an extended version of a dynamic time warping algorithm is introduced and used for alignment of chromatography mass spectrometry of a proteomics data set.[36] Another up-to-date contribution introduces an efficient two-step procedure, which is used for the alignment of three-dimensional liquid chromatography-mass spectrometry (LC-MS) chromatographic surfaces.[47] In an earlier research, liquid chromatography is paired with tandem MS in order to identify peptides from complex protein mixtures.[45]

Chromatograms and chromatographic processing are involved in several other applications. For example, the similarity of the high-performance liquid chromatographic fingerprints of traditional Chinese medicine is estimated.[42] In another publication, gas chromatograms are compared in order to classify jet fuels for environmental analysis.[41] A review article in the Encyclopedia of Analytical Chemistry presents various case studies concerning the clustering of chromatograms.[64]

3.6 Metabolomics and metabonomics: ^{1}H NMR of biological fluids

In the last decade, the scientific field of metabolomics was born. A new range of analytical methodologies and technologies were introduced related to the analysis of microbial, plant, and animal metabolomes and, more generally, of functional genomics. Metabolomes are the qualitative and quantitative collections of all low molecular weight compounds (metabolites) in a cell that are taking part in general metabolic reactions. In a review,[92] the terms of metabolomics and metabonomics are explained, the new methods and techniques are listed, and several applications are outlined.

Metabolomics means the non-biased identification and quantification of all metabolites in a biological system. Sample preparation must be so as not to exclude metabolites, and selectivity and sensitivity of the analytical technique must be high.[92] Metabonomics, on the other hand, is defined as the evaluation of tissues and biological fluids for changes in endogenous metabolite levels that result from disease or therapeutic treatments.[93]

In metabolomics (also termed metabonomics for NMR-based clinical applications), NMR spectroscopy provides a rapid, nondestructive, high-throughput method that requires minimal sample preparation.[93] Data is usually presented in a table where each row relates to a given sample or analytical experiment and each column corresponds to a single measurement in that experiment, typically individual spectral peak intensities or metabolite concentrations. Normally, the following typical steps are used in the analysis of metabonomics data:[94]

1. Post-instrument processing, i.e., polynomial baseline correction and calculation of intensity values either on each data point, on each peak, or summed over segmented regions (binning);
2. Generation of a data table from the analytical measurements such that there are *m* rows (observations, samples) and *n* columns (variables, frequencies, integrals);
3. Normalization of the data (a row operation);
4. Scaling of the data (a column operation);

5. Multivariate statistical modeling of the data.

In order to reduce the set of parameters, the NMR data is first compressed using the binning method, in which the normalized relative integrals are taken in each segment having a typical width of 0.04 ppm. As a drawback, it can be stated that the binning does not work well in crowded regions of spectra with substantial peak overlap and is not easily automated for application to large sample sets. Then, the PCA[14] or partial least squares (PLS)[63] analysis is used as a second compression step.[95]

The purpose of normalization is to identify and remove systematic variation, i.e., to remove or minimize the effects of variable dilution of the samples.

Scaling is performed on the columns of data using several methods. The most common method is the mean-centering, in which the column mean is subtracted from each value in the column. Thus, components found by PCA will have the centroid of the data as their origin. The column will have unit variance if each value of the column is divided by the standard deviation of the column values. Another popular scaling is the Pareto scaling, where each variable is divided by the square root of the standard deviation of the column values. It is concluded that data preprocessing is context-dependent and will have to be adapted in an appropriate manner according to the study and type of sample.

Metabonomics is a very popular research field with numerous applications. Strategies such as metabonomics (small molecule metabolite effects), genomics, transcriptomics (changes in gene expressions), and proteomics (changes in protein levels) are more often applied in the pharmaceutical industry, especially to measure the toxicological response.[96] However, drug-induced response can only be defined accurately if the normal degree of psychological variation in the absence of stimuli is characterized. More precisely, the effects of gender, age, diet, species, hormonal status, and stress on the metabolic composition of urine are investigated. In order to evaluate normal psychological variations, pattern recognition techniques are used for the comparison of urine NMR spectra over a given time course.[96]

In an earlier application, biomarkers of renal toxicity are identified, using data reduction techniques and PCA of proton NMR spectra of urine.[97] The PCA is also applied to a metabonomic study of mouse urine.[98]

The NMR spectra of biofluids are very complex and their signals highly overlap, making it very difficult to accurately compare this type of spectra. Thus, usually, compression methods such as binning, PCA, and PLS are applied. As an alternative procedure, a peak alignment of the spectra with the fuzzy warping technique is used[91] (for more details, see Section 3.5).

3.7 New similarity methods using shifts

The similarity measures presented in Section 3.1 do not include any information about the neighborhood; this is why they are commonly called pointwise similarity criteria.

While in the case of mass and IR spectra the classical distance metrics are capable of accurately calculating the similarities, this does not hold for NMR spectra. In addition, IR spectra can be efficiently compared with the correlation coefficient method, and mass spectra (after linear transformation) by using the dot product distance. The major problem with ^{1}H NMR spectra is that the signals are very narrow, and shifts as big as 0.2 ppm[99, 100] are expected due to changes in the experimental conditions. These shifts are too large for applying the correlation coefficient method, while the dot product cannot efficiently discriminate between spectra containing a large number of signals.

In the following section, several methods are presented that detect similarities between NMR spectra taking into account the chemical shifts. They are also referred to as similarity criteria including the neighborhoods. The idea is to compare a signal of one spectrum with the environment of the corresponding signal of another spectrum and vice versa.

3.7.1 Binning

One possible solution to the aforesaid problem is to artificially increase the line widths before further processing the ^{1}H NMR data. For example, Kalelkar et al.,[101] by means of a moving average filter, reduce the number of data points from 16384 to 820 (i.e., lowering the digital resolution from 0.46 to ca. 5 Hz). After compression, the self-organizing maps (Kohonen networks) are used to develop an automated and rapid analysis of the ^{1}H NMR spectra.

In another approach often applied in metabonomic studies, the NMR data is first compressed using a binning method, in which the normalized relative integrals having a typical width of 0.04 ppm are taken in each segment. Then, the PCA[14] or PLS analysis[63] is used as a second compression step (see Section 3.6).

3.7.2 The spectral fold

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

Karfunkel et al.,[102] is the so-called fold, a smooth similarity measure. It 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.[103]). For a graphical visualization of the weighting function, see Figure 3.7, where the black region represents the diagonal elements of 1 and the gray regions, the continuously decreasing values. The fold of spectra x and y (same resolution, defined over the same range and normalized to the same value) is mathematically represented by the equation:

$$S = (x-y)^T \mathbf{F}(x-y) \qquad 3.37$$

where the elements of **F** are:

$$\mathbf{F}_{ij} = \frac{1}{1+\alpha\left|i-j\right|^{\beta}} \qquad 3.38$$

with the coefficient $\alpha > 0$ and the integer exponent β being adjustable empirical parameters. Figure 3.8 shows how the similarity of two sample binary vectors is calculated, using the fold criterion with an **F** weighting matrix having a form as defined above.

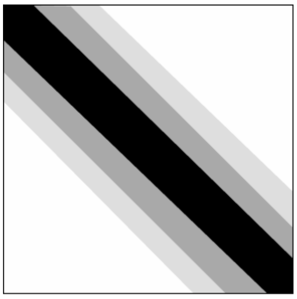


Figure 3.7. Graphical visualization of a weighting matrix **F** with diagonal elements of 1 (black region) and values continuously decreasing with increasing distance from the diagonal (the lighter gray regions symbolize the lower values) used in the spectral fold defined by Karfunkel et al.[102]

$$\begin{bmatrix}1 & 0 & 0 & 1\end{bmatrix}\begin{bmatrix}1\\0\\0\\1\end{bmatrix} = 2 \qquad \begin{bmatrix}1 & 0 & 0 & 1\end{bmatrix}\begin{bmatrix}0\\1\\1\\0\end{bmatrix} = 0$$

$$\begin{bmatrix}1 & 0 & 0 & 1\end{bmatrix}\begin{bmatrix}\mathbf{1} & 1 & 0 & 0\\1 & \mathbf{1} & 1 & 0\\0 & 1 & \mathbf{1} & 1\\0 & 0 & 1 & \mathbf{1}\end{bmatrix}\begin{bmatrix}0\\1\\1\\0\end{bmatrix} = \begin{bmatrix}1 & 0 & 0 & 1\end{bmatrix}\begin{bmatrix}1\\2\\2\\1\end{bmatrix} = 2$$

Figure 3.8. Calculating the similarity of two simple binary vectors, using the fold criterion with an **F** weighting matrix having a form as defined above.

3.7.3 Weighted cross-correlation method

More recently, de Gelder et al.[104] have shown that various similarity criteria of two functions, *f*(*x*) and *g*(*x*), including the sum of squared differences,[105] the correlation coefficient, and the overlap integral,[106] are related to the cross-correlation function, $c_{fg}(\Delta)$, at $\Delta = 0$:

$$c_{fg}(\Delta) = \int f(x)g(x+\Delta)dx \tag{3.39}$$

where Δ is the relative shift between the two functions, $f(x)$ and $g(x)$.

Thus, they cannot provide any information about patterns that are shifted relative to each other. While the integral over Δ 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, $w(\Delta)$) normalized with the product of the two weighted autocorrelation functions, c_{ff} and c_{gg}:

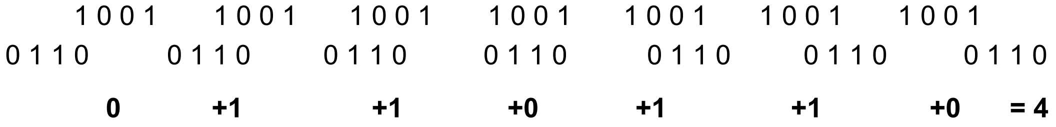
$$S_{fg} = \frac{\int w(\Delta)c_{fg}(\Delta)\,d\Delta}{\sqrt{\int w(\Delta)c_{ff}(\Delta)\,d\Delta \int w(\Delta)c_{gg}(\Delta)\,d\Delta}} \tag{3.40}$$

where $c_{ff}(\Delta)$ and $c_{gg}(\Delta)$ are defined in analogy to Equation 3.39:

$$c_{ff}(\Delta) = \int f(x)f(x+\Delta)dx \tag{3.41}$$

$$c_{gg}(\Delta) = \int g(x)g(x+\Delta)dx \tag{3.42}$$

In their studies, de Gelder et al.[104] used a triangular weighting function of width l defined as $w(\Delta) = 1 - |\Delta|/l$ if $|\Delta| < l$, and $w(\Delta) = 0$ if $|\Delta| \geq l$. Figure 3.9 demonstrates with two simple bit vectors the weighted cross-correlation method and the spectral fold.



$$\begin{bmatrix}1&0&0&1\end{bmatrix}\begin{bmatrix}1&1&1&1\\1&1&1&1\\1&1&1&1\\1&1&1&1\end{bmatrix}\begin{bmatrix}0\\1\\1\\0\end{bmatrix} = \begin{bmatrix}1&0&0&1\end{bmatrix}\begin{bmatrix}2\\2\\2\\2\end{bmatrix} = 4$$

Figure 3.9. Comparing the weighted cross-correlation (top) method and the spectral fold (bottom) using two simple bit vectors. In both cases the weighting is equal to 1.

3.8 Comparison of similarity criteria

In order to compare the most important similarity criteria presented above, we use simple simulated ^{1}H NMR spectra as shown in Figure 3.10. All three spectra have the same digital resolution of 16 K digital points, defined on the same range of 0–10 ppm, and the peak intensities are normalized to the range of 0–1. Each spectrum contains a single shifted signal, i.e., at 1.0 ppm (spectrum A), 1.1 ppm (spectrum B), and 9.0 ppm (spectrum C). It is obvious that spectra A and B with a shift of only 0.1 ppm are much more similar than spectra A and C with a shift of 8 ppm.

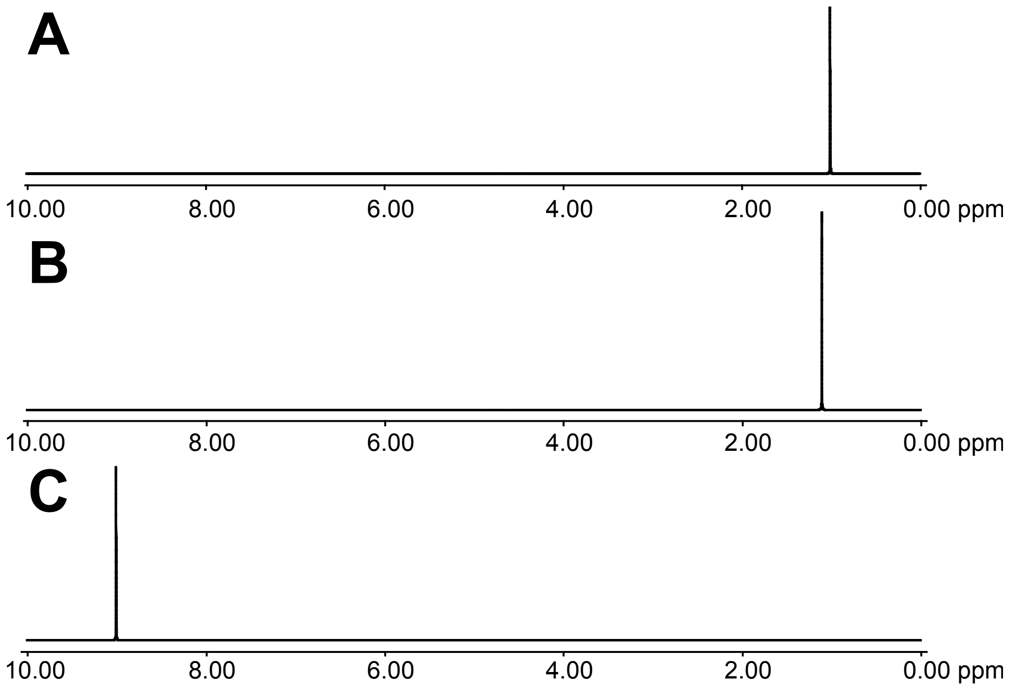


Figure 3.10. Simulated ^{1}H NMR spectra, each containing a single shifted signal: at 1.0 ppm (A), 1.1 ppm (B), and 9.0 ppm (C).

For testing purposes, two types of line widths are used: 1.8 Hz and 60 Hz. Table 3.1 lists the (dis)similarity values between spectra A and B as well as spectra A and C, using seven different criteria and the above two line widths. It is observed that the classical pointwise similarity measures are unable to deal even with a minor shift and even line broadening is not of much help. On the other hand, the similarity measures using the neighborhood can easily tackle the problem of minor shifts. The weighted cross-correlation method was performed with triangle weighting and a 1.4 ppm cut-off range, while the bin method was used with a minimal bin width of 0.4 ppm (for more details on the bin method, see Chapter 4).

Table 3.1. Comparison of various (dis)similarity criteria by different calculation methods applied to the ^{1}H NMR spectra shown in Figure 3.10.

(Dis)similarity measures	Line width, 1.8 Hz		Line width, 60 Hz	
Euclidean distance	$d_{AB} = 3.0687$	$d_{AC} = 3.0699$	$d_{AB} = 17.4259$	$d_{AC} = 19.3361$
City-block distance	$d_{AB} = 18.4107$	$d_{AC} = 18.6307$	$d_{AB} = 325.1653$	$d_{AC} = 397.0950$
Correlation coefficient	$r_{AB} = 0.0003$	$r_{AC} = -0.0011$	$r_{AB} = 0.1771$	$r_{AC} = -0.0131$
Dot product	$\cos\theta_{AB} = 0.0007$	$\cos\theta_{AC} = 0$	$\cos\theta_{AB} = 0.1878$	$\cos\theta_{AC} = 0$
Tanimoto coefficient	$T_{AB} = 0$	$T_{AC} = 0$	$T_{AB} = 0.0914$	$T_{AC} = 0$
WCC method[a]	$S_{AB} = 0.8668$	$S_{AC} = 0$	$S_{AB} = 0.9092$	$S_{AC} = 0$
Bin method	$S_{AB} = 0.9999$	$S_{AC} = 0.04$	$S_{AB} = 1$	$S_{AC} = 0.04$

[a] Weighted cross-correlation method

3.9 Performance metrics

Performance metrics are used to measure the performance of a computational method, the optimality of a procedure, or the discriminating level of a similarity criterion. Performance measurements can be considered as the complement of the corresponding error metric since the error can be seen as a negative performance.

The most important and popular performance metrics used in this work are the contingency tables and histograms.

3.9.1 Contingency tables

Contingency tables,[14, 63] a term first used by Pearson,[107, 108] are statistical tools used to verify and analyze the relationship between two or more variables, in most cases categorical variables. They show the responses of subjects to one variable as a function of another variable.

For a better understanding, here is an example. Suppose that we have two variables, i.e., gender (male or female) and handedness (right- or left-handed). We observe the values of both variables in a random sample of 100 people. Then, a contingency table as in Table 3.2 can be used to express the relationship between these two variables. Evidently, the proportion of right-handed men is about the same as that of women. However, the two proportions are not identical, and the statistical significance of the relationship between rows and columns can be tested with Pearson's chi-square test[63, 109, 110] or Fisher's exact test,[14, 63, 111] provided that the entries in the table represent random samples. If the proportions of individuals in the different columns vary between rows (and vice versa), the table is said to show contingency between the two variables. If there is no contingency, then, the two variables are independent. Pearson's chi-square model tests the null hypothesis that the relative frequencies of occurrence of observed events follow a specified frequency distribution. The events are assumed to be independent and have the same distribution, and the outcome of each event must be mutually exclusive. A simple example is the hypothesis that an ordinary six-sided dice is "fair", i.e., all six outcomes occur equally often. Fisher's exact test is used to determine if there are nonrandom associations between two categorical variables when sample sizes are small.

Table 3.2. A simple contingency table used to express the relationship between two variables, i.e., sex (male or female) and handedness (right-handed or left-handed). The values were observed for both variables in a random sample of 100 people.

	Right-handed	**Left-handed**	**Total**
Male	43	9	52
Female	44	4	48
Total	87	13	100

In order to evaluate the performance of various similarity criteria, we use the contingency table presented in Table 3.3. The totals in the right-most column and the bottom row ($tp + fp$, $fn + tn$, $tp + fn$, and $fp + tn$) are called marginal totals and the total, *N*, in the bottom right corner is the grand total. This is the simplest kind of contingency table, in which each variable has only two levels; this is called a 2×2 contingency table or 2-way contingency table. The relationship between ordinal variables, or between ordinal and categorical variables, may also be represented in contingency tables, though this is less often done since the distributions of ordinal variables can be summarized efficiently by the median.

In order to verify the quality of a measured spectrum, it will be compared with the predicted spectrum of the corresponding structure. Furthermore, if we want to test the discrimination level of a similarity measure, the measured spectra will be compared with the predicted spectrum of a foreign structure. Thus, in our case (Table 3.3), if an original spectrum is compared with the predicted spectrum of the corresponding structure, then, we have an OK real situation. In this case, if the similarity value is above a given threshold value, then, we count it as a true positive (*tp*), otherwise it will be considered as a false negative (*fn*) one. If two foreign spectra are being compared, then, we have a Not OK real situation. If the outcome is negative, i.e., if we get a similarity value below the threshold, then, we have found a true negative (*tn*) case, otherwise it will be counted as a false positive (*fp*) one.

Table 3.3. Contingency table applied to evaluate the performance of various similarity criteria.

Outcome	**Real situation**		
	OK	**Not OK**	**Total**
Positive	true positive (*tp*)	false positive (*fp*)	*tp* + *fp*
Negative	false negative (*fn*)	true negative (*tn*)	*fn* + *tn*
Total	*tp* + *fn*	*fp* + *tn*	*N*

Using the elements of the contingency table, the following four performance parameters are calculated.[14] They are useful for comparing contingency tables and for measuring their performances.

Sensitivity is defined as the proportion of true positives with respect to the total number of actual positives:

$$\text{Sensitivity} = \frac{\text{Number of true positives}}{\text{Total number of actual positives}} = \frac{tp}{tp + fn} \qquad 3.43$$

Specificity is defined as the proportion of true negatives with respect to the total number of actual negatives:

$$\text{Specificity} = \frac{\text{Number of true negatives}}{\text{Total number of actual negatives}} = \frac{tn}{tn + fp} \qquad 3.44$$

The performance of a screening assay can be measured by means of its positive predictive value (*PPV*), which is an estimate of the proportion of true positives with respect to the total number of positives:

$$PPV = \frac{\text{Number of resulting true positives}}{\text{Total number of resulting positives}} = \frac{tp}{tp + fp} \qquad 3.45$$

In a similar way, the negative predictive value (*NPV*) is defined as:

$$NPV = \frac{\text{Number of resulting true negatives}}{\text{Total number of resulting negatives}} = \frac{tn}{tn + fn} \qquad 3.46$$

Contingency tables are very useful when defining thresholds, selecting the best parameter settings, and calculating the selectivity of a method. They are also used as a statistical technique for data analysis. For example, two-way and three-way contingency tables serve to define and simulate quantum random walks.[112] Two-way contingency tables have been applied to distinguish between two variables,[113] the pair correlation method being generalized for variable selection. This method has been employed to choose between two correlated predictor variables. Recently, it has also been used to evaluate the impacts of groundwater abstraction on spring flow and base flow.[114]

3.9.2 Histograms

In statistics, a histogram (defined for the first time by Lancaster[115]) is a graphical display of tabulated frequencies. It is the graphical version of a table and shows which proportion of cases fall into each of several or many specified categories. It differs from a bar chart in that it is the area of the bar that denotes the value, not just the height. This is a crucial distinction when the categories are not of uniform width. Figure 3.11 shows an example of a histogram of 100 normally distributed random values.

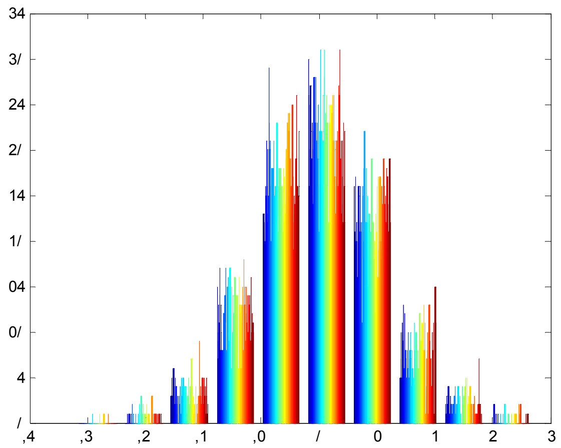


Figure 3.11. Example of a histogram of 100 normally distributed random values.

The histogram is one of the seven basic tools of quality control, which also include the Pareto chart (a special type of bar chart where the values being plotted are arranged in descending order; often represents the most common sources of defects), check sheet (a simple document used for collecting data in real-time and at the location where the data is generated), control chart (or process-behavior chart, which is a run chart of a sequence of quantitative data that helps to assess the nature of variation in a process and to facilitate forecasting and management), cause-and-effect diagram (reveals the key relationship among different variables and possible causes and provides additional insight into process behavior), flowchart (schematic representation of an algorithm or a process), and scatter diagram (or scatter plot, which is a graph used in statistics to visually display and relate two quantitative variables of a multidimensional data set by displaying the data as a collection of points).

Mathematically, a histogram is a mapping, m_i, which counts the number of observations that fall into various disjoint categories (known as bins). If n is the total number of observations and k the total number of bins, the histogram, m_i, meets the following conditions:

$$n = \sum_{i=1}^{k} m_i \tag{3.47}$$

Histograms are widely used in scientific works. They are an indispensable tool for visualizing large amounts of data and for extracting useful information. They have been employed in a number of applications: quantitative histogram analysis of images,[116] photon counting histogram analysis,[117] quantification of reaction coordinate error, i.e., the degree of freedom that quantifies the dynamical progress along a reaction pathway,[118] weighted histogram analysis method for investigating the performance of simulated and parallel tempering algorithms,[119] histogram analysis of the intensity fluctuations,[120] and histogram of apparent diffusion coefficient values.[121]

4 A novel spectra similarity measure: the bin method

In the previous chapter, it was shown that an optimal similarity measure is crucial for clustering, as clusters are formed on the basis of the calculated (dis)similarity between objects. The situation is the same when automatic spectra verification is performed, i.e., when spectra are classified on the basis of their quality and correctness. The task is even more challenging because the verification, i.e., comparison of ^{1}H NMR spectra is required. In Chapter 3, it was observed that some of the classical methods for vector-based similarity are applied efficiently to mass, IR, and UV spectra, but not to ^{1}H NMR spectra. Thus, the aim was to develop a new similarity measure that is selective and capable of discriminating between related and foreign ^{1}H NMR spectra.

In this chapter, a new method[122, 123] is described for calculating the similarity degree of two spectra. Its performance is optimized with similar, computer-generated ^{1}H NMR spectra. The novel method is compared with a recently proposed local cross-correlation method.[104] 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. Due to the generality of the approach, the novel procedure can be easily adapted and applied to comparing other spectra or patterns as well.

4.1 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 the ^{1}H NMR spectra of submicrogram samples on the order of minutes.[124-126] 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 and some of them have been used to detect similarities in spectra. All these methods are presented in Chapter 3. However, these approaches only work with spectra showing relatively broad signals since they fail to detect similarities if the positions of the signals in the two spectra differ by more than their widths. In other words, most of these measure do not detect any information about the neighborhood of a signal. Such similarity measures would not do when comparing NMR spectra because changes in signal positions of up to ca. 100 times the signal half widths are to be expected even in closely related spectra.

In this chapter, we introduce a novel method to quantify spectra similarity and compare its performance with the generalized similarity measure by de Gelder et al.[104] (presented in Section 3.7.3).

4.2 Test sets

The first test set consisted of estimated ^{1}H NMR spectra of ten arbitrarily chosen compounds (Figure 4.1). For each spectrum, two other spectra were calculated by randomly shifting signal groups using a normal distribution with a standard deviation of 0.2 or 0.4 ppm (an example is shown in Figure 4.3).

The second test set consisted of 1146 ^{1}H NMR spectra derived from a library of Chemical Concepts.[127] The ^{1}H NMR spectra are stored in JCAMP-DX format,[128-130] while the corresponding molecular structures are organized in molfiles.[131, 132] From the 5003 compounds, those database entries were selected for which the NMRPrediction 3.0 program[133] 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 (DMSO-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 $n = \overline{1, N}$ 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 completing 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 remainder corresponded to a spectral range < 0.1 ppm on the left side of the spectrum.

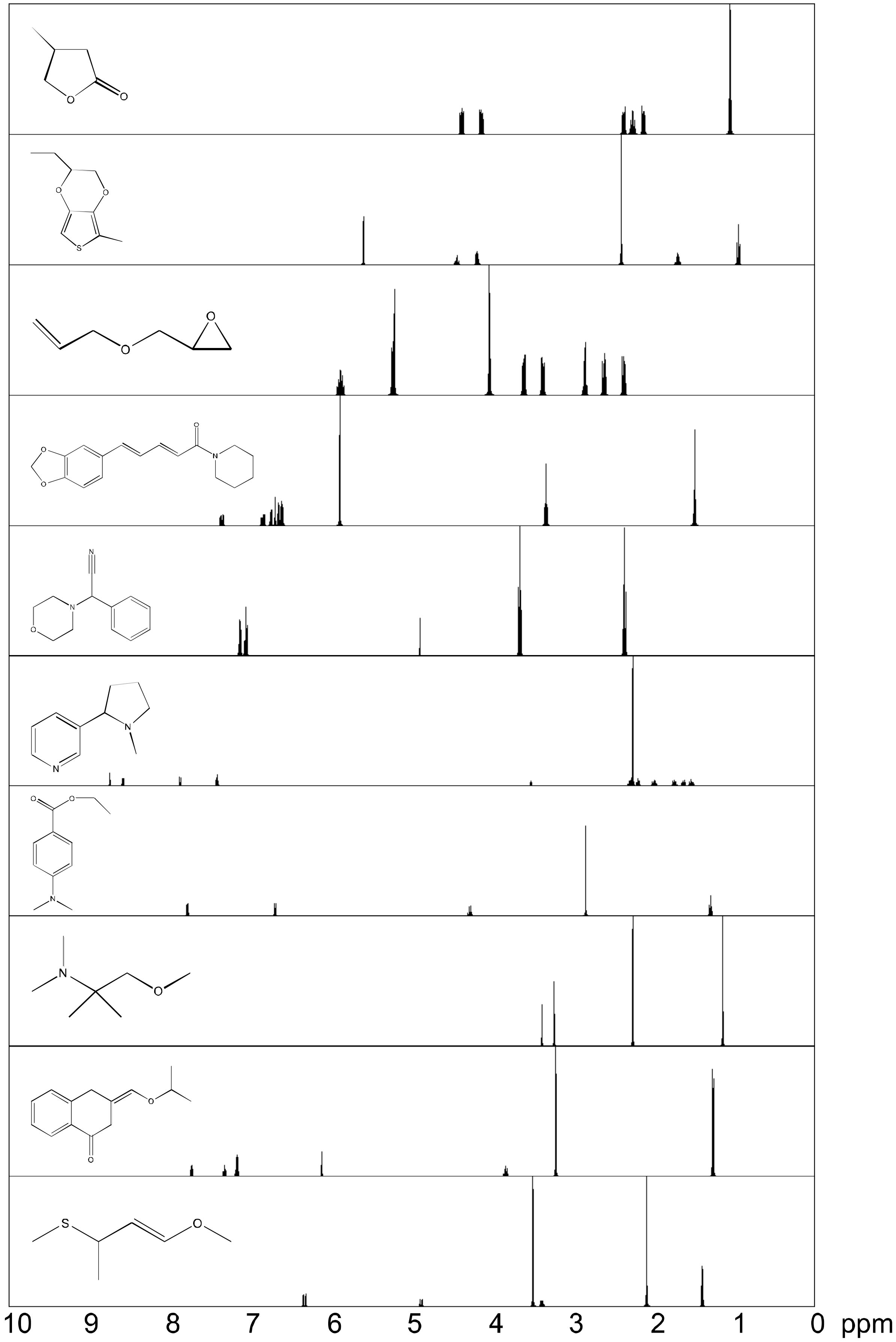


Figure 4.1. Computer-generated ^{1}H NMR spectra of ten arbitrarily chosen structures of organic compounds used as a test set (16 K digital points, 10 ppm range, 500 MHz).

The algorithms (input/output modules, similarity criteria) were implemented in Borland Delphi 5.0 programming environment.[134] The tests were performed on a Windows PC with Intel Pentium 4 2.8 GHz CPU and 512 MB RAM.

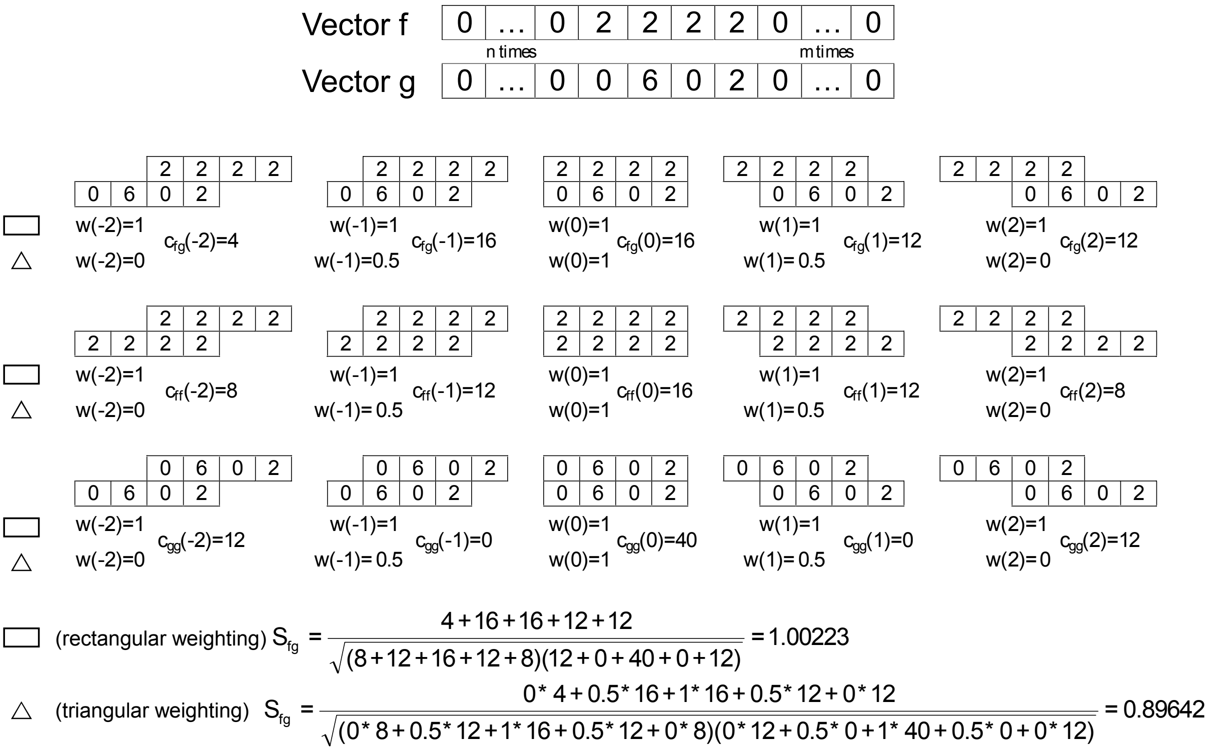
Computing times for comparing one pair of spectra, including the spectra prediction and elimination of noise and solvent signals, were on the order of 0.5 s.

4.3 Limitations of weighted cross-correlation method

The weighted cross-correlation method introduced by de Gelder et al.[104] has been presented in Section 3.7.3. Here, we investigate the performance of this method and show some of its limitations.

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 3.0 software[133] 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 on the order of 0.2 ppm.[99, 100] Similar variations in the signal positions are also expected between spectra measured in different solvents, e.g., in $CDCl_3$ and DMSO-d_6.[135]

Preliminary comparisons of measured and estimated spectra were performed with the cross-correlation method by de Gelder et al.,[104] using triangle and rectangle weighting functions, $w(\Delta)$. Surprisingly, in a series of cases, the rectangle, but not the triangle, as weighting function gave similarity values, $S_{fg} > 1$ (cf. Equation 3.40). Indeed, there is no mathematical reason why the weighted integral of the cross-correlation function should always be smaller than, or equal to, the geometric mean of the corresponding weighted integrals of the autocorrelation functions. This is illustrated by the two simple vectors in Figure 4.2. 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 Figure 4.6 in Section 4.6). For these reasons, different other similarity measures were tested. They included artificial line broadening of up to 20 Hz[101] 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 method is described in the following section.



$$S_{fg} = \frac{4+16+16+12+12}{\sqrt{(8+12+16+12+8)(12+0+40+0+12)}} = 1.00223$$

$$S_{fg} = \frac{0*4+0.5*16+1*16+0.5*12+0*12}{\sqrt{(0*8+0.5*12+1*16+0.5*12+0*8)(0*12+0.5*0+1*40+0.5*0+0*12)}} = 0.89642$$

Figure 4.2. Two simple vectors whose comparison by the cross-correlation method[104] 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 0.89642.

4.4 Novel similarity criterion

The novel similarity criterion[123] of two spectra x and y is related to the binning method applied in metabonomic studies (see Section 3.6). 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, 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 $n = \overline{1, N}$, where *N* corresponds to the maximal number of bins. For each division, the similarity index, SI_n, is calculated according to Equation 4.1:

$$SI_n = \frac{I_{xy}(n)}{I_x + I_y - I_{xy}(n)} \qquad 4.1$$

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

$$I_{xy}(n) = \sum_{i=1}^{n} \min\left(I_x(i), I_y(i)\right) \qquad 4.2$$

with $I_x(i)$ and $I_y(i)$ as the integrated intensities of the respective spectra within bin *i*. The procedure is illustrated in Figure 4.3 by two estimated spectra of piperine (cf. Figure 4.1, 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 Figure 4.3, $I_{xy}(2) = 19.000$, $I_{xy}(3) = 16.258$, and $I_{xy}(4) = 17.992$, from which one obtains $SI_2 = 1.000$, $SI_3 = 0.747$, and $SI_4 = 0.899$, respectively. Using the above example, Figure 4.4 gives the SI_n values for $n = \overline{1,50}$ bins.

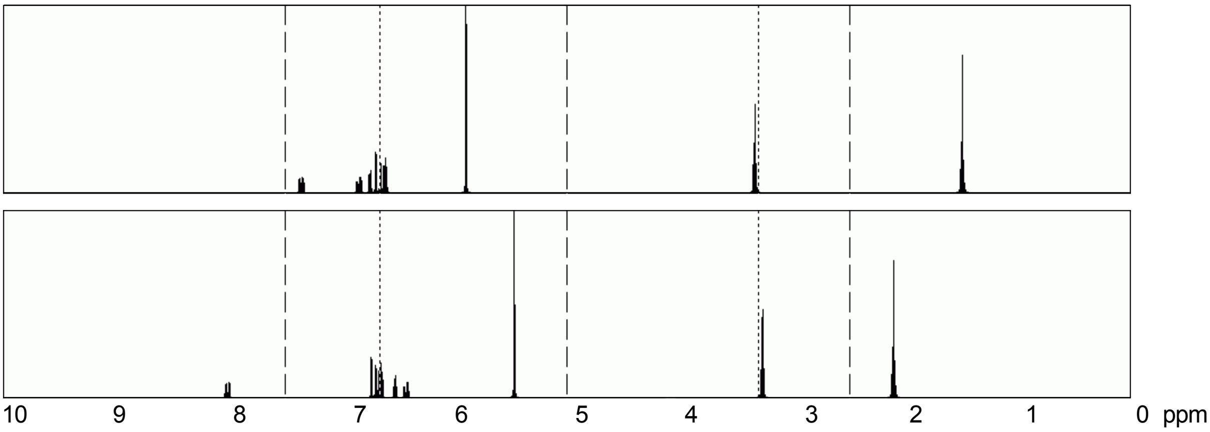


Figure 4.3. Predicted spectrum (top) of piperine (cf. Figure 4.1, entry no. 4) and the corresponding spectrum (bottom) with randomly shifted multiplets ($\mathrm{SD} = 0.4\,\mathrm{ppm}$) divided into 2 (central dashed line), 3 (dotted lines), and 4 (three dashed lines) bins. It can be observed that a finer division (i.e., more bins) may yield a higher similarity index than a coarser one.

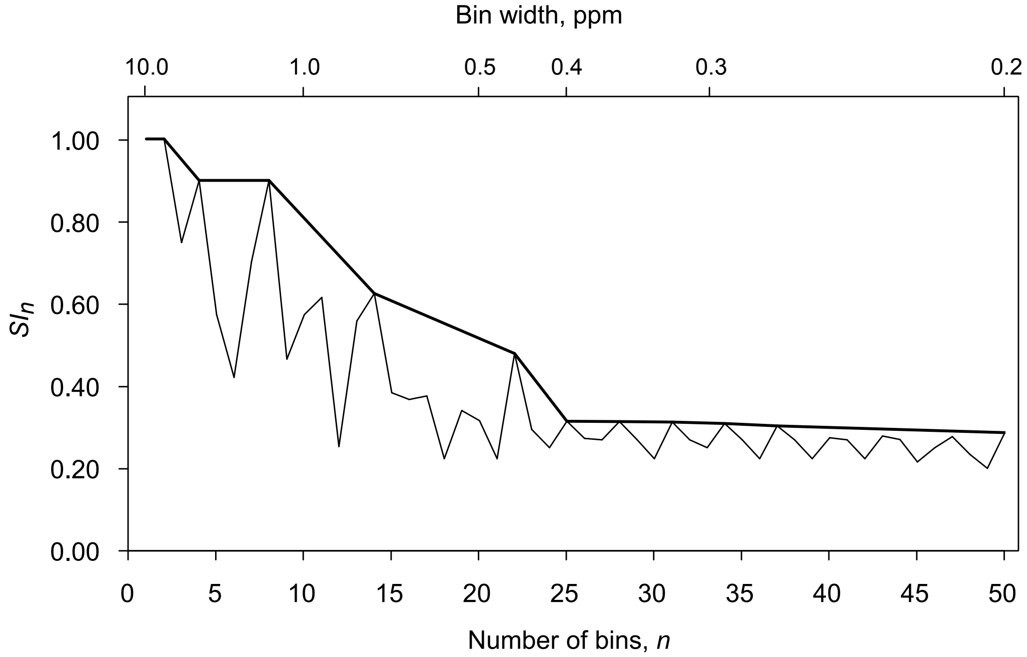


Figure 4.4. Thin line: Similarity index, SI_n, calculated by Equation 4.1 using divisions of up to 50 bins (minimal bin width, 0.2 ppm) for comparing the predicted spectrum of piperine (cf. Figure 4.3, top) with the corresponding one having randomly shifted signal groups (normal distribution, SD = 0.4 ppm). Thick line: The negative effect of the oscillating values of the similarity index is compensated by connecting the remaining global maxima of SI_n.

Apparently, it may happen that a finer division (i.e., more bins) provides a higher value of the similarity index than a coarser one. The non-monotonous changes in the SI_n values (thin line in Figure 4.4) 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 Figure 4.3). To reduce the influence of such artifacts, the overall similarity, *S*, is defined according to Equation 4.3 as the normalized integral of the function, SI_n^*, connecting the remaining global maxima (thick line in Figure 4.4) rather than the average of the SI_n values:

$$S = \frac{1}{N}\sum_{n=1}^{N} SI_n^* \qquad 4.3$$

where

$$SI_n^* = \max\left(SI_n, \frac{SI_a(n-b) - SI_b(n-a)}{a-b} \right) \qquad 4.4$$

with

$$SI_a = SI_{n-1}^* \quad \text{and} \quad SI_1^* = 1 \qquad 4.5$$

$$SI_b = \left\{ \max SI_i \middle| i = \overline{n, N} \right\} \qquad 4.6$$

According to Equations 4.1 and 4.3, 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 (see Section 3.1.6) and does not reward similarities due to the absence of signals (as, e.g., does the correlation coefficient). The only parameter to be selected 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 Figure 4.4, a too fine division, i.e., one that leads to windows smaller than the expected differences in signal positions ($\mathrm{SD} = 0.4$ 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 Section 4.2), a minimal bin width of 0.4 ppm has proven to be optimal. This nicely fits the expectation of tolerable deviations between predicted and observed chemical shifts 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, i.e., information provided by spin coupling, do not influence the result.

4.5 Overlapping bins

A further idea is to use overlapping bins. This could be efficient in discriminating better between spectra with a large number of signals. Thus, some regions (a given percentage) of the spectra are shared by two bins. The overlaps are evenly distributed among the bins, i.e., for each bin, the same percentage overlaps on both sides (so that the total overlap is equal to the input value), except for the bins at the edges of the spectrum, where only one side overlaps.

Figure 4.5 exemplifies the application of overlapping bins. The artificial (predicted) ^{1}H NMR spectrum shown in Figure 4.3 (top) is used to demonstrate this method. The corresponding spectrum with randomly shifted multiplets is the same as in Figure 4.3 (bottom), but is not shown in Figure 4.5. Moreover, the same divisions are applied also to this spectrum. In Figure 4.5 with the artificial spectrum, three different overlap sizes (three groups in the figure) are applied: 10%, 30%, and 70% of bin size. Then, for each overlap size (within each of the three groups in the figure), the spectra are successively divided into n bins ($n = \overline{2,4}$). It can be observed that with the exception of the bins at the edges of the spectra, each bin overlaps with the neighboring bins on both sides (left and right). The total overlap is equal to the overlap size given as input. Thus, for example with an overlap size of 10%, the bins overlap with each other by 5% on both sides, except for the two bins at the edges of the spectrum, which overlap by only 5% on one single side. The obtained similarity values for the two artificial spectra are listed in Table 4.1.

Table 4.1. Similarity values, *S*, obtained by comparing the artificial spectra (the predicted spectrum and the corresponding one with randomly shifted multiplets) of Figure 4.3 using different numbers of bins with different overlap sizes.

Overlap	**Number of bins**								
between bins	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
0%	1.0000	0.9159	0.9622	0.8843	0.8069	0.8786	0.9307	0.8788	0.8754
10%	1.0000	0.9767	0.9603	0.8880	0.8706	0.8575	0.9320	0.8959	0.8659
30%	0.9255	1.0000	0.9837	0.9238	0.9482	0.9122	0.9534	0.9241	0.8988
70%	0.9732	0.9397	0.8953	0.9510	0.9402	0.9238	0.8779	0.8976	0.8792
90%	1.0000	0.9611	0.9999	0.9999	0.9768	0.9523	0.9127	0.9173	0.8931

It can be observed that the larger the overlap used, the higher are the similarity values (taken, on average, over the different numbers of bins, especially with bins overlapping by 30% and 90%). The question is: In which proportion will the similarities of unrelated ^{1}H NMR spectra be higher? In order to answer the question, tests have been conducted in Chapter 5.

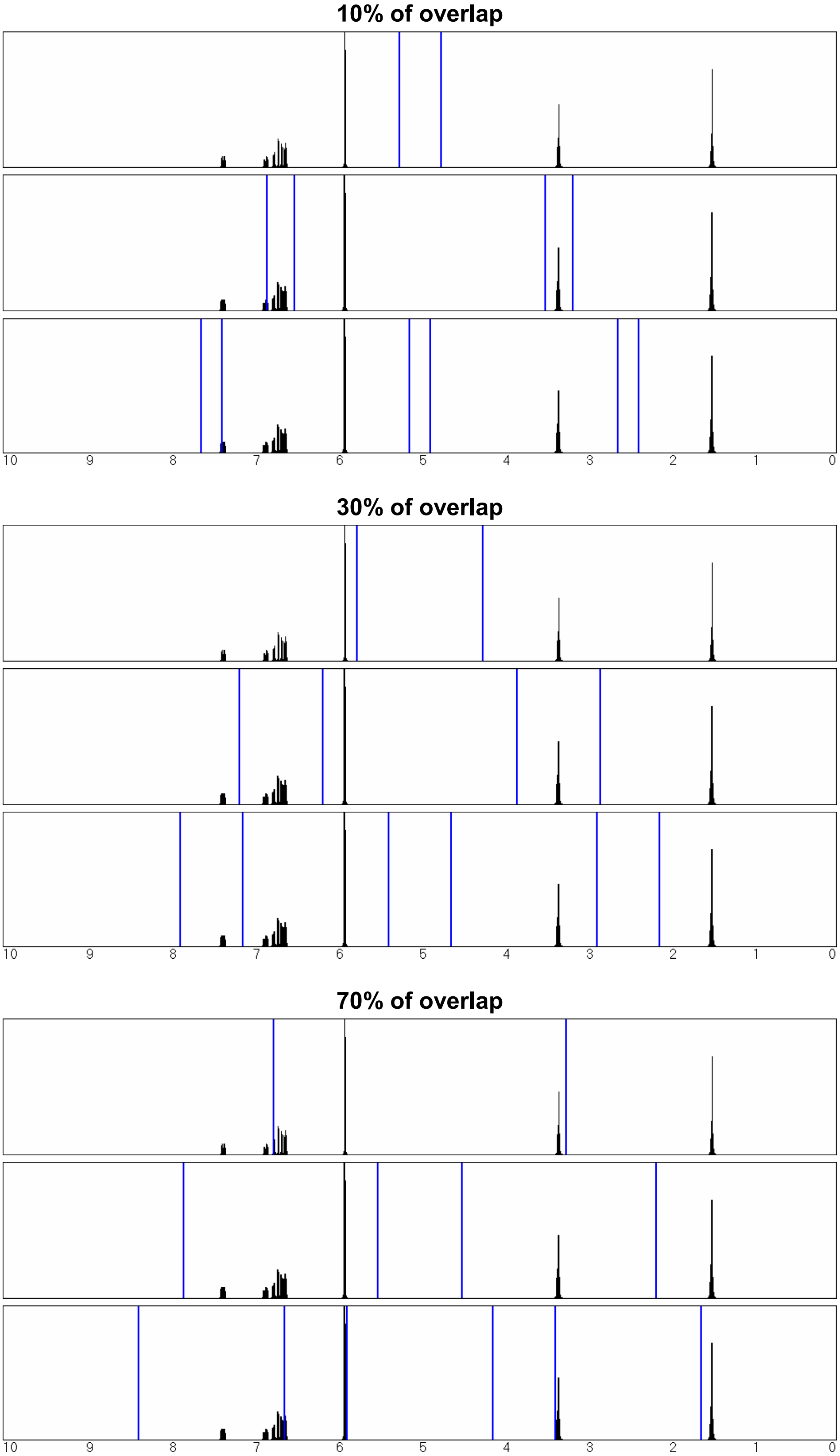


Figure 4.5. Calculating the similarity between the artificial ^{1}H NMR spectra of Figure 4.3: predicted spectrum and the corresponding one with randomly shifted multiplets (not shown here). Overlapping bins are used with overlaps of 10%, 30%, and 70% (of bin size). The spectra are successively divided into n bins, with $n = \overline{2,4}$ (spectra within each of the three groups), yielding the following similarity values: 1.0000, 0.9767, and 0.9603 (10% overlap); 0.9255, 1.0000, and 0.9837 (30% overlap); 0.9732, 0.9397, and 0.8953 (70% overlap).

4.6 Tests with artificial spectra

A comparison of the performance of the new similarity measure with the cross-correlation method[104] and correlation coefficient was first tested with the ten compounds of Figure 4.1, whose spectra were predicted with the computer program NMRPrediction 3.0 program.[133] Additionally, for each structure, two further spectra were calculated in which the multiplets were randomly shifted using a normal distribution with $\mathrm{SD} = 0.2$ or 0.4 ppm, one example with $\mathrm{SD} = 0.4$ ppm being shown in Figure 4.3 (bottom). The spectral similarities calculated by the correlation coefficient, the cross-correlation method,[104] and the new bin method proposed here are shown in Figure 4.6 as dotted, dashed, and thick lines, respectively. 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 $\mathrm{SD} = 0.4$ ppm. Analogously, for the last ten entries (56–65), the random shifts correspond to $\mathrm{SD} = 0.2$ 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, $w(\Delta)$ (cf. Section 3.7.3), for the cross-correlation method and the minimal bin width for the bin method have been systematically varied.

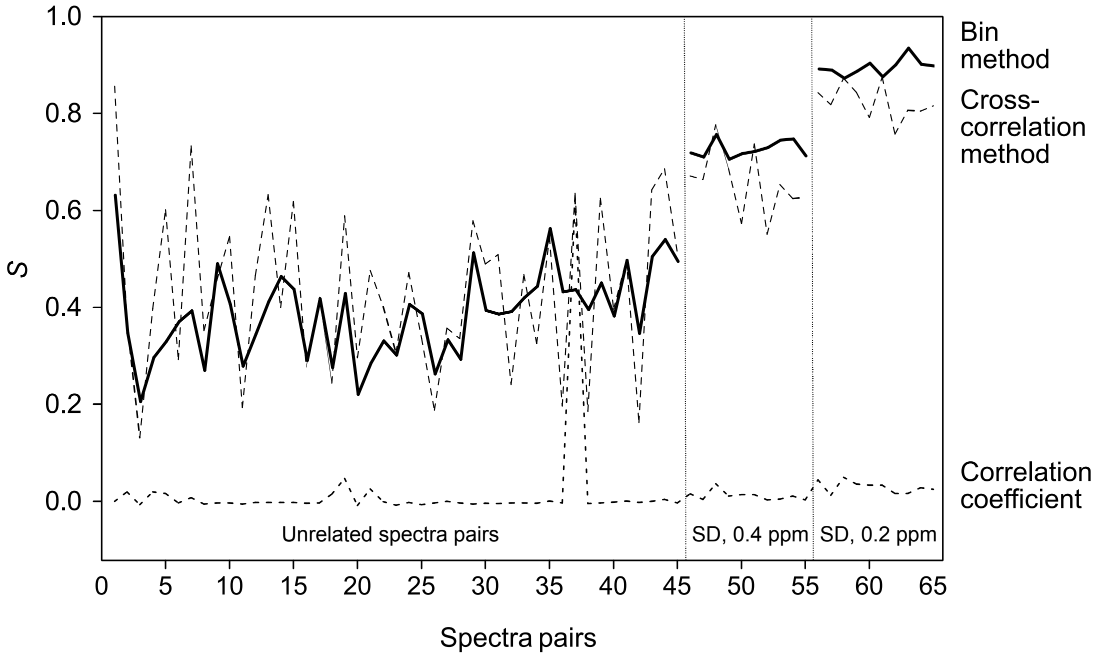


Figure 4.6. Similarity achieved by the three measures investigated here: correlation coefficient (dotted line), cross-correlation method (dashed line), and bin method (thick line). The ten spectra (Figure 4.1) are compared with those corresponding to other structures (entries 1–45), and with those having randomly shifted signal groups (entries 46–55: $\mathrm{SD} = 0.4$ ppm; entries 56–65: $\mathrm{SD} = 0.2$ ppm). The last two sets correspond to an average of the results obtained from 100 spectra comparisons.

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 Figure 4.6, the correlation coefficient is not an adequate measure, but the other two methods discussed here fulfill the stated requirement. The overall discrimination is, however, better with the bin method. For example, for the spectra with $\mathrm{SD} = 0.4$ 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.

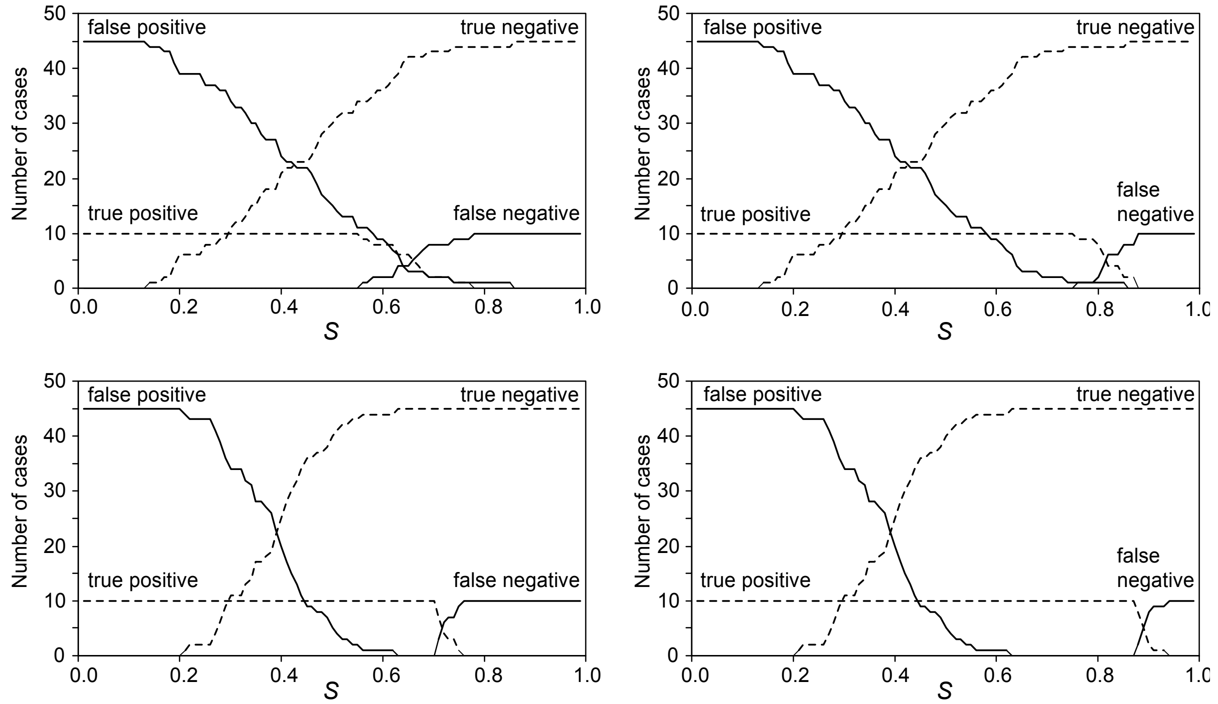


Figure 4.7. Entries of the contingency tables as a function of the threshold value of the similarity, *S*. Top: weighted cross-correlation method (triangle weighting, 1.4 ppm cut-off range). Bottom: bin method (minimal bin width, 0.4 ppm). The ten true positive pairs result from comparing the original spectra with those having randomly shifted signal groups, corresponding to $\mathrm{SD} = 0.4$ ppm (left) and $\mathrm{SD} = 0.2$ ppm (right).

A more detailed comparison is possible on the basis of the contingency diagrams shown in Figure 4.7. 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 numbers 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.[104]

4.7 Tests with measured spectra

Further tests were conducted with 1146 entries of a ^{1}H NMR spectral library (see Section 4.2). 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). Ideally, all normal comparisons should lead to a high, and the random ones to a low similarity value. As indicated in Figure 4.8, the similarity measure used has an influence on the results.

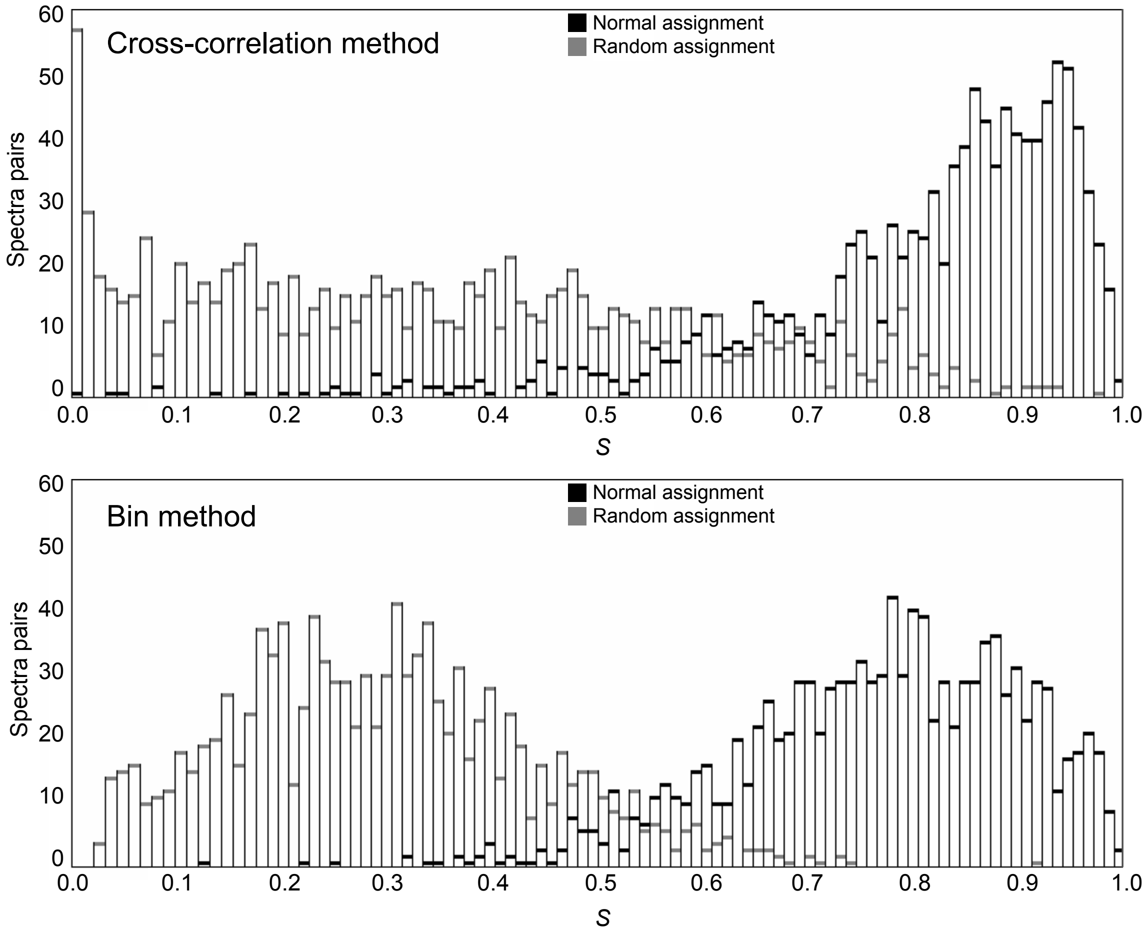


Figure 4.8. Histograms of similarity values, *S*, of measured and calculated ^{1}H NMR spectra using correct and random structure assignments. Top: weighted cross-correlation method (triangle weighting, 1.4 ppm cut-off range). Bottom: bin method (minimal bin width, 0.4 ppm).

With the weighted cross-correlation method[104] 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.

4.8 Conclusions

The similarity of related ^{1}H NMR spectra has been successfully detected by a novel method based on dividing the spectra into $n = \overline{1, N}$ 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 in this case 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. Here, it has only been tested with one-dimensional ^{1}H NMR spectra, but the application of the new method with spectra of two or more dimensions including image analysis is straightforward. This will be shown in Chapters 5–8.

5 Similarity of ^{1}H NMR spectra

The new similarity measure, the bin method, presented in Chapter 4, will now be applied to the comparison of ^{1}H NMR spectra using two further test sets. The first test set consists of 289 chemical structures and the corresponding ^{1}H NMR spectra. Additionally, the two-dimensional HSQC (heteronuclear single quantum correlation) spectra are available so that experiments with both types of spectra are possible. The second test set consists of 96 compounds obtained by parallel synthesis and is presented in Chapter 6.

Apparently, the ^{1}H NMR spectra have some limitations, which generate problems of various types as follows:

1. Limited accuracy of measured chemical shifts: reproducibility problems due to different environments, devices, solvents etc.
2. Limited accuracy of predicted shift values: In the case of ^{1}H NMR spectra, the mean absolute deviation is ca. 0.3 ppm for protons attached to carbon atoms, while for ^{13}C NMR spectra, the mean absolute deviation is approximately 2.0 ppm. Finally, the prediction of X–H chemical shifts (i.e., hydrogens not bonded to carbon) is unreliable. A possible way to avoid the adverse effects of unreliable shift estimations of X–H protons is presented in Section 5.4.2.
3. Noise and impurities: Most spectra have a noise, which has to be removed; otherwise, integral values would be distorted.
4. Solvent signals: The automatic removal of solvent signals is difficult because their positions may vary and because they may overlap with those of the sample.

Solutions for the last two problems are presented in the following sections.

5.1 Elimination of noise

Most ^{1}H NMR spectra contain a noise, which has to be removed so as not to distort the integral values of the sample signals. In order to do this, first, the standard deviation of the intensities in a given interval (i.e., 0.5% of the total range of the spectrum) at both edges of the spectrum is calculated. It is assumed that no signals occur in these regions. Finally, the intensities that are lower than three times the above calculated standard deviation are eliminated (zeroed). This is a very efficient, simple, and fast method for noise removal.

5.2 Elimination of solvent signals

Solvent signals including those of water are so varying in their position that it is sometimes difficult to identify them. Moreover, they may overlap with other signals of the sample.

The signals of most common solvents are removed: Dimethyl sulfoxide-*d*$_5$ (DMSO-*d*$_5$) and $CHCl_3$ occur as impurities in the corresponding deuterated solvents (DMSO-*d*$_6$ and $CDCl_3$). The signal of the reference, TMS, is also eliminated by the same procedure as for solvent signals. For DMSO-*d*$_5$, the theoretical signal is constructed as a Lorentz curve, using the corresponding coupling constant (J), intensity ratio, and line width, the values of the parameters being defined empirically. Finally, the above constructed signal is removed from the spectrum at the thus found chemical shift (in the case of DMSO-*d*$_5$, usually around 2.50 ppm). The chemical shifts of the solvents were calculated empirically by examining several experiments.[135] As solvent signals may have shifted, a tolerance window is applied (e.g., ±0.02 ppm for DMSO-*d*$_5$). The size of this window is important because if it is too narrow, the signal is not completely removed, and if it is too wide, real sample signals may be eliminated. The same procedure is done for the signals of $CHCl_3$, TMS, and water, with the difference that they are singlets and not multiplets as that of DMSO-*d*$_5$. Obviously, signals of solvent and water are removed at the same time.

In the following, the removal procedures of the most common solvent signals, i.e., those of DMSO-*d*$_5$, $CHCl_3$, water, and the reference, are summarized.

1. Removing the DMSO-*d*$_5$ signal (impurity in the solvent DMSO-*d*$_6$):
    - find the exact position and intensity at 2.50 ± 0.02 ppm; if the structure contains $SiCH_3$ groups, then at 2.50 ± 0.1 ppm
    - multiplet structure
        - J = 1.8 Hz
        - intensity ratio: 1:2:3:2:1
        - line width at half-height: 2 Hz
    - create the theoretical signal by constructing the five Lorentz curves (use signal range of 20 Hz) and subtract from the measured spectrum at the corresponding position.
2. Removing the water signal (since DMSO-*d*$_6$ is hygroscopic, water is always present):
    - find the exact position and intensity at 3.31 ± 0.10 ppm; if the solvent contains also CCl_4, then, at 3.13 ± 0.10 ppm
    - find the line width at half-height

- construct the corresponding Lorentz curve (using a signal range of 20 Hz) and subtract

3. Removing the $CHCl_3$ signal (impurity in the solvent $CDCl_3$):
   - find the exact position and intensity at 7.26 ± 0.10 ppm
   - find the line width at half-height
   - construct the corresponding Lorentz curve (using a signal range of 20 Hz) and subtract.
4. Removing the signal of water present in $CDCl_3$:
   - find the exact position and intensity at 1.55 ± 0.03 ppm
   - find the line width at half-height
   - construct the corresponding Lorentz curve (using a signal range of 20 Hz) and subtract.
5. Removing the TMS signal:
   - find the exact position and intensity at 0 ± 0.005 ppm
   - if the structure contains $SiCH_3$, then, do not remove the TMS signal
   - construct the corresponding Lorentz curve (using a signal range of 20 Hz) and subtract.

In order to demonstrate the effects of removing solvent and noise signals, we consider the structure in Figure 5.1 and its corresponding spectra (measured and predicted) presented in Figure 5.2. Figure 5.3 shows the measured spectrum in three different states: a) original measured spectrum, b) after eliminating solvent signals, and c) after removing the noise as well. Obviously, the similarities, *S*, will differ when the predicted spectrum is compared with the measured one being in the three different states: a) $S = 0.3660$, b) $S = 0.5587$, and c) $S = 0.5909$.

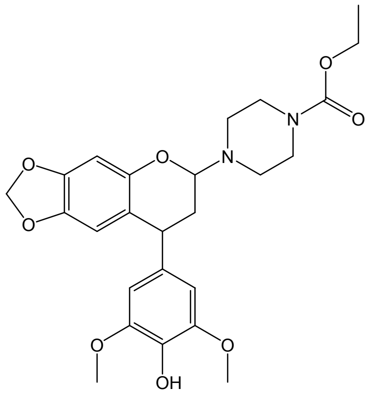


Figure 5.1. Example of a structure (the corresponding measured and predicted spectra are given in Figure 5.2) used to demonstrate the effects of eliminating solvent and noise signals.

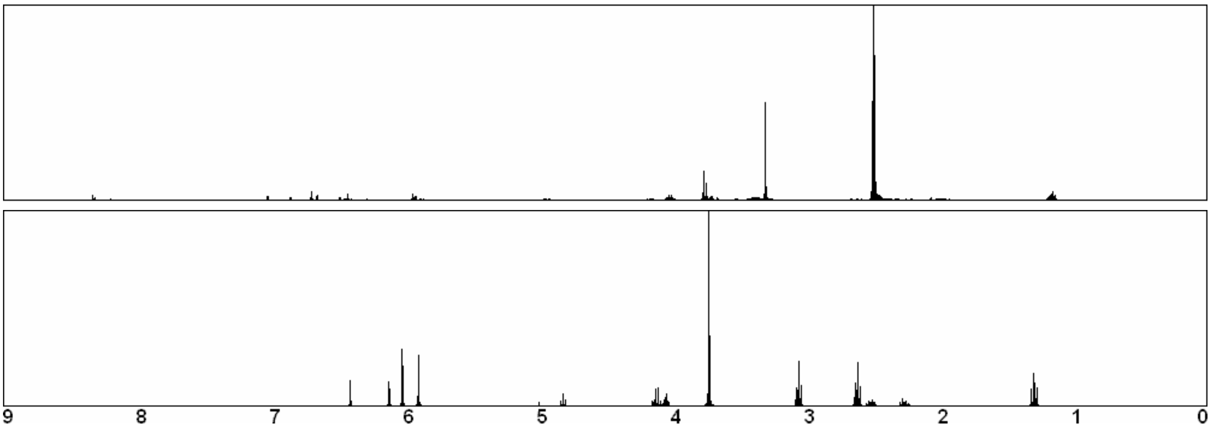


Figure 5.2. Measured (top) and predicted (bottom) spectra of the structure shown in Figure 5.1 used to demonstrate the effects of eliminating solvent and noise signals. The measured spectrum is unprocessed (solvent and noise signals are not removed).

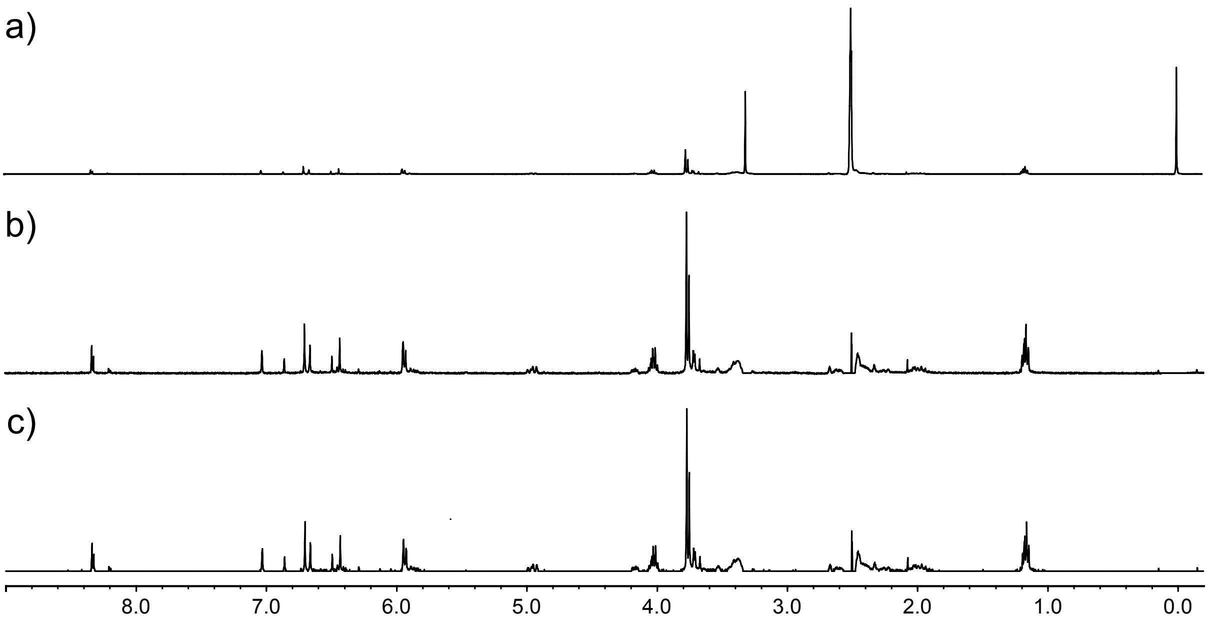


Figure 5.3. From the original measured spectrum (a), first, the solvent signals are eliminated to give (b); then, the noise is removed as well (c). Comparing the three spectra with the predicted one yields the following similarities, *S*: a) 0.3660, b) 0.5587, and c) 0.5909. For comparing the spectra, the bin method was used with a minimal bin width of 0.2 ppm (i.e., a maximal number of 45 bins).

5.3 Test set

The test set consisted of 289 structures and the corresponding ^{1}H NMR and HSQC spectra. From the 447 compounds originally available in the database, those entries were selected for which the NMRPrediction 3.0 software[133] is capable of predicting all chemical shifts with optimal accuracy. Additionally, 21 spectra containing obvious errors identified on the basis of the test runs were removed.

The ^{1}H NMR spectra were recorded in the range of -1–21 ppm with 32 K digital points.

During tests, the original sizes of the spectra were used, and not only the region where signals are present. The signals of DMSO-d_5, $CHCl_3$, and contained water are removed prior to comparison. No other solvents were used for this data set.

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). 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 are analyzed with histograms having 100 clusters/bins. The overlap between two histograms (in 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. In the following, several tests were conducted with different parameter settings of the bin method.

Besides, of 289 compounds, 250 contained at least one proposed (incorrect) structure (in a few cases, more): 37 cases were excluded from the tests because they did not have proposed structures, one compound had an unreadable file of the proposed structure, and another one had the same elucidated (correct) and proposed structure. In some cases, there was more than one proposed structure so that a total of 261 were present. In separate tests, each measured spectrum was compared with that predicted for the corresponding correct structure as well as with the predicted spectrum of the proposed structure. Higher average similarities are expected to be achieved when the correct structures are used for prediction. Note that these structures were proposed by a specialist, hence, they do not obviously contradict the spectra and in most cases are very similar to the correct ones.

5.4 Results and discussion

5.4.1 Influence of parameters on the performance of compatibility tests with ^{1}H NMR spectra

In the following, several tests will be conducted using the above presented data set. Different bin widths (i.e., applying different numbers of bins) were tried. Figure 5.4 shows the overlap percentage between the histograms of similarity values, *S*, of measured and calculated ^{1}H NMR spectra using correct and random structure assignments. It is observed that the lowest overlap of 11.4% is achieved with a minimal bin width of 0.4 ppm (i.e., a maximal number of 55 bins). Thus, all further comparisons were made using this value of minimal bin width. For this case, the histograms of similarity values of measured and calculated ^{1}H NMR spectra are presented in Figure 5.5. Note that metrics of axes for these and all forthcoming histograms are the same as indicated in Figure 4.8 (Section 4.7).

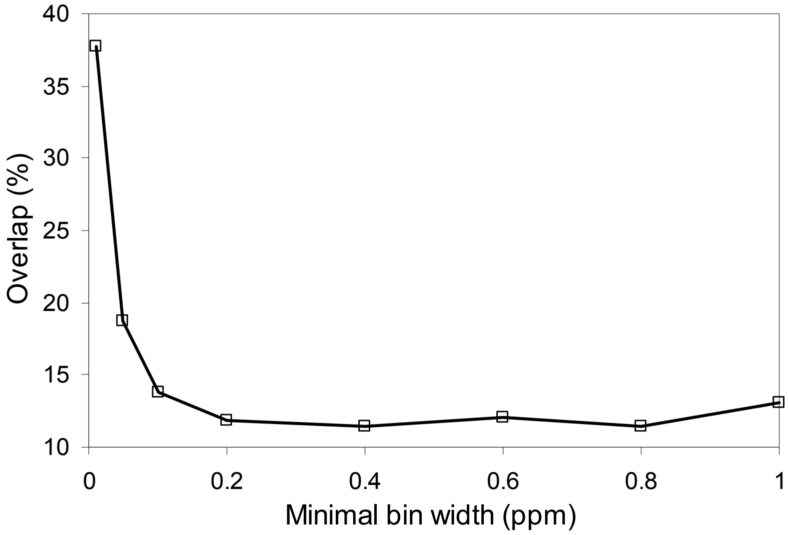


Figure 5.4. Testing the bin method with various minimal bin widths (0.01, 0.05, 0.1, 0.2, 0.4, 0.6, 0.8, and 1 ppm): overlap percentage between the histograms of similarity values, *S*, of measured and calculated ^{1}H NMR spectra using correct and random structure assignments.

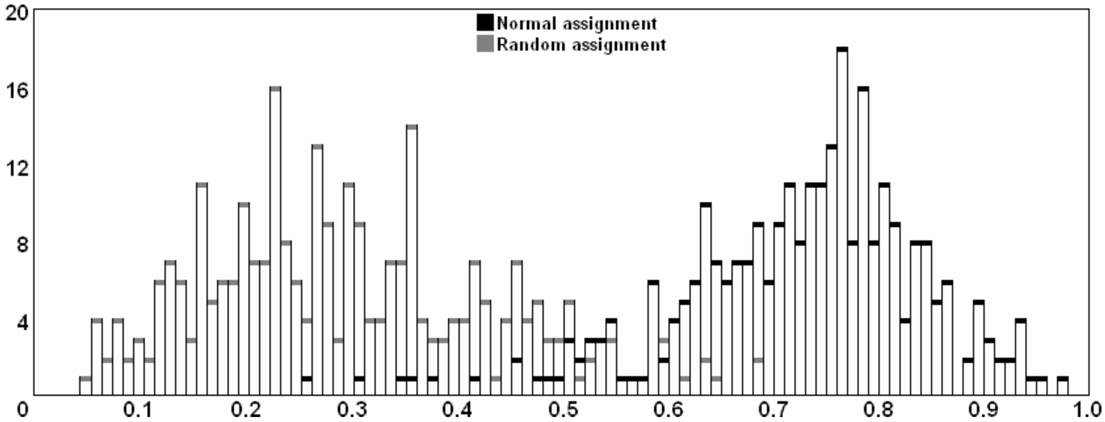


Figure 5.5. Histograms of similarity values of measured and calculated ^{1}H NMR spectra using normal and random structure assignments: the overlap is 11.4% (33 spectra pairs). For comparing the spectra, the bin method was used with a minimal bin width of 0.4 ppm. Metrics of axes as in Figure 4.8.

As discussed in Section 4.5, overlapping bins may improve the performance of similarity measure. Thus, tests are conducted using the bin method with a maximal number of 55 and 110 bins and various sizes of overlaps: 10%, 30%, 50%, 70%, and 90% of bin size (Figure 5.6). It can be seen that the lowest spectra pairs overlap of 10.7% was achieved with 55 bins and an overlap of 90% between bins. It is observed that with less bins, better results (lower overlapping between normal and random assignments of spectra pairs) are obtained if overlapping bins are used. However, the overlap is not significantly lower than the 11.4% achieved with non-overlapping bins, using the same number of bins. In conclusion, it can be stated that overlapping bins do not significantly improve the selectivity of the bin method and, therefore, will not be used.

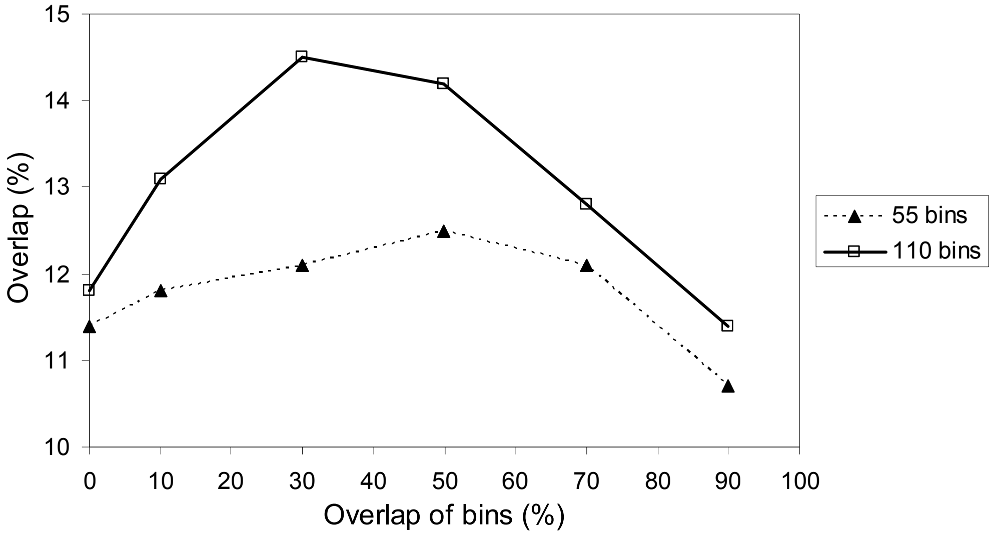


Figure 5.6. Investigating the performance of overlapping bins. Tests were conducted using minimal bin widths of 0.2 and 0.4 ppm (i.e., 110 and 55 bins, respectively) and various sizes of overlaps: 10%, 30%, 50%, 70%, and 90% of bin size. The lowest spectra pairs overlap of 10.7% was achieved with 55 bins and an overlap of 90% between bins.

In a next test, experiments involving the proposed (incorrect) structures were performed. Similarities are expected to be lower when the incorrect structures are used. In a first step, each of the 250 measured spectra was compared with two predicted spectra: one based on the corresponding incorrect structure (normal assignment) and the other, on a randomly selected incorrect structure from the dataset (random assignment). Figure 5.7 shows the resulting two histograms. As expected, here, the overlap is much higher, namely 26.7% in comparison with 11.4% in the case of correct structures. Since the previous two tests were conducted under the same conditions with the same parameter settings, comparison of results is justified. This experiment shows how selective the bin method is, as it clearly discriminates between correct and incorrect structures. Note that the incorrect structures are the expected proposed ones. They were proposed by a specialist, therefore, they should not obviously contradict the spectra and, in most cases, are very similar to the correct ones (in some situations, the only differences arise owing to the compounds being diastereomers or bearing charges, which are not handled by the prediction program).

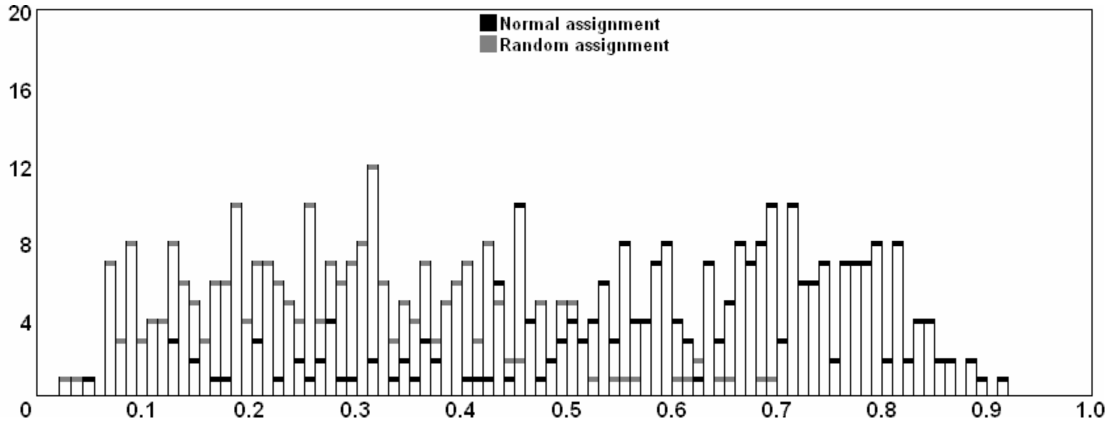


Figure 5.7. Histograms of similarity values of measured and predicted (based on incorrect structures) ^{1}H NMR spectra using normal and random structure assignments: the overlap is 26.7% (70 spectra pairs). For comparing the spectra, the bin method was used with a minimal bin width of 0.4 ppm.

Figure 5.8 reveals important information about the selectivity of the similarity measure. Each of the 250 measured ^{1}H NMR spectra was compared with the corresponding predicted spectrum of the elucidated structure (black histogram). In another test, each of the 250 measured spectra was compared with the corresponding predicted spectrum of the proposed structure (gray histogram). In some cases, there was more than one proposed structure, which explains for the total of 261 comparisons. Differences in diastereomers and charges cannot be detected by the method, as the prediction does not handle them. Notwithstanding, they are also included in the test set. Evidently, the similarities are higher with the correct structures.

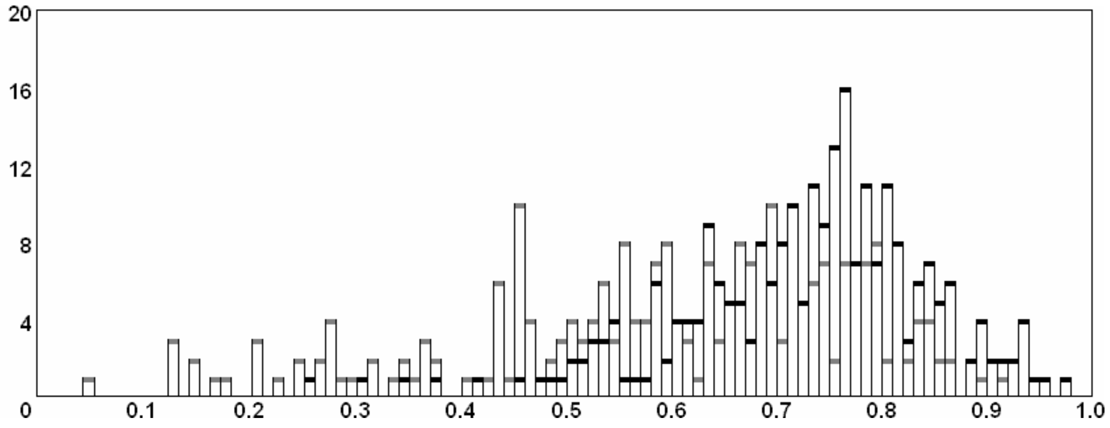


Figure 5.8. Comparing the similarities when correct (black histogram, 250 spectra pairs) and incorrect structures (gray histogram, 261 spectra pairs) are used to predict their ^{1}H NMR spectra. The predicted spectra are compared with measured ones using the bin method with a minimal bin width of 0.4 ppm.

Figure 5.9 presents another type of visualization of the above results. Here, 261 measured HSQC spectra are compared with those predicted for the correct and incorrect structures. Again, as expected, in most cases, higher similarities are achieved with correct structures. The average similarity of 261 comparisons is 0.73 (with correct structures) and 0.61 (with incorrect structures). Since even for the elucidated (correct) structures, the similarity can be lower than 0.3, the method is not selective enough if related structures must be excluded.

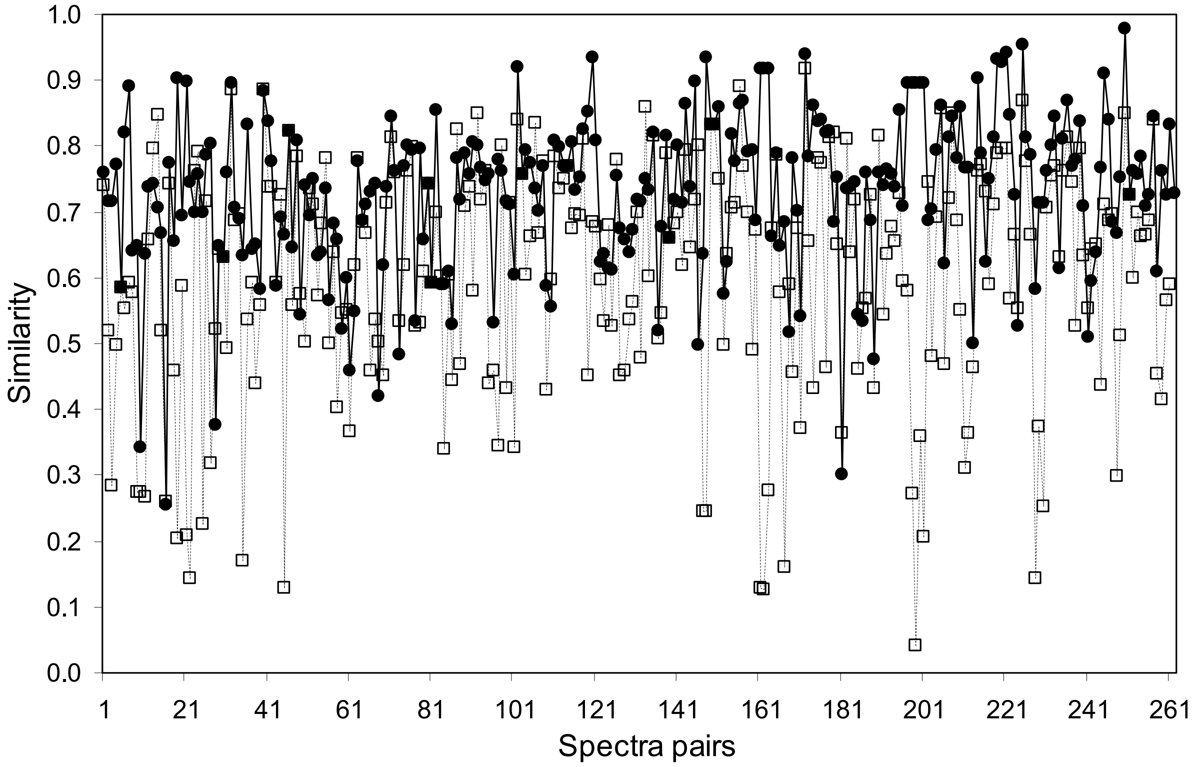


Figure 5.9. Comparing the similarities when correct (solid line) and incorrect structures (dashed line) are used to predict their ^{1}H NMR spectra. The predicted ^{1}H NMR spectra are compared with the measured ones using the bin method with a minimal bin width of 0.4 ppm (i.e., a maximal number of 55 bins). The average similarity of 261 comparisons is 0.73 in the case of correct structures and 0.61 with incorrect ones.

As a conclusion, it can be stated that this method is able to discriminate between correct and incorrect structures by comparing their predicted spectra with the corresponding measured ones. When correct structures are used, the overlap between spectra pairs using normal and random assignments is as low as 11.4%. On the other hand, when incorrect structures are used, the overlap between the two cases is as high as 26.7%. The best results were achieved with a minimal bin width of 0.4 ppm (i.e., a maximal number of 55 bins).

5.4.2 Eliminating X–H signals from ^{1}H NMR spectra

The aim is to exclude the X–H signals (i.e., hydrogens not bonded to carbon) both from measured and predicted ^{1}H NMR spectra, as the chemical shifts of these signals highly depend on the experimental conditions and, therefore, are difficult to be predicted accurately. It is expected that the performance of the similarity measure (the bin method) will improve, i.e., the method will be more selective and will better discriminate between related and foreign spectra.

In order to remove X–H signals from the measured proton NMR spectrum, the corresponding measured HSQC spectrum is used because only hydrogens attached to carbons lead to signals in the latter. Figure 5.10 contains the algorithm that removes the X–H signals both in measured and predicted ^{1}H NMR spectra. First, the HSQC spectrum is read and the signals are defined using the Method1 algorithm (see Figure 7.3, Section 7.2.1). Obviously, the solvent signals are removed. The ^{1}H NMR coordinates of the HSQC signals are obtained by projecting the HSQC signals on the x axis of the HSQC spectrum. In a next step, the ^{1}H NMR spectrum is read from file and preprocessed: First, the integrated intensities are normalized to the 0–1 range, then, the solvent signals and the noise are eliminated. For each of the HSQC signals, the corresponding signal is searched in the proton NMR spectrum: In a given range, the signal peak is identified and, then, the edges of the signal are found. After all the found signals in the HSQC have been retained from the ^{1}H NMR to a new empty ^{1}H NMR spectrum (represented by a vector), the intensities in the new spectrum are normalized to the total number of protons bonded to carbon atoms.

In the predicted ^{1}H NMR spectrum, the X–H signals are removed as well. Then, the two processed (measured and predicted) ^{1}H NMR spectra are compared using the bin method in the same manner as in the case of simple measured and predicted one-dimensional spectra.

```
•  read the HSQC spectrum
•  find the signals using the Method1 algorithm
•  remove solvent signals
•  get the $^1$H NMR shifts of HSQC signals (project signals) in ppm
•  read the $^1$H NMR spectrum
•  normalize the integrated intensities to the range of 0—1
•  remove solvent signals
•  eliminate noise
•  find the $^1$H NMR signals from the $^1$H NMR spectrum using the
   signal positions found in HSQC spectra and a window range of
   ±0.15 ppm; find the edges of the signal by going to the left
   and right until the ratio of the actual intensity and the
   maximum intensity of the spectrum is > 0.003
•  retain all found signals in the HSQC from the $^1$H NMR to a new
   (empty) $^1$H NMR spectrum (represented by a vector)
•  normalize the integrated intensities in the new $^1$H NMR
   spectrum to the number of C-H signals
•  remove the X-H signals from the predicted $^1$H NMR spectrum
```

Figure 5.10. Algorithm for excluding the X–H signals in the measured and predicted ^{1}H NMR spectra.

In the following, analogously to the previous section, several tests are conducted using the above data set. The bin method was tested with various numbers of bins. Again, the best performance (overlap between spectra pairs of 5.9%) was achieved with a minimal bin width of 0.4 ppm (i.e., a maximal number of 55 bins). In this case, the histograms of similarity values of measured and calculated ^{1}H NMR spectra (without X–H signals) are presented in Figure 5.11. This is a significant improvement in comparison with the previous method (using the original ^{1}H NMR spectra), where an overlap as high as 11.4% was achieved (cf. Figure 5.5).

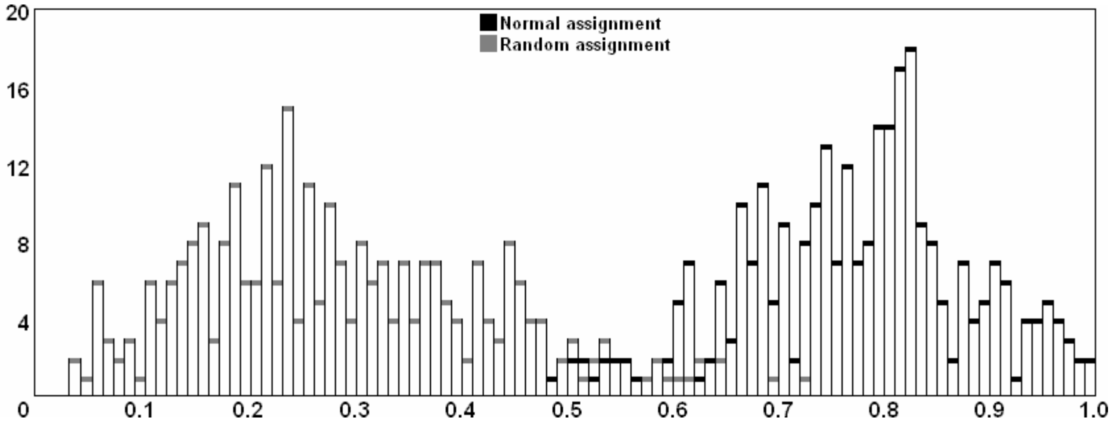


Figure 5.11. Histograms of similarity values of measured and calculated ^{1}H NMR spectra (without X-H signals) using normal and random structure assignments: the overlap is 5.9% (17 spectra pairs). For comparing the spectra, the bin method was used with a minimal bin width of 0.4 ppm.

Since in the case of the original ^{1}H NMR spectra (containing X–H signals), the bin method with overlapping bins did not significantly improve the performance, it was not tested here.

In a next test, experiments involving the proposed (incorrect) structures were performed. As before, each of the 250 measured spectra was compared with two predicted ones: one based on the corresponding incorrect structure (normal assignment) and the other based on a randomly selected incorrect structure from the dataset (random assignment). Figure 5.12 shows the resulting two histograms. As expected, here, the overlap is much higher, namely 26.0% in comparison with 5.9% in the case of correct structures. This test proves that the bin method can discriminate between correct and incorrect structures. With the latter, the aim is to get a high overlap, while in the case of correct structures, it should be as low as possible. In the previous section, the overlap for incorrect structures was slightly higher (26.7%, see Figure 5.7), but did not increase at the same rate as the overlap in the case of correct structures. Thus, it can be stated that the ^{1}H NMR spectra without X–H signals really do improve the performance of the bin method.

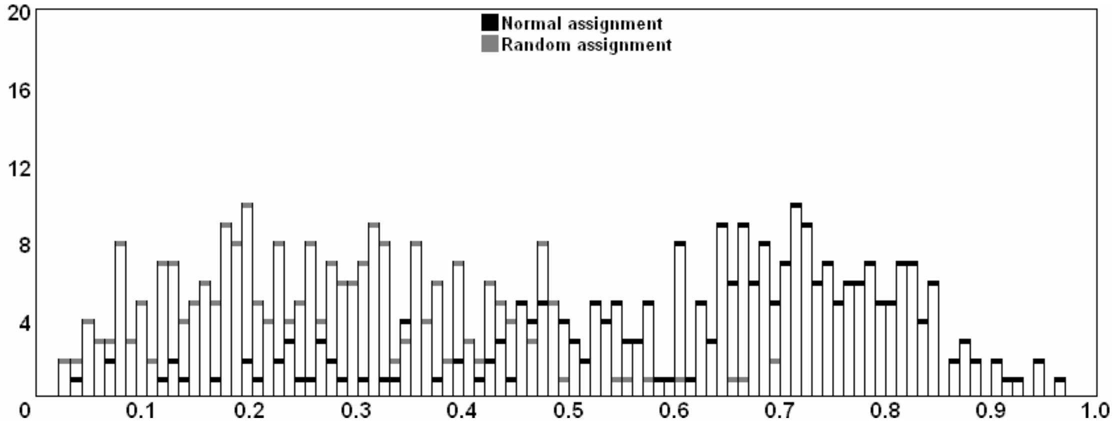


Figure 5.12. Histograms of similarity values of measured and predicted (based on incorrect structures) ^{1}H NMR spectra using normal and random structure assignments: the overlap is 26.0% (68 spectra pairs). For comparing the spectra, the bin method was used with a minimal bin width of 0.4 ppm (i.e., a maximal number of 55 bins).

In Figure 5.13, each of the 250 measured ^{1}H NMR spectra (without X–H signals), on the one hand, was compared with the corresponding spectrum predicted for the elucidated structure (black histogram) and, on the other hand, with the corresponding spectrum predicted for the proposed structure (gray histogram). It is clear that the correct structures yield much higher similarities. The lowest similarity is slightly less than 0.5, and the highest similarity of 1 is reached by several spectra pairs. This is a much better performance than that achieved with the original ^{1}H NMR spectra (see Figure 5.8), where the lowest similarity is 0.25 and none of the spectra pairs managed to reach the maximal similarity of 1.

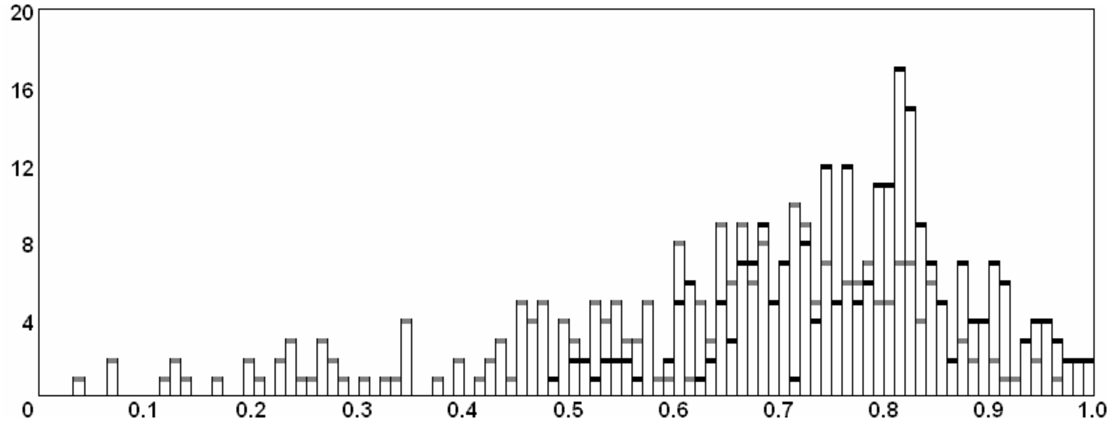


Figure 5.13. Comparing the similarities when correct (black histogram, 250 spectra pairs) and incorrect structures (gray histogram, 261 spectra pairs) are used to predict their ^{1}H NMR spectra. The predicted spectra are compared with the measured ones using the bin method with a minimal bin width of 0.4 ppm.

Figure 5.14 presents the above results in another form. The 261 measured HSQC spectra are compared with the spectra predicted for the correct and incorrect structures. As expected, in most cases, higher similarities are achieved with the correct structures, the average similarity of 261 comparisons being 0.78 (with correct structures) and 0.62 (with incorrect structures).

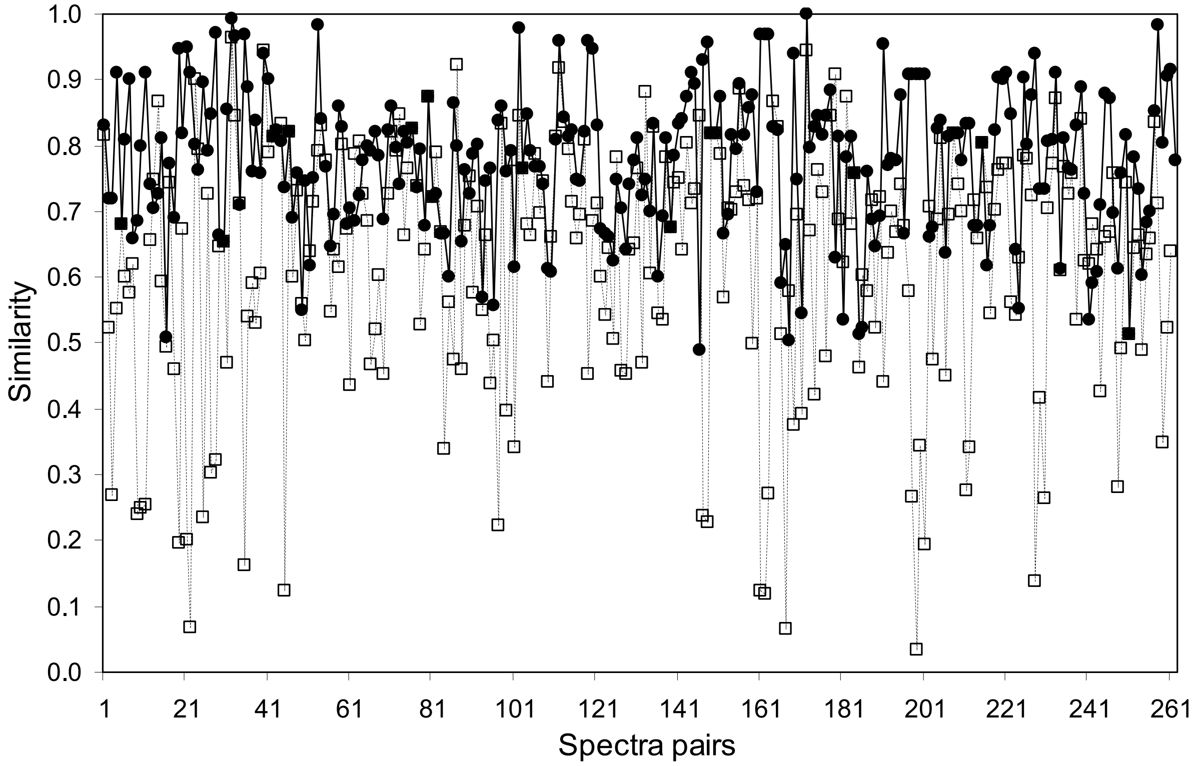


Figure 5.14. Comparing the similarities when correct (solid line) and incorrect structures (dashed line) are used to predict their ^{1}H NMR spectra. The predicted ^{1}H NMR spectra are compared with the measured ones (without X–H signals) using the bin method and a minimal bin width of 0.4 ppm (i.e., a maximal number of 55 bins). Average similarity of 261 comparisons: 0.78 (with correct structures) and 0.62 (with incorrect structures).

As a conclusion, it can be stated that this method produces better, i.e., more selective results than the previous one, where the original ^{1}H NMR spectra (with X–H signals) was used. Optimal results were achieved with a minimal bin width of 0.4 ppm (i.e., a maximal number of 55 bins). The only drawback is that the method is somewhat slower as the corresponding HSQC signals are needed.

5.5 Conclusions

In Chapter 4, it has been shown that the bin method is an efficient and generally applicable similarity measure that can also successfully compare ^{1}H NMR spectra. This is confirmed by the results of the present chapter, where it was applied to a further real-life data set containing 289 ^{1}H NMR spectra. By excluding the X–H signals (i.e., hydrogens not bonded to carbon) from both measured and predicted spectra, the selectivity of the criterion is increased. The overlap between normal and random assignments is almost halved: 5.9% instead of 11.4% when the X–H signals are retained in the spectra.

6 Automatic spectra verification of a combinatorial library

In the previous chapter, the bin method has been successfully applied to the comparison of ^{1}H NMR spectra. Also here, ^{1}H NMR spectra will be used, but the data is a product of parallel syntheses generated from repeated building blocks (combinatorial library) for which methods of automatic spectra verification will be presented. Automatic spectra and structure verification is a novel research area and, so far, only few papers have been published in this field. A recent article[136] presents an automated structure verification method based on ^{1}H NMR prediction. The results are promising but statistically not significant because the data sets used are small. Moreover, the parameters and threshold values were manually customized for each test set and none of the test sets was a combinatorial library.

In this work, the aim is to classify the ^{1}H NMR spectra based on their quality and correctness. Thus, the problem, on the one hand, is more difficult as the chemical structures and, therefore, their spectra are highly similar. On the other hand, since the data set is a combinatorial library, the measured spectra can be compared not only with the predicted ones but also with other spectra of the library, e.g., with product spectra, or combined or individual spectra of the educts (reagents). Finally, the outcome of the automatic spectra verification is compared with the results of the manual analysis.

6.1 Introduction

Nowadays, especially in the process of drug discovery, new compounds can easily be generated by parallel syntheses.[124] High-throughput instruments are able to automatically register ^{1}H NMR spectra on the order of minutes. Thus, the bottleneck is the interpretation of molecular spectra. So far, tools for a reliable automatic interpretation are not yet available.

The bin method (see Chapter 4) is applied here along with a few other approaches to detect erroneous entries in a combinatorial library. They are all based on the comparison of measured spectra with: (1) estimated ones, (2) with other spectra of the combinatorial library, (3) combined spectra of the educts, or (4) individual spectra of the educts. The performance of the methods is characterized with contingency diagrams and the optimal thresholds are defined. Finally, for each spectrum of the combinatorial library, the results of the different approaches are represented in traffic light style: green (here, light gray due to grayscale graphics) if the spectrum is correct, and red (here, dark gray) if the spectrum is incorrect.

6.2 Combinatorial library

The combination of 14 different educts (A1–A6, B1–B4, and C1–C4, see Figure 6.1) is used to generate a total of 96 ($6 \times 4 \times 4$) products.[137] The structures of the products ("topology") are listed in Table 6.1, and Figure 6.2 presents the product number 2 as an example. The ^{1}H NMR spectra of the educts and products are available as well. They are homogeneous in that all of them have 16384 digital points in a range of approximately 20 ppm and the solvent used was DMSO, whose signals along with the noise by the procedure described in Section 5.1 are removed (see Section 5.2 for details).

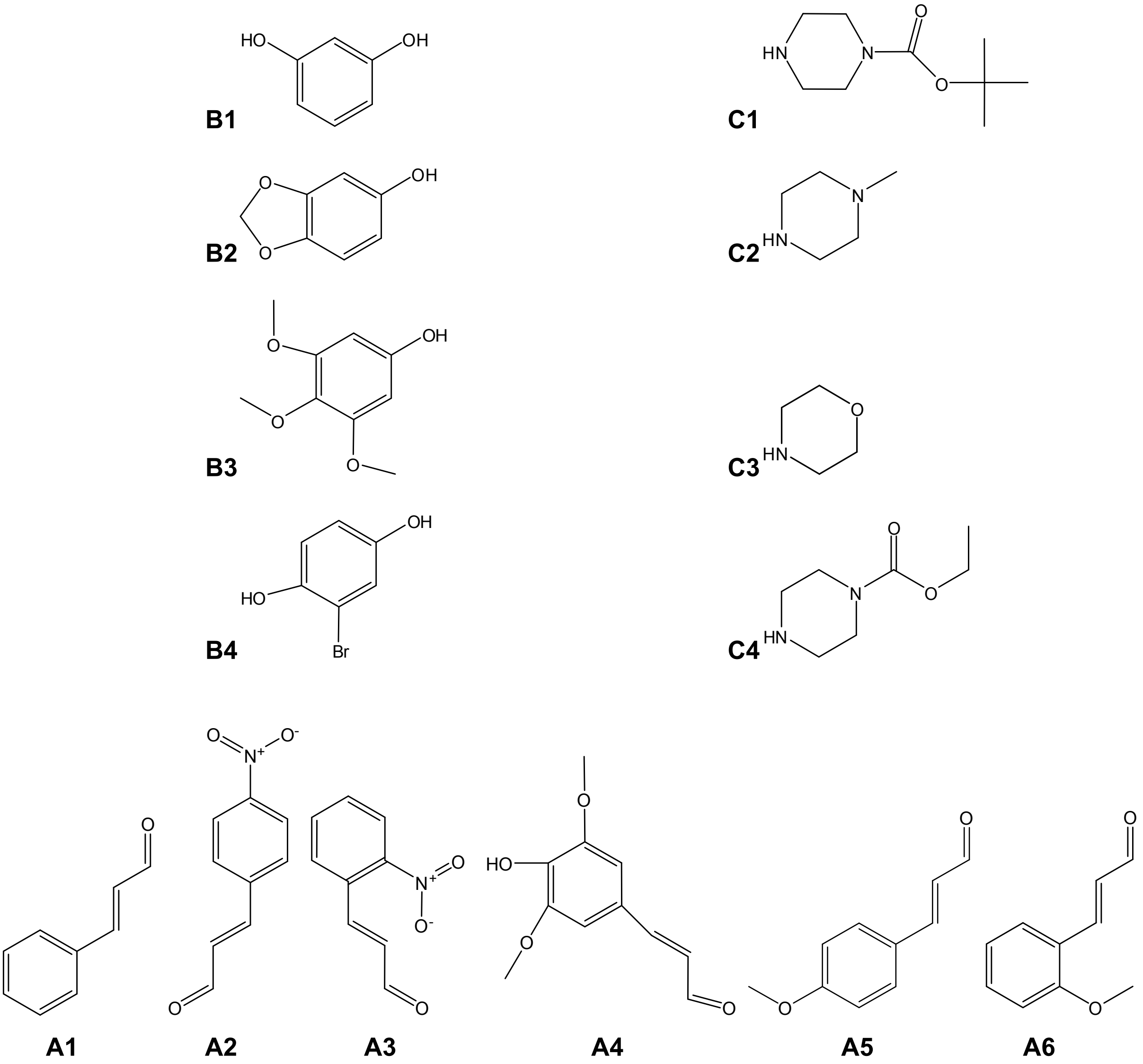


Figure 6.1. Educts used to generate the combinatorial library for the study presented here.

Table 6.1. List of the 96 products in the combinatorial library.

1 A1B2C1	15 A2B3C2	29 A4B2C2	43 A6B3C1	57 A2B2C3	71 A3B3C4	85 A5B2C4
2 A1B1C1	16 A2B4C2	30 A4B1C2	44 A6B4C1	58 A2B1C3	72 A3B4C4	86 A5B1C4
3 A1B3C1	17 A3B2C1	31 A4B3C2	45 A6B2C2	59 A2B3C3	73 A4B2C3	87 A5B3C4
4 A1B4C1	18 A3B1C1	32 A4B4C2	46 A6B1C2	60 A2B4C3	74 A4B1C3	88 A5B4C4
5 A1B2C2	19 A3B3C1	33 A5B2C1	47 A6B3C2	61 A2B2C4	75 A4B3C3	89 A6B2C3
6 A1B1C2	20 A3B4C1	34 A5B1C1	48 A6B4C2	62 A2B1C4	76 A4B4C3	90 A6B1C3
7 A1B3C2	21 A3B2C2	35 A5B3C1	49 A1B2C3	63 A2B3C4	77 A4B2C4	91 A6B3C3
8 A1B4C2	22 A3B1C2	36 A5B4C1	50 A1B1C3	64 A2B4C4	78 A4B1C4	92 A6B4C3
9 A2B2C1	23 A3B3C2	37 A5B2C2	51 A1B3C3	65 A3B2C3	79 A4B3C4	93 A6B2C4
10 A2B1C1	24 A3B4C2	38 A5B1C2	52 A1B4C3	66 A3B1C3	80 A4B4C4	94 A6B1C4
11 A2B3C1	25 A4B2C1	39 A5B3C2	53 A1B2C4	67 A3B3C3	81 A5B2C3	95 A6B3C4
12 A2B4C1	26 A4B1C1	40 A5B4C2	54 A1B1C4	68 A3B4C3	82 A5B1C3	96 A6B4C4
13 A2B2C2	27 A4B3C1	41 A6B2C1	55 A1B3C4	69 A3B2C4	83 A5B3C3	
14 A2B1C2	28 A4B4C1	42 A6B1C1	56 A1B4C4	70 A3B1C4	84 A5B4C3	

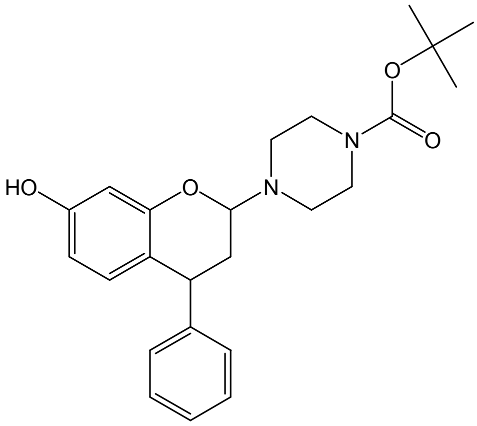


Figure 6.2. The structure of product number 2: A1B1C1.

Originally, it was assumed that each of the 96 product spectra are correct, but apparently, in preliminary tests (using the bin method for comparison of spectra), the combinatorial library was found to contain incorrect and impure spectra. Therefore, the spectra were manually analyzed and interpreted by a specialist. Table 6.2 categorizes them based on their quality. Thus, a perfect test set was constructed so that the performance and reliability of the bin method and the different other approaches could be checked and quantified.

Table 6.2. Analyzing the spectra of products based on their quality.

Quality of spectra	**Product spectra ID**	**Count**
Correct	1, 2, 9, 10, 11, 14, 16, 17, 18, 19, 22, 25, 26, 30, 32, 33, 34, 35, 38, 48, 49, 53, 56, 58, 61, 62, 63, 65, 73, 78, 79, 85, 89, 90, 92	35
Incorrect	3, 4, 7, 20, 31, 39, 43, 45, 47, 60, 64, 68, 71, 74, 75, 83, 87, 88, 91, 93, 94	21
Impure	5, 6, 8, 12, 13, 15, 21, 23, 24, 27, 28, 29, 36, 37, 40, 41, 42, 44, 46, 50, 51, 52, 54, 55, 57, 59, 66, 67, 69, 70, 72, 76, 77, 80, 81, 82, 84, 86, 95, 96	40

6.3 Generating reference spectra

Automatic spectra verification is usually based on the comparison of spectra. For all comparisons made here, the bin method was used. In the following, various approaches to detecting erroneous entries in a combinatorial library are presented. The measured product spectra are compared with different reference spectra, such as:

1. predicted spectrum of the product molecule,
2. measured spectra of the products,
3. sum of the measured spectra of the educts (reagents),
4. individual measured spectra of the educts (reagents).

6.3.1 Comparison with predicted spectra

This is the easiest and most straightforward method to test whether a spectrum is correct or incorrect. Of course, it has its own drawback: To use the predicted spectra adds uncertainty to the system. The predicted spectrum has to have the same properties as the measured one, i.e., the same digital resolution, range, line width, etc. If the similarity is lower than a given threshold value, then, the measured spectrum is claimed to be incorrect. A false positive is generated when an incorrect spectrum has a similarity greater than the threshold. If a correct spectrum yields a similarity lower than the given threshold, this will be accounted for as a false negative case.

6.3.2 Comparison with other measured product spectra

In this method, each measured spectrum is compared with every other one. For each product spectrum, the results are ranked by descending similarity. Spectra pairs that differ in only one educt module should receive a higher similarity value (lower rank number), while those differing in all three modules receive a lower similarity value (higher rank). For correct spectra, the absolute differences between these average ranks (AvgRankDiff) must be larger than for incorrect ones.

The results are evaluated by calculating the false positives and false negatives based on their average rank differences using various thresholds. If in the case of a correct spectrum this difference is less than, or equal to, a given threshold, then it will be accounted for as a false negative. A false positive is generated when an incorrect spectrum has an AvgRankDiff value greater than the threshold.

6.3.3 Comparison with the sum of the measured spectra of the educts

The measured spectra of the educts contained in the product under study (based on the "topology", see Table 6.1) are, first, normalized with the total number of protons and, then, summed up. Each thus artificially constructed spectrum is now compared with each of the measured product spectra. For each artificially built spectrum, the comparisons are ranked by descending similarity. An artificial spectrum should have the highest similarity (rank of 1) with the corresponding measured spectrum, and get a lower similarity (higher ranking number) when it is compared with foreign spectra.

In order to select the best result, we count the number of missed correct (false negative) and missed incorrect (false positive) spectra based on their rank and using different thresholds. In the case of correct spectra, missed means that the rank is greater than a certain threshold (rank), while in the case of incorrect spectra, the rank is less than, or equal to a given threshold.

6.3.4 Comparison with the individual measured spectra of the educts

In this method, each measured product spectrum is compared with each of the educt spectra. For each product spectrum, the comparisons are ranked by descending similarity. It is expected that a product spectrum has higher similarity values with the spectra of those educts that are contained in the product than with those that are not. Some of the educt types (e.g., type A in the library, cf. Figure 6.1 and Figure 6.2) are strongly modified in the products and, thus, the similarities with these educts are not relevant.

Using this method, a false positive is generated if an incorrect spectrum has a rank less than, or equal to the threshold. If a correct spectrum has a rank greater than the given threshold, then it will be counted as false negative.

6.4 Results and discussion

The measured ^{1}H NMR spectra of the library were compared with the generated reference spectra. Note that the tests were carried out with each of the 96 products, including also the incorrect and impure spectra besides the correct ones. Then, for each method, the optimal thresholds were defined in order to generate a response in the traffic light style. When comparing the spectra, the bin method was tested with different numbers of maximal bins, the minimal bin width of 0.2 ppm proving to be the optimal parameter. The integrated intensities of the spectra were normalized to the total number of protons. Prior to the comparison, the noise and solvent signals (DMSO and contained water) were removed (for details, see Sections 5.1 and 5.2).

In the following, detailed results are presented for each of the four methods. First, the measured spectra were compared with the predicted ones (cf. Section 6.3.1). The histograms of similarity values of measured and calculated ^{1}H NMR spectra using normal (black histogram) and random (gray histogram) structure assignments are shown in Figure 6.3. The overlap between the two histograms is 29.2%, which evidently is high because also the incorrect and impure spectra were included in the test. In order to have a more realistic result, tests were also made with the correct spectra only. The resulting histograms in the case of normal and random structure assignments are presented in Figure 6.4. It is observed that the lowest similarity value, achieved when related structures are compared, now is 0.5 (in contrast to 0.1 when incorrect spectra were included). At the same time, the similarities in the gray histogram are lower than 0.6. Thus, only the spectra pairs between 0.5–0.6 are overlapping. This is due to the fact that also foreign structures are very similar since they were generated from repeated building blocks.

Because the random configuration is arbitrary, the overlap between the normal and random assignments changes for each random assignment. Therefore, in the next test, every possible random assignment was generated. The measured spectra were compared with every predicted spectrum of each structure. The histograms of the distribution of the similarities for normal and random assignments (every possible configuration) are shown in Figure 6.5. As expected, the similarities of related spectra are in the upper range of the histograms; the lowest similarity value is at approximately 0.48, but there is a sharp overlap with the similarities of random assignments. Observe that the black histogram is the same as in Figure 6.3.

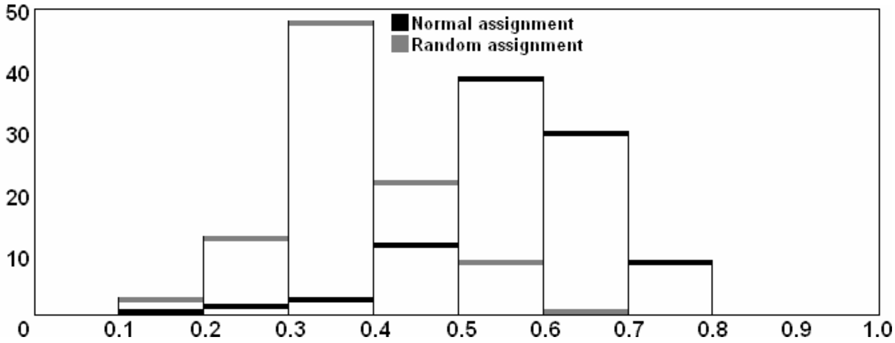


Figure 6.3. Histograms of similarity values of measured and calculated ^{1}H NMR spectra using normal and random structure assignments: the overlap is 29.2% (28 spectra pairs).

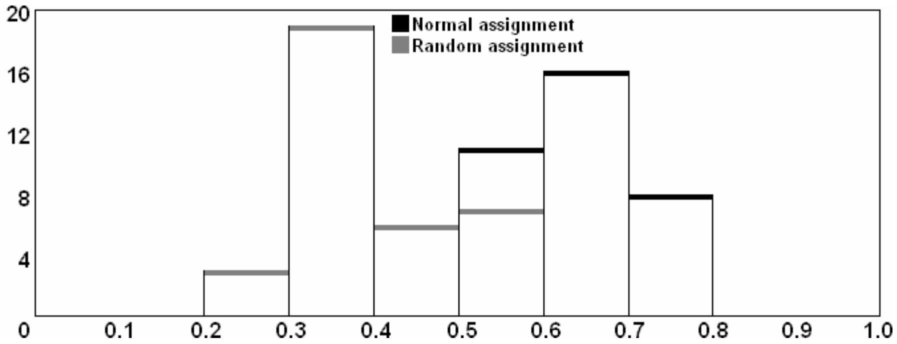


Figure 6.4. Histograms of similarity values of measured and calculated ^{1}H NMR spectra using normal and random structure assignments: the overlap is 20.0% (7 spectra pairs). Note that only the 35 correct spectra (cf. Table 6.2) were used for this test.

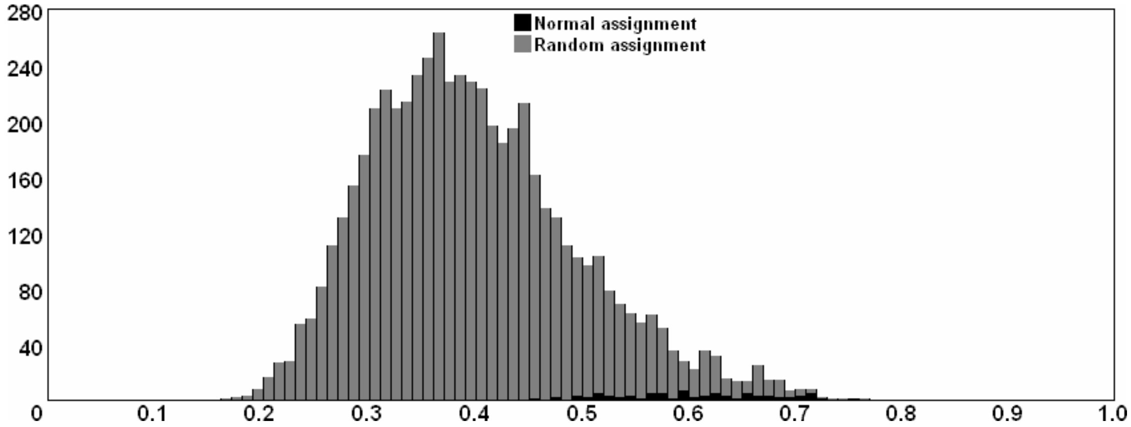


Figure 6.5. Histograms of similarity values of measured and calculated ^{1}H NMR spectra using normal and random structure assignments. Here, each measured spectrum was compared with each predicted one of every structure.

With the next method (cf. Section 6.3.2), each measured product spectrum was compared with every other measured product spectrum of the library and the results were ranked by descending similarity. Thus, two types of average rankings were calculated for each product spectrum: one for spectra pairs that differ in only one educt module and another for those that differ in all three modules. Table 6.3 lists the above mentioned average rank numbers (4th column) for the 21 incorrect spectra, while for the 35 correct ones, the average ranks are given in Table 6.4. For the full list of correct and incorrect structures, see Table 6.2. It is obvious that in the case of correct spectra, the average rank number should be low when the compared spectra pairs differ in only one educt module. A high rank number is expected if the spectra pairs differ in all of the three modules. This is confirmed by the results. For example, spectrum no. 1 of Table 6.4 receives an average rank of 19 when it is compared with the spectra of products that differ in only one module, and a rank of 68 in case of differing in all three modules (2nd column). It can be observed that these ranks are quite far apart from each other (difference of 49). On the other hand, in the case of incorrect spectra, the values of the average ranking are much closer to each other. For instance, spectrum no. 4 of Table 6.3 achieves average ranks of 37 and 52 when the compared spectra differ in one and three educt modules, respectively. Here, the difference is only 15 in contrast to the above mentioned 49 in the case of a correct spectrum.

Table 6.3. Average ranks (for one and three distinct module cases) obtained for the 21 incorrect spectra when the measured product spectra are compared with every other one.

Measured spectrum no.	Number of distinct modules	Average similarity	Average rank	Product
3	1	0.51	27	A1B3C1
3	3	0.41	62	A1B3C1
4	1	0.32	37	A1B4C1
4	3	0.27	52	A1B4C1
7	1	0.47	35	A1B3C2
7	3	0.36	56	A1B3C2
20	1	0.40	53	A3B4C1
20	3	0.41	47	A3B4C1
31	1	0.58	15	A4B3C2
31	3	0.33	66	A4B3C2
39	1	0.59	18	A5B3C2
39	3	0.40	62	A5B3C2
43	1	0.46	29	A6B3C1
43	3	0.33	57	A6B3C1
45	1	0.56	16	A6B2C2
45	3	0.38	64	A6B2C2
47	1	0.56	33	A6B3C2
47	3	0.46	54	A6B3C2
60	1	0.34	42	A2B4C3
60	3	0.32	56	A2B4C3
64	1	0.33	38	A2B4C4
64	3	0.28	58	A2B4C4
68	1	0.41	51	A3B4C3
68	3	0.42	49	A3B4C3
71	1	0.19	42	A3B3C4
71	3	0.19	54	A3B3C4
74	1	0.50	33	A4B1C3
74	3	0.40	55	A4B1C3
75	1	0.59	12	A4B3C3
75	3	0.30	68	A4B3C3
83	1	0.63	13	A5B3C3
83	3	0.38	66	A5B3C3
87	1	0.59	19	A5B3C4
87	3	0.36	65	A5B3C4
88	1	0.55	23	A5B4C4
88	3	0.44	59	A5B4C4
91	1	0.55	14	A6B3C3
91	3	0.32	64	A6B3C3
93	1	0.58	27	A6B2C4
93	3	0.47	58	A6B2C4
94	1	0.39	38	A6B1C4
94	3	0.35	53	A6B1C4

In order to visualize the above presented results, the measured spectra were clustered on the basis of the calculated similarities. This was made by the built-in hierarchical clustering method of the SPSS 11.0 statistical software package.[138] Figure 6.6 presents the dendrogram of the clustering using Ward's method (described in Section 3.2.2.2) with the squared Euclidean distance measure. Besides the spectra ID, the structure configuration and the spectra quality are given: correct (__), incorrect (_X), and impure (_O).

Table 6.4. Average ranks (for one and three distinct module cases) obtained for the 35 correct spectra when the measured product spectra are compared with every other one.

Measured spectrum no.	Number of distinct modules	Average similarity	Average rank	Product
1	1	0.65	19	A1B2C1
1	3	0.37	68	A1B2C1
2	1	0.65	20	A1B1C1
2	3	0.38	67	A1B1C1
9	1	0.68	12	A2B2C1
9	3	0.35	66	A2B2C1
10	1	0.69	10	A2B1C1
10	3	0.36	66	A2B1C1
11	1	0.63	17	A2B3C1
11	3	0.37	63	A2B3C1
14	1	0.62	10	A2B1C2
14	3	0.34	67	A2B1C2
16	1	0.57	16	A2B4C2
16	3	0.33	67	A2B4C2
17	1	0.66	12	A3B2C1
17	3	0.35	66	A3B2C1
18	1	0.66	15	A3B1C1
18	3	0.37	66	A3B1C1
19	1	0.54	30	A3B3C1
19	3	0.37	63	A3B3C1
22	1	0.62	10	A3B1C2
22	3	0.34	67	A3B1C2
25	1	0.68	12	A4B2C1
25	3	0.39	65	A4B2C1
26	1	0.68	12	A4B1C1
26	3	0.39	64	A4B1C1
30	1	0.61	15	A4B1C2
30	3	0.40	65	A4B1C2
32	1	0.61	16	A4B4C2
32	3	0.42	61	A4B4C2
33	1	0.71	12	A5B2C1
33	3	0.41	66	A5B2C1
34	1	0.72	11	A5B1C1
34	3	0.41	64	A5B1C1
35	1	0.62	15	A5B3C1
35	3	0.35	66	A5B3C1
38	1	0.61	12	A5B1C2
38	3	0.40	66	A5B1C2
48	1	0.60	14	A6B4C2
48	3	0.42	64	A6B4C2
49	1	0.63	11	A1B2C3
49	3	0.43	65	A1B2C3
53	1	0.67	11	A1B2C4
53	3	0.45	65	A1B2C4
56	1	0.55	31	A1B4C4
56	3	0.43	62	A1B4C4
58	1	0.58	19	A2B1C3
58	3	0.40	65	A2B1C3
61	1	0.60	20	A2B2C4
61	3	0.41	66	A2B2C4
62	1	0.60	23	A2B1C4
62	3	0.42	66	A2B1C4
63	1	0.58	25	A2B3C4
63	3	0.43	62	A2B3C4
65	1	0.59	14	A3B2C3
65	3	0.40	66	A3B2C3
73	1	0.61	18	A4B2C3
73	3	0.40	63	A4B2C3
78	1	0.62	20	A4B1C4
78	3	0.43	63	A4B1C4
79	1	0.61	15	A4B3C4
79	3	0.33	69	A4B3C4
85	1	0.67	13	A5B2C4
85	3	0.46	64	A5B2C4
89	1	0.64	18	A6B2C3
89	3	0.45	64	A6B2C3
90	1	0.61	26	A6B1C3
90	3	0.45	63	A6B1C3
92	1	0.61	26	A6B4C3
92	3	0.45	61	A6B4C3

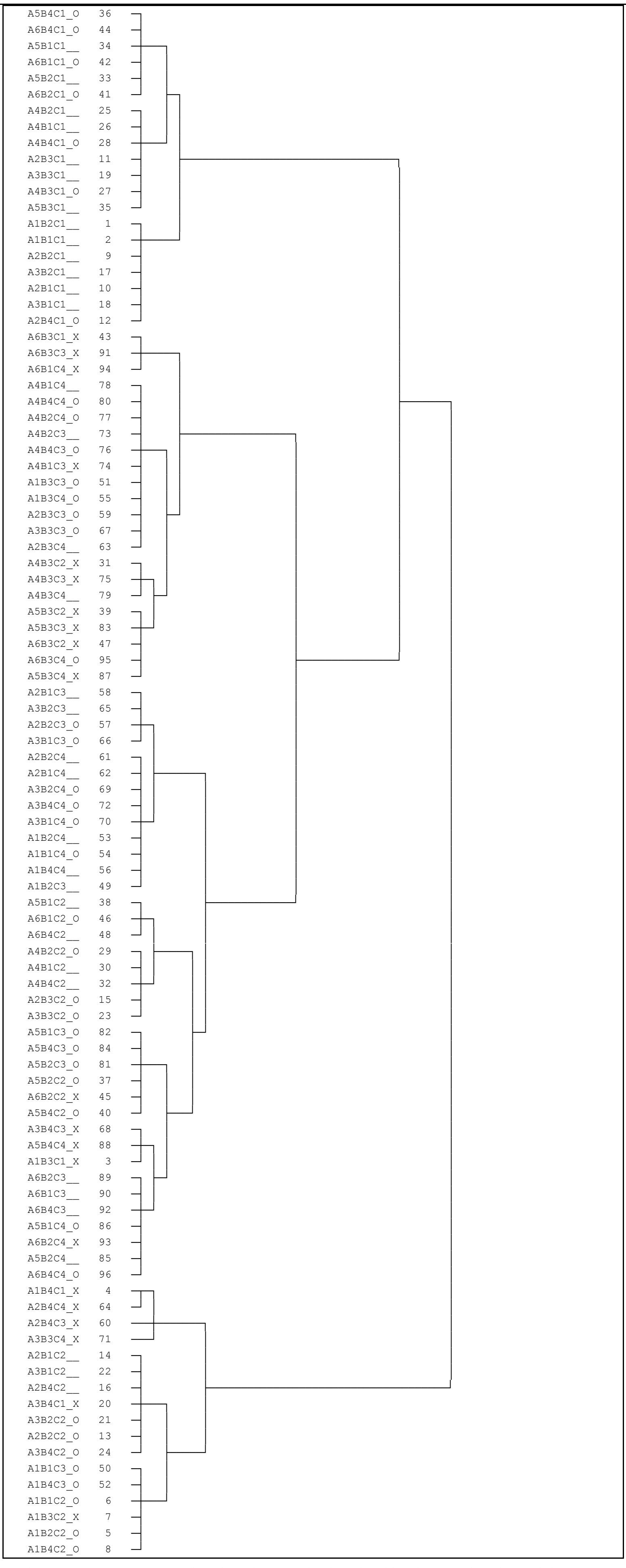


Figure 6.6. Clustering the structures on the basis of their similarity calculated between each spectrum. Quality of spectra: correct (__), incorrect (_X) and impure (_O).

It can be observed that there are clear clusters based on a single educt module or on the quality of the spectra. For example, the first 3rd-level cluster contains only correct and a few impure spectra with structures having a common C type educt (i.e., C1). The last 1st-level cluster contains impure and incorrect spectra and the structures include a common A1 module.

Table 6.5. The ranks obtained for the 35 correct and 21 incorrect spectra. The measured product spectra were compared with the sum of the measured spectra of the educts.

Artificial spectrum no.	Measured spectrum no.	Similarity	Rank	Product	Quality of spectrum
1	1	0.59	3	A1B2C1	correct
2	2	0.59	2	A1B1C1	correct
9	9	0.60	1	A2B2C1	correct
10	10	0.60	1	A2B1C1	correct
11	11	0.61	4	A2B3C1	correct
14	14	0.59	1	A2B1C2	correct
16	16	0.59	2	A2B4C2	correct
17	17	0.61	1	A3B2C1	correct
18	18	0.60	2	A3B1C1	correct
19	19	0.64	2	A3B3C1	correct
22	22	0.57	2	A3B1C2	correct
25	25	0.61	7	A4B2C1	correct
26	26	0.61	4	A4B1C1	correct
30	30	0.60	3	A4B1C2	correct
32	32	0.58	5	A4B4C2	correct
33	33	0.64	3	A5B2C1	correct
34	34	0.65	1	A5B1C1	correct
35	35	0.66	1	A5B3C1	correct
38	38	0.66	2	A5B1C2	correct
48	48	0.60	2	A6B4C2	correct
49	49	0.55	2	A1B2C3	correct
53	53	0.54	3	A1B2C4	correct
56	56	0.55	3	A1B4C4	correct
58	58	0.55	2	A2B1C3	correct
61	61	0.52	1	A2B2C4	correct
62	62	0.52	1	A2B1C4	correct
63	63	0.57	3	A2B3C4	correct
65	65	0.58	2	A3B2C3	correct
73	73	0.60	3	A4B2C3	correct
78	78	0.54	4	A4B1C4	correct
79	79	0.63	3	A4B3C4	correct
85	85	0.57	2	A5B2C4	correct
89	89	0.64	1	A6B2C3	correct
90	90	0.60	1	A6B1C3	correct
92	92	0.59	2	A6B4C3	correct
3	3	0.56	10	A1B3C1	incorrect
4	4	0.27	75	A1B4C1	incorrect
7	7	0.41	49	A1B3C2	incorrect
20	20	0.33	37	A3B4C1	incorrect
31	31	0.63	1	A4B3C2	incorrect
39	39	0.57	2	A5B3C2	incorrect
43	43	0.45	37	A6B3C1	incorrect
45	45	0.54	11	A6B2C2	incorrect
47	47	0.56	5	A6B3C2	incorrect
60	60	0.36	41	A2B4C3	incorrect
64	64	0.34	53	A2B4C4	incorrect
68	68	0.37	37	A3B4C3	incorrect
71	71	0.26	91	A3B3C4	incorrect
74	74	0.42	52	A4B1C3	incorrect
75	75	0.66	1	A4B3C3	incorrect
83	83	0.65	1	A5B3C3	incorrect
87	87	0.62	2	A5B3C4	incorrect
88	88	0.70	1	A5B4C4	incorrect
91	91	0.58	5	A6B3C3	incorrect
93	93	0.54	4	A6B2C4	incorrect
94	94	0.40	51	A6B1C4	incorrect

Next, the measured product spectra were compared with the sum of the measured spectra of the corresponding educts (method presented in Section 6.3.3). Each artificially built spectrum was compared with each of the measured product spectra and the comparisons were ranked by descending similarity. The highest similarity (lowest ranking) should be achieved when the measured spectrum is compared with the corresponding artificial spectrum.

Table 6.6. Ranks obtained for the 35 correct and 21 incorrect spectra. The measured spectra are compared with those of the individual educts.

Product spectrum no.	Educt spectrum no.	Similarity	Rank	Module	Quality of spectrum
1	5	0.49	1	C1	correct
2	5	0.48	1	C1	correct
9	5	0.51	1	C1	correct
10	5	0.50	1	C1	correct
11	5	0.42	1	C1	correct
14	6	0.49	1	C2	correct
16	6	0.47	1	C2	correct
17	5	0.52	1	C1	correct
18	5	0.52	1	C1	correct
19	5	0.47	1	C1	correct
22	6	0.47	1	C2	correct
25	5	0.44	1	C1	correct
26	5	0.45	1	C1	correct
30	6	0.40	1	C2	correct
32	6	0.39	1	C2	correct
33	5	0.47	1	C1	correct
34	5	0.45	1	C1	correct
35	5	0.40	1	C1	correct
38	6	0.44	1	C2	correct
48	6	0.43	1	C2	correct
49	7	0.41	1	C3	correct
53	8	0.40	1	C4	correct
56	8	0.47	1	C4	correct
58	7	0.42	1	C3	correct
61	8	0.41	1	C4	correct
62	8	0.40	1	C4	correct
63	8	0.33	1	C4	correct
65	7	0.43	1	C3	correct
73	7	0.34	1	C3	correct
78	8	0.34	1	C4	correct
79	8	0.31	1	C4	correct
85	8	0.34	1	C4	correct
89	7	0.37	1	C3	correct
90	7	0.37	1	C3	correct
92	7	0.38	1	C3	correct
3	5	0.29	1	C1	incorrect
4	5	0.10	3	C1	incorrect
7	6	0.32	1	C2	incorrect
20	5	0.21	3	C1	incorrect
31	6	0.28	1	C2	incorrect
39	6	0.24	1	C2	incorrect
43	5	0.14	3	C1	incorrect
45	6	0.25	1	C2	incorrect
47	6	0.19	4	C2	incorrect
60	7	0.41	2	C3	incorrect
64	8	0.14	2	C4	incorrect
68	7	0.30	2	C3	incorrect
71	8	0.11	1	C4	incorrect
74	7	0.30	2	C3	incorrect
75	7	0.25	1	C3	incorrect
83	7	0.27	1	C3	incorrect
87	8	0.26	1	C4	incorrect
88	8	0.40	1	C4	incorrect
91	7	0.18	1	C3	incorrect
93	8	0.30	1	C4	incorrect
94	8	0.17	2	C4	incorrect

Table 6.5 lists the ranks (4th column) for the 35 correct and 21 incorrect spectra. It can be seen that in the case of correct spectra, the ranks are low (the highest is 7), while with incorrect spectra, the ranks are much higher (even 91 for spectrum no. 71). Nevertheless, there are some incorrect spectra that receive a very low ranking (e.g., spectra no. 31, 75, 83, and 88 have a ranking number of 1). Note that the overall highest possible rank is 96 (i.e., the number of product spectra).

Finally, the measured spectra were compared with the spectra of the individual educts (for details, see Section 6.3.4). Each product spectrum was compared with each of the educt spectra and the comparisons were ranked by descending similarity. The highest similarity (lowest rank) should be achieved when the spectrum of the product is compared with that of an educt contained in the product. The 4th column of Table 6.6 lists the ranks in the case of the 35 correct and 21 incorrect spectra, showing only the comparisons with the spectra of type C educts. It is observed that for each correct spectrum, the rank is 1. Ideally, the ranks for incorrect spectra should be higher than 1. Unfortunately, there is a row of incorrect spectra that achieve the ranking number of 1, thus, these spectra are classified wrongly.

After comparing the measured spectra of products with the four types of generated reference spectra (see Section 6.3), the automatic spectra verification system was constructed. For each product spectrum of the library, an answer in a traffic light style is obtained regarding the quality of the spectra, i.e., in the sense of their correctness. For this purpose, the optimal thresholds have to be defined for each of the above presented four approaches. In order to find the optimal thresholds, the false positive and false negative cases were calculated for different threshold values, leading to contingency diagrams (introduced in Section 3.9.1). Figure 6.7 presents the false positives and false negatives for different threshold values when the measured product spectra are compared with different reference spectra. The conditions for achieving a false positive or a false negative situation in the case of the four methods is described in Section 6.3. The aim is to get the lowest possible false positives and false negatives at the same time, i.e., to minimize the sum of the two cases. Evidently, the impure spectra were excluded from the calculations because it was not possible to decide whether they are correct or incorrect and, hence, cannot be counted as false positives or false negatives. Under these circumstances, the optimal threshold values for the four methods described in Sections 6.3.1, 6.3.2, 6.3.3, and 6.3.4 were found to be: a) 0.57, b) 36, c) 4, and d) 1, respectively. Note that Figure 6.7d shows the performance based on the comparison with the spectra of type C educts. In this case, the maximum threshold is 4, as four different type C educts are used. It is seen that there are no false negatives when this method is applied.

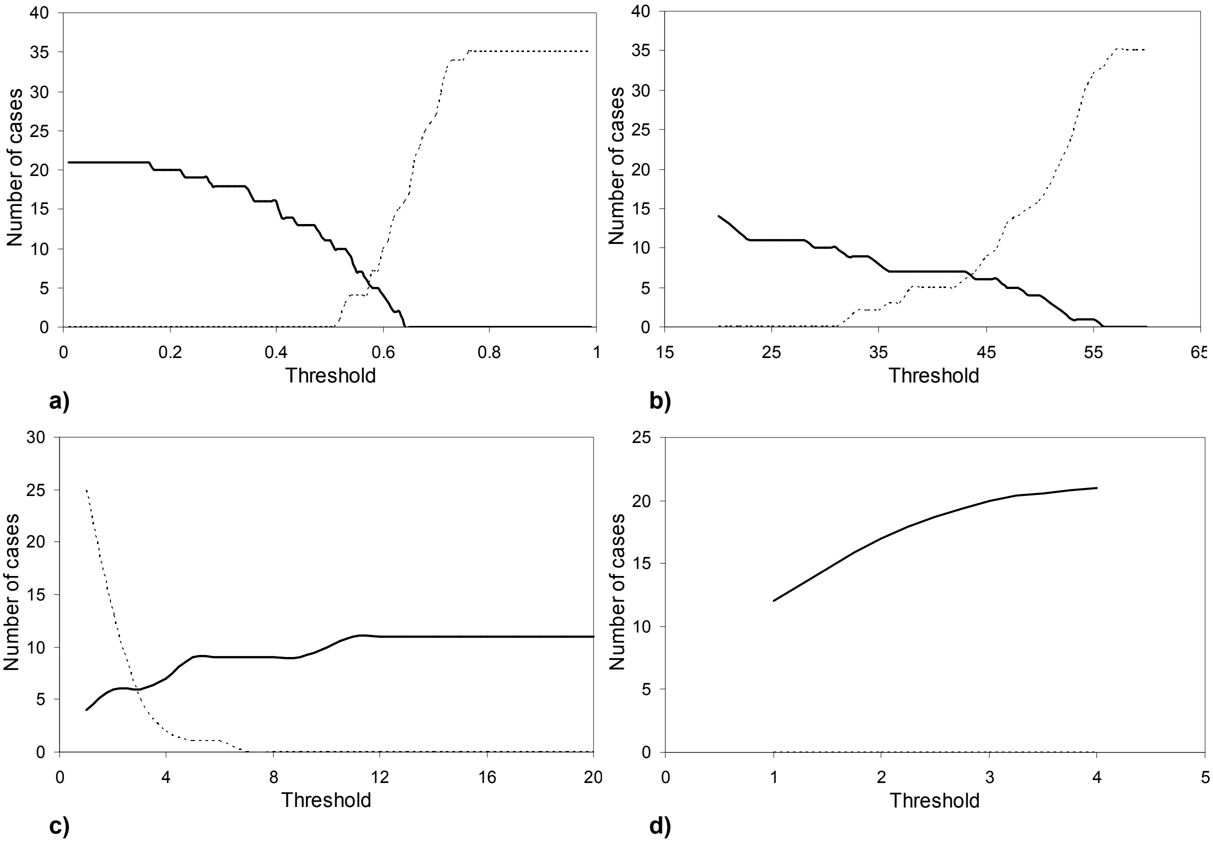


Figure 6.7. Number of false positives (solid line) and false negatives (dashed line) using different thresholds when the measured spectra are compared with: a) the corresponding predicted ones, b) every other measured product spectrum, c) combined spectra of educts, and d) the individual measured spectra of type C educts. The impure spectra (cf. Table 6.2) were excluded from the calculations.

The final responses obtained with the above methods are given in Figure 6.8. The thresholds used for evaluation were chosen on the basis of contingency diagrams in order to simultaneously retain the lowest possible number of false positives and false negatives. Thus, the thresholds by the four methods, when the measured spectra are compared with the reference ones, are as follows: a) 0.57, b) 36, c) 4, and d) 1. The number of false positives and false negatives are listed in Table 6.7 for each method separately. It is seen that Figure 6.8 also shows the outcome for the impure spectra, although they were not included in the calculation of the false positives and false negatives, as obviously no decision can be made whether they are correct or incorrect.

If the aim is to achieve the lowest possible number of false positives, then, other threshold values must be used. In this case, the responses are shown in Figure 6.9, while Table 6.8 gives the number of false positives and false negatives separately for each method. Here, the threshold values used with the four methods are defined as follows: a) 0.64, b) 56, c) 1, and d) 1. Note that in the case of the fourth method, the threshold is the same as above for Table 6.7, hence, the number of false positives and false negatives is the same too.

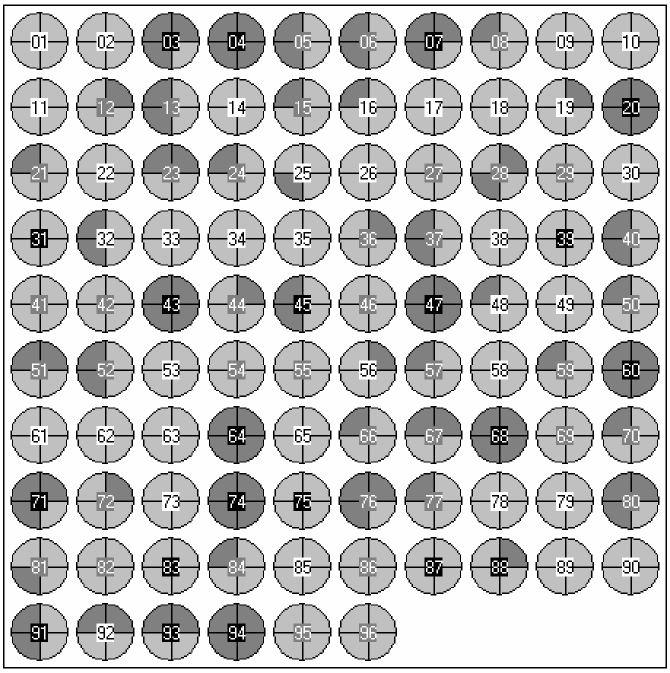


Figure 6.8. Final responses (light gray: OK, dark gray: not OK) for each spectrum of the library given by each method when the measured spectra are compared with: a) estimated spectra (upper left section), b) other spectra of the combinatorial library (upper right section), c) combined spectra of the educts (lower left section), and d) individual spectra of educts (lower right section). The thresholds used are: a) 0.57, b) 36, c) 4, and d) 1. The actual quality of spectra is also visualized: white background for correct spectra, black for incorrect, and gray for impure ones. (cf. text)

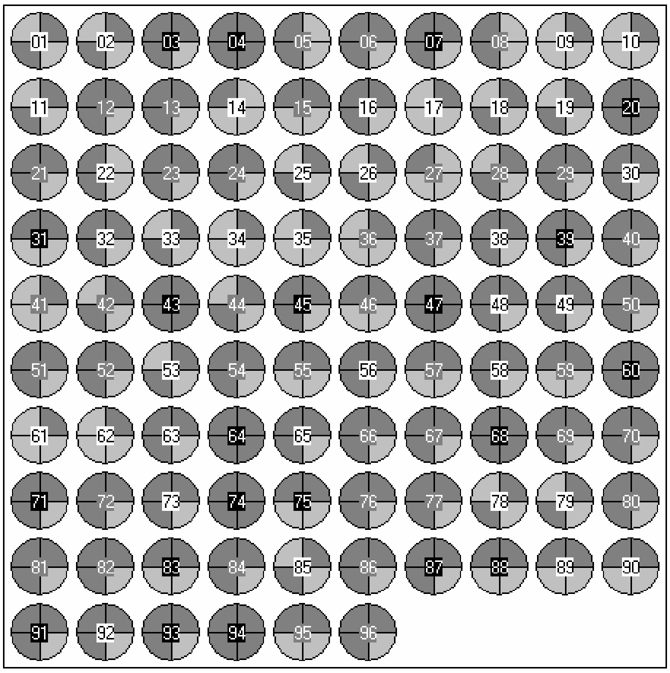


Figure 6.9. Final responses (light gray: OK, dark gray: not OK) for each spectrum of the library given by each method when the measured spectra are compared with: a) estimated spectra (upper left section), b) other spectra of the combinatorial library (upper right section), c) combined spectra of the educts (lower left section), and d) individual spectra of educts (lower right section). The thresholds used are: a) 0.64, b) 56, c) 1, and d) 1. The actual quality of spectra is also visualized: white background for correct spectra, black for incorrect, and gray for impure ones. (cf. text)

Table 6.7. Number of false positives and false negatives for each comparison method based on the final responses shown in Figure 6.8. The thresholds were so defined as to achieve the lowest number of false positives and false negatives at the same time.

	a)	b)	c)	d)
False positives	6	7	7	12
False negatives	4	3	2	0

Table 6.8. Number of false positives and false negatives for each comparison method based on the final responses shown in Figure 6.9. The thresholds were so defined as to achieve the lowest number of false positives with each method.

	a)	b)	c)	d)
False positives	0	0	4	12
False negatives	16	33	25	0

6.5 Conclusions

In this chapter, the test set was a combinatorial library containing very similar chemical structures (generated from the same educt structures) and spectra. Also under this condition, the bin method is able to cope with the problem. Since the library is complex, several reference spectra could be generated with which the measured spectra were compared. Gross errors in the test set were easily detected with all these procedures. However, none of them alone was capable of achieving high values of true positives and true negatives simultaneously. The combination of the various methods provides the most promising approach.

7 Similarity of HSQC spectra

The similarity measure introduced in Chapter 4 is extended here to the use with two-dimensional spectra. Because of their high practical relevance, the heteronuclear single quantum correlation (HSQC) spectra were chosen. For testing the compatibility of such a spectrum with the proposed chemical structure, first, the spectrum is predicted on the basis of that structure. Then, it is compared with the measured one. In this context, topics of optimization are the automatic peak picking, the use of various signal intensity measures, and the optimization of the two-dimensional bin method.

7.1 Introduction

Since their introduction in the 1980s, several hundreds of different two-dimensional (2D) NMR methods have been proposed. Only a few of them are routinely used in analytical laboratories. Such spectra are obtained by recording a series of interferograms, $I = f(t_2)$, by stepwise changing a second time-dependent parameter (t_1). First, a series of spectra are obtained by Fourier transformation. Then, a second Fourier transformation produces the 2D spectrum, in which the intensity depends on two frequencies. Thus, in 2D NMR spectra, the signal intensities are plotted as a function of the two frequencies, F_1 and F_2.

In the HSQC or ^{13}C-^{1}H-correlated spectra, the F_1 and F_2 axes correspond to the ^{13}C and ^{1}H chemical shifts, respectively. The signals are singlets and do not provide any coupling information. The HSQC is a so-called polarization transfer method. This means that the ^{13}C NMR information is obtained on the basis of magnetization of the ^{1}H nuclei, which have a sensitivity that is about 6000 times higher than the ^{13}C nuclei. Hence, the time required to register an HSQC spectrum is much smaller than for a 1D ^{13}C NMR spectrum. This explains the high practical relevance and popularity of this method. In fact, in many laboratories, the NMR methods used for structure verification are 1D ^{1}H NMR and 2D HSQC spectra. Therefore, an automatic verification system as proposed in this work is of high practical relevance.

However, HSQC spectra have also several drawbacks. Signal intensities depend on the transfer of magnetic polarization from protons to ^{13}C atoms making use of the coupling of the two nuclei over one bond. Quaternary carbon atoms, which do not have such coupling partners, are invisible in such spectra. Moreover, the coupling constant between ^{1}H and ^{13}C may vary in a large range (about 100–250 Hz). Even if the extreme cases, i.e., carbon atoms with triple bonds and unusual substituents, are excluded, the range of coupling constants still is about 120–200 Hz. As the method is based on one single coupling constant (usually assumed to be around 150), the efficiency of the polarization transfer and, thus, the intensity of the signals vary. Since signal intensities contain highly relevant structural information, various methods have been tested in this work to regenerate them artificially.

Another problem of HSQC spectra is that due to imperfections in the excitation process, various artifacts may occur. To reduce them, a number of heuristic techniques have been tested here. Finally, to save time, routine spectra often have poor signal to noise ratios so that it is not trivial to automatically distinguish signals from the noise. Also for this purpose, various algorithms have been used in this work.

In NMR spectroscopy, HSQC spectra are very up-to-date experiments, though only few articles have been published involving them. Being a relatively cheap and quick experiment, the HSQC is useful to screen candidates for structure determination by NMR. These spectra are particularly useful in the field of protein NMR spectroscopy.[139] One-dimensional protein spectra are far too complex for interpretation as most of the signals overlap heavily. By the introduction of additional spectral dimensions, these spectra are simplified and some extra information is obtained.

Automatic evaluation of chemical structures using both ^{1}H NMR and HSQC spectra is a very important and promising research field.[140, 141] In both these articles, a low false positive rate is reported, but the results are not statistically significant as the data sets are small: only 14 compounds and corresponding spectra in the first contribution[140] and 25 structures with spectra in the second one.[141]

7.2 Finding the signals in the measured HSQC spectra

One of the major advantages of the HSQC spectra is the short recording time, i.e., they are recorded faster than one-dimensional ^{13}C NMR spectra. Another advantage is that the ^{1}H/^{13}C shift information is available simultaneously and X–H signals are not present in the spectra. Nevertheless, HSQC spectra have also some disadvantages. It is obvious that there is no coupling information and no signals for quaternary carbon atoms present. But, the major problem with HSQC spectra lies in that it is very difficult to find the signals. The intensity of signals highly depends on J_{CH}, parameter setting, and artifacts in the spectra. Artifacts can occur during recording and are due to impurities or other types of errors and disturbing factors. The heteronuclear correlation experiments suffer from noise artifacts that appear as vertical streaks.

Basically, there are two problems with HSQC signals:

1. finding the signals and
2. calculating the signal intensities.

The aim is to find an optimal threshold that clearly separates the signals from the noise. It is important not to retrieve noise as signal and not to lose signals. It is expected that signals are more intense than noise, thus, signals are defined on the basis of their intensities and ratios of signal intensities. There are two ways of representing the intensities of the identified signals: using either a uniform intensity of 1 or the integral of the corresponding ^{1}H NMR signal as the intensity measure.

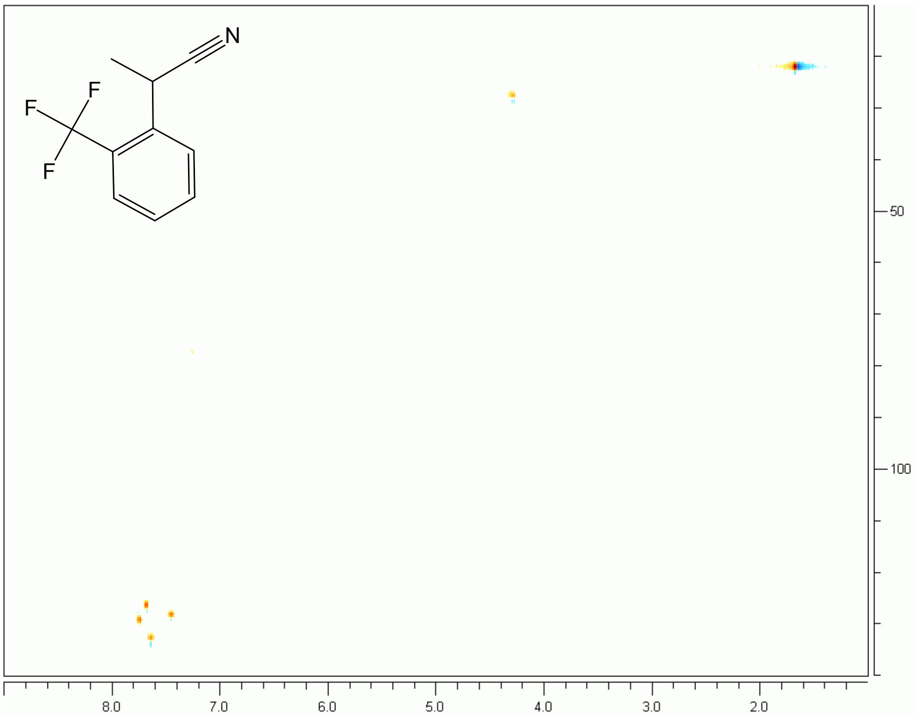


Figure 7.1. Example of a structure and the corresponding HSQC spectrum visualized with the MestReC 4.5.6 software.[142] The proton and carbon chemical shifts are on the x and y axes, respectively. This is the original spectrum; the solvent signal and artifacts have not yet been removed.

The HSQC spectra used here are stored in Bruker format. Processing and acquisition parameters are saved in text files, while the intensity values of the spectrum are expressed in a binary file and grouped into several blocks of various dimensions for fast zooming. After the content of the binary file is read into a matrix, the signals have to be defined. For an example, see Figure 7.1 and Figure 7.2 in which the same HSQC spectrum is visualized with the MestReC 4.5.6 software[142] and with our internally developed software, respectively. The signal at (1.68, 22.26) ppm is broad, and our method identifies three different signals. In the figures, the raw, unprocessed spectra are shown. The solvent signal at (7.26, 77.60) ppm and the two artifact signals accompanying the signal at (1.68, 22.26) ppm are to be removed automatically. This is a challenging task and the methods developed for these purposes are presented in the following sections.

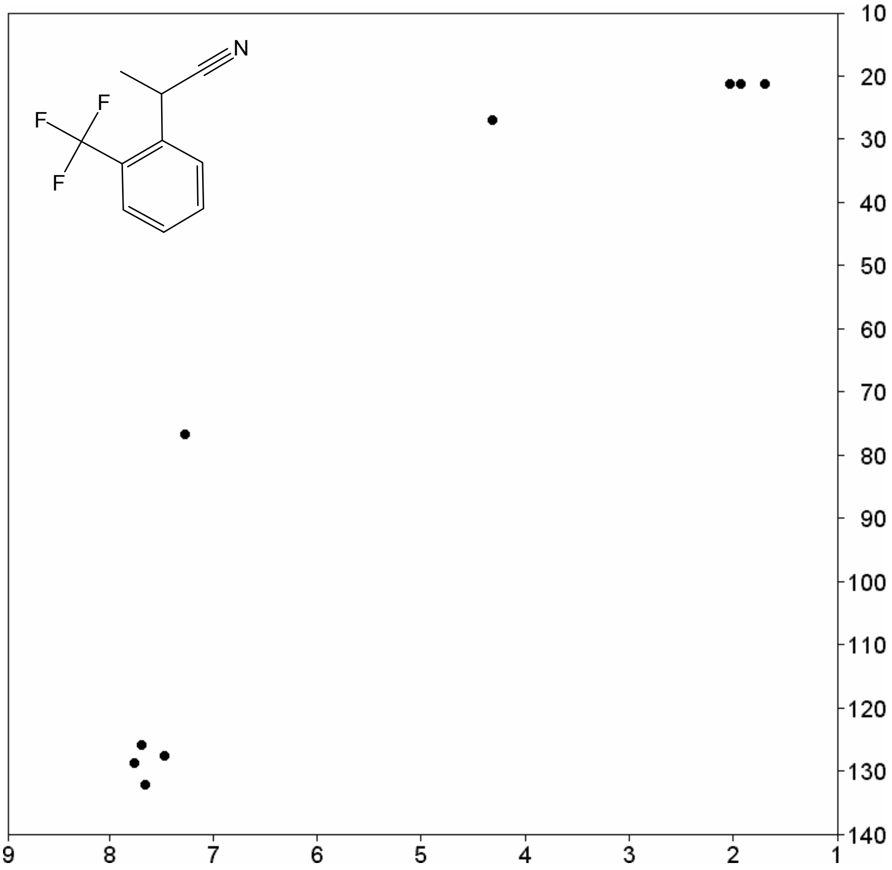


Figure 7.2. Example of a structure and the corresponding HSQC spectrum visualized with our internally developed software. On the x axis are the proton shifts and on the y axis, the carbon shifts. Also here, the solvent signal and artifacts have not yet been removed.

7.2.1 Using uniform intensity values of 1

As discussed above, the aim is to find the signals in the HSQC spectrum excluding noise or artifact signals. For this purpose, a threshold level is defined that clearly separates signals from noise. Then, the retrieved signals are analyzed and based on their intensities, some of them are removed. Finally, the uniform intensity value of 1 is assigned to the remaining HSQC signals. This is the easiest and most straightforward method for HSQC signal identification.

It was observed empirically that the signal-to-noise ratio very often is too low so that it is impossible to correctly identify all signals in an HSQC spectrum. Therefore, the Bruker TopSpin 1.3 processing software[143] was used to manually tune the parameters of spectra in order to increase the signal-to-noise ratio. The parameters of the so-called window function and linear prediction were customized. The signal-to-noise ratio can be increased by using a narrower window at the cost of lower digital resolution, while the linear prediction procedure improves the resolution of 2D spectra. After several trials, it can be stated that the above software in some cases is useful but the optimal method and optimal parameters may be different for individual cases (spectra). Thus, the method is difficult to automate and, consequently, will not be used.

The algorithm (here called Method1) that identifies the signals in an HSQC spectrum using the above described principals, is given in Figure 7.3. First, the intensities of the spectrum are normalized to the range of 0–1 by dividing them by the maximal intensity value. This is important in order to define a threshold range, which can have values between 0 and 1. In the next step, the average intensity value is calculated in the area of the upper left and bottom right corners of the HSQC spectrum in a range of 0.05 ppm/11 digital points (direction of ^{1}H NMR) by 4 ppm/9 digital points (direction of ^{13}C NMR), assuming that in these two regions, no signals are present. This is an important step to estimate the noise, which will serve as an initial value for the threshold. This calculated average noise will be the starting value of the threshold that defines the signals in the spectrum. The level of the threshold is increased in small steps (0.001) until so-called isolated signals are present. These are signals that are only one digital point above the threshold value, surrounded by points having intensities that are below the threshold level.

When recording HSQC spectra, artifacts can occur due to impurities or other types of errors and disturbing factors. They can appear in the form of a vertical line with strong signals along the same ^{1}H NMR chemical shift axis. In order to tackle this problem, only the most intense signal is kept along this coordinate, using a tolerance of ± 0.016 ppm = 3 digital points for the ^{1}H NMR coordinates. The minimal ^{13}C NMR distance between two signals along the same ^{1}H NMR value has to be > 3.7 ppm = 10 digital points (direction of ^{13}C NMR). Note that this is a heuristic approach that may lead to a loss of signals.

By continuously increasing the threshold value, large signals can be divided in two or more “islands” of signals. This is an artifact since, actually, only one signal is present in such a small region and it is avoided by keeping only the most intense signal in a range of 0.06 ppm

= 10 digital points (^{1}H NMR) and 1.2 ppm = 3 digital points (^{13}C NMR).

It has been observed that the above calculated threshold value will retain more peaks than there are in reality. Thus, statistics are performed on the signals retained considering the maximum intensity value within each signal, the number of digital points of each signal, and the sum of the intensities within a signal (for an example, see Table 7.1). With the help of this information, the algorithm is fine-tuned, and in the second part, further signals that are not peaks, are excluded. Based on empirical results, only those signals are kept whose summed up intensity is greater than 1/18th of the average sum of intensities of all remaining signals. Finally, the signal is converted to a single digital point located at the maximum intensity value of each signal, and its intensity is assigned the value of 1. Thus, a 0 and 1 valued matrix is generated with the value 1 representing signals.

```
• 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 $^1$H NMR value
• keep only the strongest signal in a range of 0.06 ppm = 10 digital points ($^1$H NMR) and 1.2 ppm = 3 digital points ($^{13}$C NMR)
• keep only those signals whose summed intensity is greater than the 1/18th of the average sum of intensities of all remaining signals
• convert to pointwise signals: the remaining signals will be represented by the value of 1 in a matrix storing the spectrum
```

Figure 7.3. Method1 algorithm: Signal definition for HSQC spectra. The calculated threshold is used to define the signals, which in the second part of the algorithm are analyzed further. Each identified signal will be represented with the value of 1 in the matrix.

The algorithm is demonstrated with the HSQC spectrum shown in Figure 7.2. The found threshold value is used to define the signals listed in Table 7.1. It can be seen that some intensities are not signals but artifacts. Applying the fine-tuning procedures presented in the second part of the algorithm, signals no. 2 and 3 are excluded. Evidently, these signals have the lowest sum of their intensity values, thus, their ratio to the average sum of intensities is very low, i.e., below the given threshold. Signal no. 5 is not retained either as it is the signal of $CHCl_3$ (impurity in the solvent $CDCl_3$). The final measured HSQC spectrum is displayed in Figure 7.4.

Table 7.1. Statistics made on the retained signals of the measured HSQC spectrum in Figure 7.2. Signal no. 5 from the solvent impurity $CHCl_3$ is removed, so are signals no. 2 and 3 because of their low intensity.

Signal No.	^1H NMR (ppm)	^{13}C NMR (ppm)	Number of digital points of the signal	Maximum intensity value of the signal	Sum of the intensities of the signal
1	1.68	22.26	118	1.000	18.945
2	1.90	22.26	2	0.043	0.161
3	2.01	22.26	3	0.042	0.116
4	4.28	27.83	40	0.231	3.872
5	7.26	77.60	5	0.063	0.252
6	7.67	127.63	36	0.456	6.821
7	7.44	128.49	44	0.399	5.939
8	7.75	129.60	38	0.446	6.342
9	7.64	132.94	45	0.325	5.230

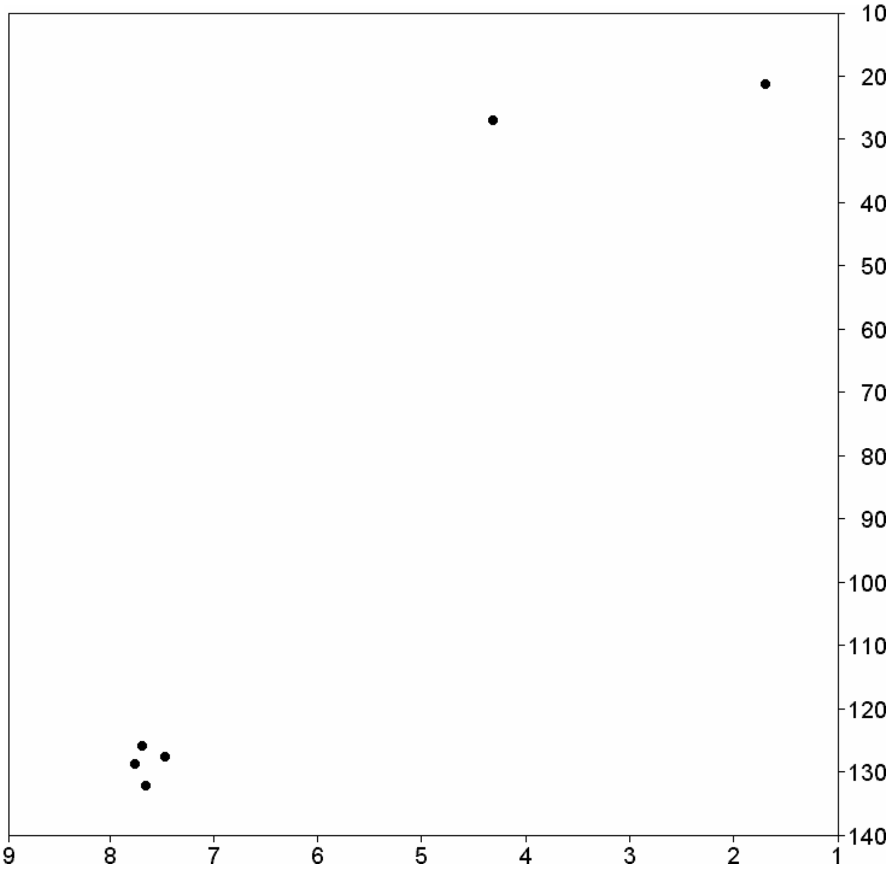


Figure 7.4. The final measured and processed HSQC spectrum. Each signal has an intensity of 1. Signals were identified by the Method1 algorithm, which is summarized in Figure 7.3.

7.2.2 Counting the number of protons for the measured HSQC signals

Since the measured signal intensities in HSQC spectra are arbitrary, a method is proposed here to calculate signal intensities that are proportional to the number of protons. This will allow a more powerful comparison of measured and predicted spectra.

The problem with the above described method is that the HSQC intensities are distorted since at this stage, each signal has the intensity of 1, i.e., the same weighting, although in reality, it is not the case. In order to tackle this issue, the true intensities are estimated on the basis of the corresponding ^{1}H NMR spectrum. Moreover, from the estimated spectrum, intensity information is available.

Thus, for each measured HSQC signal, the integral of the corresponding ^{1}H NMR signal is used as intensity measure instead of the value 1. For this purpose, three different methods were developed and are presented in the following sections.

In the case of the predicted HSQC signals, the intensities are replaced with the number of protons. This is expected to increase the selectivity of the method and to better discriminate between similar and foreign spectra.

7.2.2.1 Using the integral of the corresponding ^{1}H NMR signals: Integrating the intensities within the defined edges

The algorithm, called Method2, developed to count the number of protons for measured HSQC signals is listed in Figure 7.5. Here, we also use the ^{1}H NMR spectrum belonging to the same compound as the HSQC spectrum in question. Thus, the ^{1}H NMR spectrum is read from file and preprocessed: First, the integrated intensities are normalized to the 0–1 range, then, the solvent and noise are eliminated and, finally, the integrated intensities are normalized to the number of protons. Next, the HSQC spectrum is processed. After the signals have been identified in a HSQC spectrum and represented in the spectrum matrix with the value of 1 using the Method1 algorithm (Figure 7.3), the signals of the solvent impurities (dimethyl sulfoxide-d_5 (DMSO-d_5) or $CHCl_3$, see Section 7.3) are removed. Then, the ^{1}H NMR chemical shifts of HSQC signals are noted (i.e., the signals are projected on the x axis) and ordered by descending ppm positions. For each of these shifts, the corresponding signal is searched in the proton NMR spectrum: In a given range, first, the signal peak is identified and, then, the edges of the signal are defined. The intensities between these edges are summed up (equal to the integral of the ^{1}H NMR signal) and stored. In the next step, each calculated integral is divided by *(sum + X–HNo)/ProtonNo*, where *sum* is the sum of the integrals of all identified signals, *X–HNo* is the number of X–H, i.e., hydrogens not bonded to carbon, and *ProtonNo* is the total number of protons in the structure. This procedure normalizes the intensities only between the selected ranges. Thus, protons are not lost among impurities, i.e., intensities that belong to impurities are not summed up and, consequently, signals are not divided by this value. In other words, after this normalization procedure, the sum of the intensities between the selected ranges are equal to the total number of protons that are bonded to carbon atoms. Hydrogens not bonded to carbon are not taken into account during normalization as they are not present in HSQC spectra.

Finally, the intensity of the corresponding HSQC signal (currently 1) is replaced by the normalized value of the corresponding integral.

```
• read the $^1$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 using the Method1 algorithm
• remove solvent signals
• get the $^1$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 $^1$H NMR spectrum
    • in a given range (±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 the half
      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
      (e.g., 0.0001)
    • 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: $(sum + X\text{-}HNo)/ProtonNo$,
      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) with the normalized value of the
      corresponding integral
```

Figure 7.5. Method2 algorithm: For each measured HSQC signal, it assigns as intensity the integral of the corresponding ^{1}H NMR signal. At the signal positions in the matrix (which represents the spectrum), the integral of the corresponding ^{1}H NMR signal is stored.

In order to demonstrate and exemplify the above Method2 algorithm, we again use the structure and corresponding HSQC spectrum shown in Figure 7.4. Here, the intensities of the HSQC spectrum are calculated from the corresponding ^{1}H NMR spectrum (see Figure 7.6 and Figure 7.7 for more details). There are six signals in the HSQC spectrum. The corresponding signals in the ^{1}H NMR spectrum are identified and the intensities are added up between the calculated edges of the signals. Then, the sum of the intensities is normalized to the total number of protons (without the hydrogens that are not bonded to carbon).

Unfortunately, with this algorithm, the integrals are not correctly distributed between the signals (see Table 7.2). The reason is that the edges of the signals are not perfectly identified, especially not in the case of overlapping signals (e.g., for signals no. 2 and 3, at 7.67 and 7.64 ppm). In the case of the signal at 1.67 ppm, a too large range is taken into account so that the intensities of impurities are counted as well. It can be observed that for this spectrum, if the rounded values (to the closest integer) of the intensities were used, the exact integral values would be considered. The algorithm described in the following section uses this improvement.

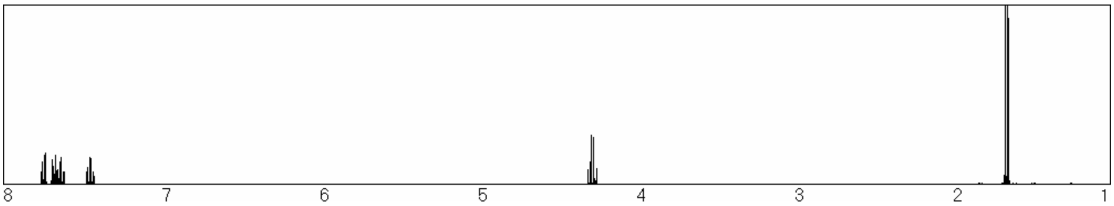


Figure 7.6. ^{1}H NMR spectrum (range 1–8 ppm) belonging to the HSQC spectrum in Figure 7.4.

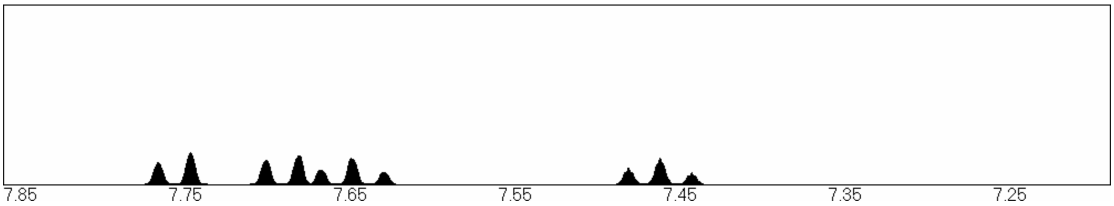


Figure 7.7. Detailed visualization of the signals between 7.85 and 7.17 ppm of the ^{1}H NMR spectrum in Figure 7.6.

Table 7.2. Calculated intensities of the HSQC signals (Figure 7.4) using the corresponding ^{1}H NMR spectrum (Figure 7.6).

Signal No.	HSQC signal (ppm)		Edges of the signal in ^{1}H NMR spectrum (ppm)	Relative intensities
	^{1}H NMR	^{13}C NMR		
1	7.74	129.61	7.77–7.72	0.94
2	7.67	126.63	7.70–7.65	1.22
3	7.64	132.96	7.65–7.61	0.73
4	7.45	128.49	7.49–7.42	0.98
5	4.29	27.70	4.32–4.24	0.97
6	1.67	22.13	1.70–1.64	3.16

7.2.2.2 Using the integral of the corresponding ^{1}H NMR signals: Distributing the integral of the signal groups among the HSQC signals

As observed with the previous Method2 algorithm, the integrals are not correctly distributed between the signals. This is because the chemical shifts are not correctly defined, and the edges of the signals are not perfectly identified.

In order to eliminate these errors, a new algorithm, named Method3 (see Figure 7.8), for calculating the intensities of HSQC signals is presented. As in the previous algorithm, also here, the ^{1}H NMR spectrum is read from file and is preprocessed. The spectrum is divided into groups of signals, i.e., the “islands” of signals are identified. In each group, the neighboring signals are closer than a given threshold (in Hz) to each other. Then, the HSQC spectrum is processed. The signals are projected on the x axis and ordered by descending ppm positions. In the next step, the corresponding ^{1}H NMR signal group is defined for each of these shifts. If at least one HSQC signal belongs to a ^{1}H NMR signal group, then, the integral of the group is calculated. On the other hand, if more than one HSQC signal belongs to a group, the integral of this group is redistributed equally among these signals. Finally, the intensity of the corresponding HSQC signal (currently 1) is replaced with the above value.

```
•  read the $^{1}$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
•  group the signals based on a threshold (range of $x$ Hz)
•  read the HSQC spectrum
•  find the signals using the Method1 algorithm
•  remove solvent signals
•  get the $^{1}$H NMR shifts of HSQC signals (project signals) in ppm
•  order signals by ascending ppm position
•  for each signal do the following steps:
   •  get the position of the signal in the $^{1}$H NMR spectrum
   •  define to which $^{1}$H NMR signal group it belongs
   •  if at least one signal belongs to a group, then calculate
      the integral of the group and store the rounded value
   •  if more than one signal belongs to a group, then
      redistribute equally the integral of the group among these
      signals
   •  replace the intensity of the corresponding HSQC signal
      (currently 1) with the above value
```

Figure 7.8. The Method3 algorithm assigns the integral of the corresponding ^{1}H NMR signal as intensity to each measured HSQC signal.

The algorithm is demonstrated with the same example as above in Figure 7.4 for the HSQC spectrum and Figure 7.6 and Figure 7.7 for the ^{1}H NMR spectrum. The ^{1}H NMR spectrum is divided into groups of signals, in which the neighboring signals are closer than 20 Hz to each other. Thus, four non-overlapping groups are identified and the integrals are calculated (see Table 7.3). It can be seen that group no. 4 contains three HSQC signals. The integral (here 3) is redistributed among these signals, i.e., it is divided by 3 and the integral of 1 is assigned to each of them. Thus, each HSQC signal has an intensity of 1, which is the correct result. Therefore, the similarity between the measured and predicted HSQC spectra is higher than in the case of the Method2 algorithm.

Table 7.3. Identification of signal groups and calculation of the integral of these groups for the ^{1}H NMR spectrum (cf. Figure 7.6 and Figure 7.7), after which each HSQC signal is assigned to the corresponding group.

Group No.	Edges of the signal groups in ^{1}H NMR spectrum (ppm)	Contained HSQC signals	Integral of the group
1	1.69–1.64	1.67	3
2	4.31–4.25	4.29	1
3	7.48–7.43	7.45	1
4	7.77–7.61	7.64, 7.67, 7.74	3

Notwithstanding, this method is limited mainly because it supposes that the integral of the group is evenly distributed among the signals belonging to the group. For example, if the integral of the group is 3 and there are two signals contained in this group, then the intensity of 1.5 is assigned to each signal, although the respective intensities in reality are 1 and 2. Since it is difficult to determine which of the two signals has the intensity of 1 or 2, the above described heuristic method is used. The method presented next tries to solve this problem.

7.2.2.3 Using the integral of the corresponding ^{1}H NMR signals: Calculating the integral of each HSQC signal and possibly merging the closer ones

In order to exclude the limitations occurring with the above methods, a new algorithm was developed. It has been shown that using a uniform intensity of 1 for each HSQC signal is not optimal, as the HSQC intensities are distorted and the weighting between signal intensities is artificially altered. On the other hand, to use the integral of the corresponding ^{1}H NMR signal as HSQC intensity is not easy either. It is difficult to find the edges of the ^{1}H NMR signal and to define the correct number of integrals. The method presented in the following excludes all these errors and limitations.

The new algorithm, named Method4, is very similar to the previous one (Method3). The only difference is the manner of calculating the integrals of HSQC signal when more than one HSQC signal belong to a ^{1}H NMR signal group. As one cannot be sure that the integral of this signal group is equally distributed among the HSQC signals, the chemical shift of each proton in the ^{1}H NMR spectrum is estimated. Since, usually, there are more proton than HSQC signals, the integrals of the closest proton signals are merged. The algorithm is given in Figure 7.9. Each HSQC signal is checked to determine to which ^{1}H NMR signal group it belongs. If at least one HSQC signal belongs to a group, then it is retained in a new empty vector that represents the spectrum. The intensities in this group are normalized to the rounded integral value of the group. Then, the chemical shift of each proton in the ^{1}H NMR spectrum is defined as follows: Sum up the intensities until 0.5, 1.5, 2.5, 3.5 … etc. is reached. The position of the first chemical shift will be at an integral value of 0.5, as a proton in the ^{1}H NMR spectrum (previously normalized to the total number of protons) will have an integral of exactly 1. The chemical shift of the next proton will be at an integral value higher by 1 compared with the previous one. This is continued until the maximum number of protons is reached. Thus, the chemical shift of each proton is estimated. Usually, more protons are present than HSQC signals. Hence, the closest ^{1}H NMR protons are merged until their number is equal to that of the HSQC signals.

```
•  read the $^1$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
•  group the signals based on a threshold (range of $x$ Hz)
•  read the HSQC spectrum
•  find the signals using the ReadBrukerHSQCSpectr1 algorithm
•  remove solvent signals
•  get the $^1$H NMR shifts of HSQC signals (project signals) in ppm
•  order signals by ascending ppm position
•  for each signal do the following steps:
   •  get the position of the signal in the $^1$H NMR spectrum
   •  define to which $^1$H NMR signal group it belongs
   •  if at least one signal belongs to a group, then, calculate
      and round the integral of the group and normalize the
      intensities within this group to this value
•  get the estimated chemical shift of each proton in the $^1$H NMR
   spectrum
•  merge the closest protons until the number of protons is
   equal to that of the HSQC signals
•  replace the intensity of the corresponding HSQC signal
   (currently 1) with the number of protons counted after
   merging
```

Figure 7.9. Method4 algorithm: Unlike the Method3 algorithm, it estimates the chemical shift of each proton, upon which the signals of the required protons (the closest ones) are merged. In the matrix (which represents the spectrum), the integral of the corresponding ^{1}H NMR signal is stored.

The algorithm is exemplified using the same sample spectra (HSQC and ^{1}H NMR shown in Figure 7.4 and in Figure 7.6, respectively) as used above. Here, after the signal groups are identified in the ^{1}H NMR spectrum, the chemical shift of each proton is calculated (Table 7.4). The structure contains eight protons, while the HSQC spectrum has only six signals. Thus, the closest three protons (at 1.65, 1.66, and 1.67 ppm) are merged and form the HSQC signal at (1.67, 22.13) ppm having a total integral of 3.

Table 7.4. Calculated intensities of the HSQC signals of Figure 7.4 using the corresponding ^{1}H NMR spectrum of Figure 7.6.

Proton No.	HSQC signal (ppm)		Chemical shift of the proton in the ^{1}H NMR spectrum (ppm)	Integral of the proton
	^{1}H NMR	^{13}C NMR		
1	7.74	129.61	7.74	1
2	7.67	126.63	7.68	1
3	7.64	132.96	7.64	1
4	7.45	128.49	7.45	1
5	4.29	27.70	4.28	1
6	1.67	22.13	1.67	1
7	1.67	22.13	1.66	1
8	1.67	22.13	1.65	1

7.3 Elimination of solvent signals

After defining the signals in the HSQC spectrum using the threshold found with the above described algorithm, the signals of the solvent impurities dimethyl sulfoxide-d_5 (DMSO-d_5) and $CHCl_3$ signals have to be removed. The signal of DMSO-d_5 is expected at (2.50, 39.50) ppm and that of $CHCl_3$ at (7.26, 77.0) ppm. In both cases, a ±0.02 ppm (direction of ^{1}H NMR) and ±1.5 ppm (direction of ^{13}C NMR) tolerance range is applied. In Table 7.1, signal no. 5 is that of $CHCl_3$.

7.4 Prediction of the HSQC spectra

In order to check the correctness of a spectrum and to make structure-spectra compatibility tests, the measured HSQC spectrum is compared with the predicted one of the corresponding structure. The NMRPrediction 3.0 program[133] is capable of predicting not only the ^{1}H NMR shifts but also the ^{13}C NMR signals. The combination of these two estimations results in the corresponding predicted HSQC spectrum.

7.4.1 Using intensity values of 1

The easiest and fastest method to generate the predicted HSQC spectrum is to predict the proton and carbon chemical shifts (for an example, see Table 7.5). Then, using the atom IDs (identification numbers), all proton-carbon couplings over one bond must be found and the positions with the value of 1 in the matrix (representing the predicted HSQC spectrum at the corresponding ppm values) must be marked. For example, the first C–H coupling (atom ID = 1) is at position (7.40, 124.5) ppm. The predicted HSQC spectrum is displayed in Figure 7.10 (triangles).

Table 7.5. Predicted ^{1}H NMR and ^{13}C NMR chemical shifts for the structure shown in Figure 7.2.

Predicted ^{1}H NMR shifts				**Predicted ^{13}C NMR shifts**			
shift	atom_ID	parent_ID	prediction_quality	shift	atom_ID	parent_ID	prediction_quality
7.40	1	1	0	131.5	6	6	0
7.05	3	3	0	126.9	4	4	0
7.01	2	2	0	124.5	1	1	0
7.21	5	5	0	129.3	3	3	0
3.95	7	7	0	128.1	2	2	0
1.69	9	9	0	131.0	5	5	0
				116.6	10	10	0
				117.7	8	8	0
				24.4	7	7	0
				17.7	9	9	0

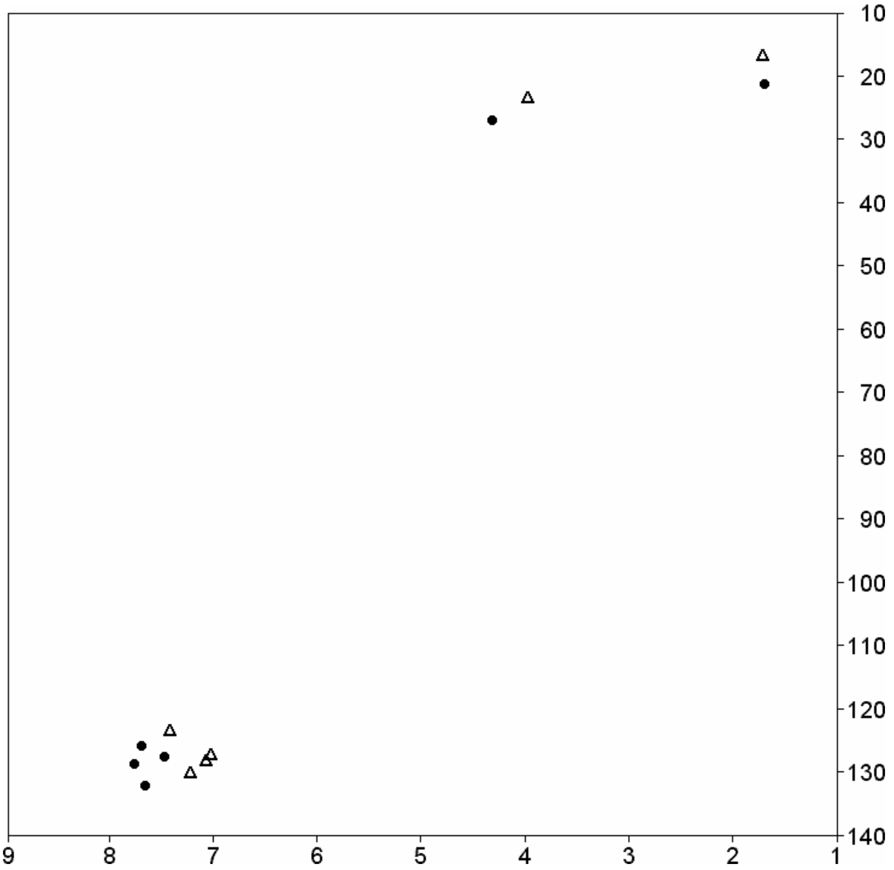


Figure 7.10. The signals of both measured (dots) and predicted (triangles) spectra are represented with the intensity value of 1.

7.4.2 Counting the number of protons for the predicted HSQC signals

The same as with the measured HSQC spectra, also in the case of predicted HSQC signals, we want to assign to each predicted signal, the real number of protons instead of value 1. Thus, the weight of the signals will be different. For this purpose, we need an extra output file of the NMRPrediction 3.0 software,[133] the so-called protocol of the ^{1}H NMR prediction (see Table 7.6). Based on the above mentioned files (chemical shifts and protocol files), the ^{1}H NMR – ^{13}C NMR assignments and the corresponding number of protons are calculated (Table 7.7). For example, in the matrix representing the predicted HSQC spectrum (see Figure 7.10) at position (1.69, 17.7) ppm, the value of 3 denotes the 3 protons of the methyl group.

Table 7.6. Protocol of the ^{1}H NMR prediction for the structure shown in Figure 7.2.

Node	Shift	Base + Inc.	Comment
C	131.5	128.5	1-benzene
		2.5	1 -C(F)(F)-F
		0.5	1 -C-C+N
C	126.9	128.5	1-benzene
		-3.2	1 -C(F)(F)-F
		1.6	1 -C-C+N
CH	124.5	128.5	1-benzene
		-3.2	1 -C(F)(F)-F
		-0.8	1 -C-C+N
CH	129.3	128.5	1-benzene
		0.3	1 -C(F)(F)-F
		0.5	1 -C-C+N
CH	128.1	128.5	1-benzene
		0.3	1 -C(F)(F)-F
		-0.7	1 -C-C+N
CH	131.0	128.5	1-benzene
		3.3	1 -C(F)(F)-F
		-0.8	1 -C-C+N
C	116.6	-2.3	aliphatic
		24.3	1 alpha -1:C*C*C*C*C*C*1
		210.3	3 alpha -F
		-2.5	1 gamma -C
		-0.5	1 delta -C+N
		0.3	1 delta -C
		-113.0	general corrections
C	117.7	117.7	1-nitrile
		0.0	1 -C-C
CH	24.4	-2.3	aliphatic
		24.3	1 alpha -1:C*C*C*C*C*C*1
		4.3	1 alpha -C+N
		9.1	1 alpha -C
		-2.5	1 gamma -C
		0.0	3 delta -F
		-8.5	general corrections
CH3	17.7	-2.3	aliphatic
		9.1	1 alpha -C
		9.3	1 beta -1:C*C*C*C*C*C*1
		2.4	1 beta -C+N
		0.3	1 delta -C
		-1.1	general corrections

Table 7.7. ^{1}H NMR – ^{13}C NMR assignments and corresponding numbers of protons.

^{1}H NMR – ^{13}C NMR couplings	Node
7.40, 124.5	CH
7.05, 129.3	CH
7.01, 128.1	CH
7.21, 131.0	CH
3.95, 24.4	CH
1.69, 17.7	CH3

7.5 Bin method applied to HSQC spectra

The bin method introduced in Chapter 4 and successfully applied to ^{1}H NMR spectra is adapted to two-dimensional spectra and demonstrated with HSQC spectra. Basically, it is a simple generalization of the bin method. In this case, the spectra are divided into rectangles instead of vector sections.

First, the total integral of each individual spectrum is normalized to the total number of protons. Then, the spectra are successively divided into n^2 rectangle bins (n divisions along rows and columns) with $n = \overline{1, N}$; N corresponds to the maximal number of divisions in both directions (^{1}H NMR and ^{13}C NMR) and is obtained through dividing the spectral range in the direction of ^{13}C NMR by the ppm value of the desired minimal bin width (used as an input parameter). Minimal bin widths of 10 ppm, 5 ppm, and 2.5 ppm were used for testing purposes, corresponding to minimal bin widths of 0.6315 ppm, 0.3157, and 0.1578 ppm in the direction of ^{1}H NMR. Note that these calculations are made on the original HSQC spectra, which have a range of 190 ppm and 12 ppm in the direction of ^{13}C NMR and ^{1}H NMR, respectively.

For each division, the similarity index, SI_n, is calculated (cf. Section 4.4):

$$SI_n = \frac{I_{xy}(n)}{I_x + I_y - I_{xy}(n)} \qquad 4.1$$

I_x and I_y being the total integrals of the spectra x and y and:

$$I_{xy}(n) = \sum_{i=1}^{n} \min\left(I_x(i), I_y(i)\right) \qquad 4.2$$

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^*, connecting the remaining global maxima rather than the average of the SI_n values:

$$S = \frac{1}{N} \sum_{n=1}^{N} SI_n^* \qquad 4.3$$

where

$$SI_n^* = \max\left(SI_n, \frac{SI_a(n-b) - SI_b(n-a)}{a-b} \right) \tag{4.4}$$

with

$$SI_a = SI_{n-1}^* \quad \text{and} \quad SI_1^* = 1 \tag{4.5}$$

$$SI_b = \left\{ \max SI_i \,\middle|\, i = \overline{n, N} \right\} \tag{4.6}$$

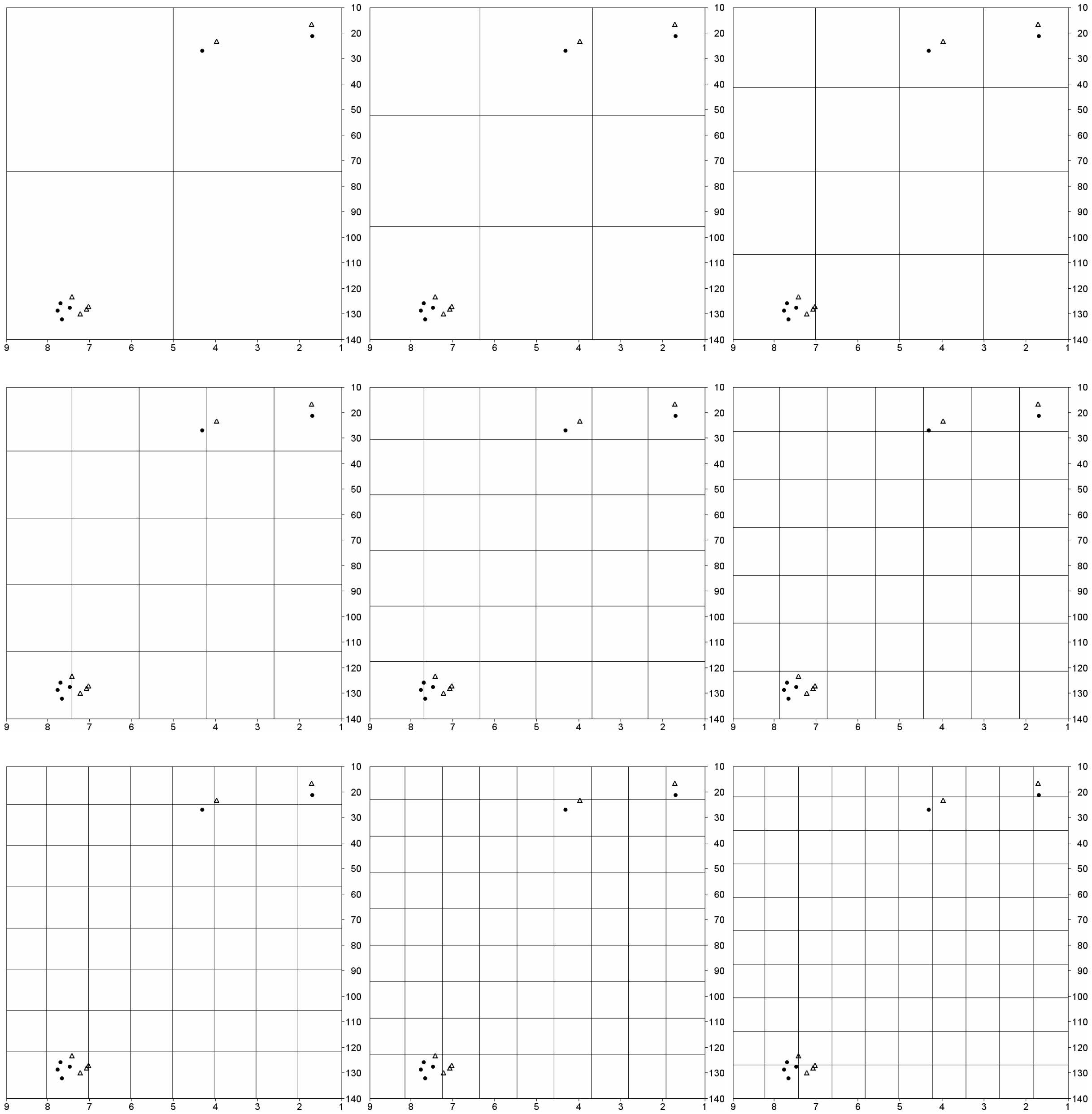


Figure 7.11. Calculating the similarity between the measured and predicted HSQC spectra from Figure 7.10. The spectra are successively divided into n^2 bins ($n = \overline{2,10}$), yielding, respectively, the similarity values of 1.0000, 1.0000, 1.0000, 0.8182, 0.9286, 1.0000, 0.9375, 0.8704, and 0.7924.

In order to demonstrate the method, the similarities between the measured and predicted HSQC spectra from Figure 7.10 are calculated using a different maximal number of bins. The signals in both spectra are represented with intensity values of 1. Both spectra are successively divided into n^2 bins ($n = \overline{2,10}$), yielding different similarity values: 1.0000, 1.0000, 1.0000, 0.8182, 0.9286, 1.0000, 0.9375, 0.8704, and 0.7924 (see Figure 7.11). A division of 10 bins in both directions (i.e., a partition of totally 100 bins) will result in a minimal bin width of 13 ppm in the direction of ^{13}C NMR (range: 130 ppm) and 0.8 ppm in that of ^{1}H NMR (range: 8 ppm).

It is observed that also here, as in the one-dimensional case, the similarity values fluctuate, i.e., a finer division with more bins may provide a higher value of similarity than a coarser one. The similarity with 7 divisions in both directions (i.e., a total of 49 bins) is equal to 1.0000, while with 5 divisions, it is only 0.8182.

Usually, for testing purposes, a larger number of bins is used: Optimal results have been achieved with 38^2 bins or 76^2 bins, 38^2 bins corresponding to 3.42 ppm and 0.21 ppm of minimal bin width in the direction of ^{13}C NMR and ^{1}H NMR, respectively, while 76^2 bins correspond to minimal bin widths of 1.71 ppm and 0.11 ppm in the direction of ^{13}C NMR and ^{1}H NMR, respectively. The similarity of the two spectra in Figure 7.10 is 0.3195 with 38^2 bins and 0.1968 with 76^2 bins.

7.5.1 Overlapping bins

A basic limitation of the bin method is that the signals in two spectra (e.g., measured and calculated ones) may fall in different ranges (bins) even if the difference in their position is very small. 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. Again, the above measured and predicted HSQC spectra (Figure 7.10) are used to visualize this method. In Figure 7.12, the spectra are successively divided into n^2 bins ($n = \overline{2,4}$) represented by the spectra in the three different columns of the figure (from left to right). Three different overlap sizes are applied to the spectra in the rows of the figure (from top to bottom): 10%, 30%, and 70%. The similarity values are listed in Table 7.8. It can be stated that, in general, the larger the overlap used, the lower is the similarity value. It is also obvious that without overlap, the similarities are much higher. The question is: In which proportion will the similarities of foreign HSQC spectra be lower than those of related ones? In order to answer this question, tests were conducted and the results are shown in Section 7.8.

Table 7.8. Similarities obtained by comparing the measured and predicted HSQC spectra of Figure 7.10 using different numbers of bins with different overlap sizes.

Overlap between bins	Number of bins		
	2	3	4
0%	1.0000	1.0000	1.0000
10%	1.0000	1.0000	0.9375
30%	1.0000	0.9524	0.8810
70%	1.0000	0.9000	0.9625

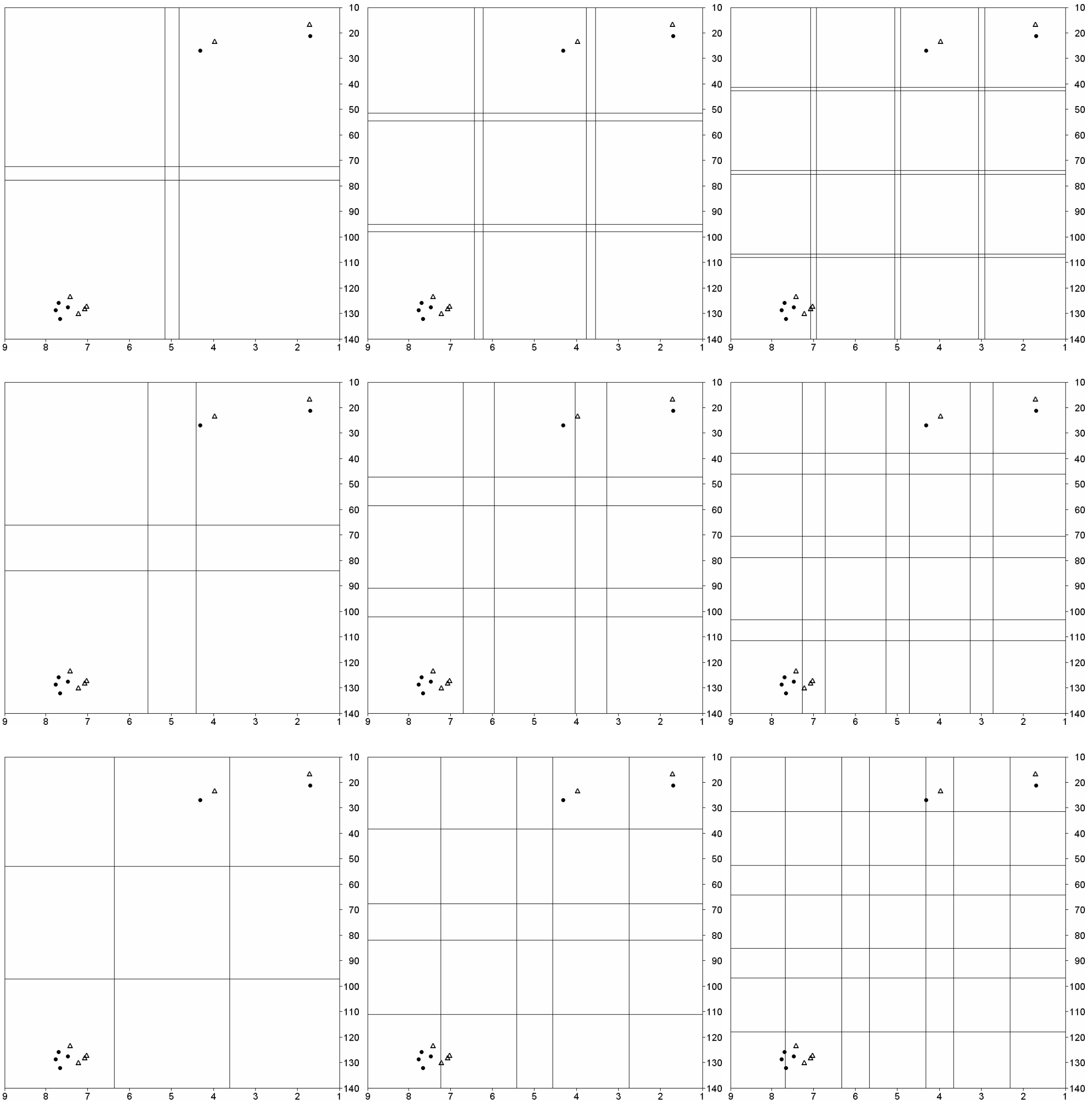


Figure 7.12. Calculating the similarity between the measured and predicted HSQC spectra (Figure 7.10) using overlapping bins. The spectra are successively divided into n^2 bins ($n = \overline{2,4}$; spectra in the figure from left to right) using overlaps of 10%, 30%, and 70% (spectra from top to bottom), yielding the following similarity values: 1.0000, 1.0000, and 0.9375 (10% overlap); 1.0000, 0.9524, and 0.8810 (30% overlap); 1.0000, 0.9000, and 0.9625 (70% overlap).

7.6 Rotating the HSQC spectra

Since the ^{1}H and ^{13}C chemical shifts of $\mathrm{CH_n}$ groups have a certain 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 on the signals' main trend. Moreover, the probability is higher that measured and predicted signals close to each other will not be separated in different bins.

As the rotation is done using the ppm coordinates, first, the spectrum is stretched along the x axis, i.e., along the ^{1}H NMR coordinates, in order to obtain a square. After the spectrum has been transformed to a square, it is rotated clockwise by 45°. In Figure 7.13, this is shown for the measured and predicted HSQC spectra of Figure 7.10.

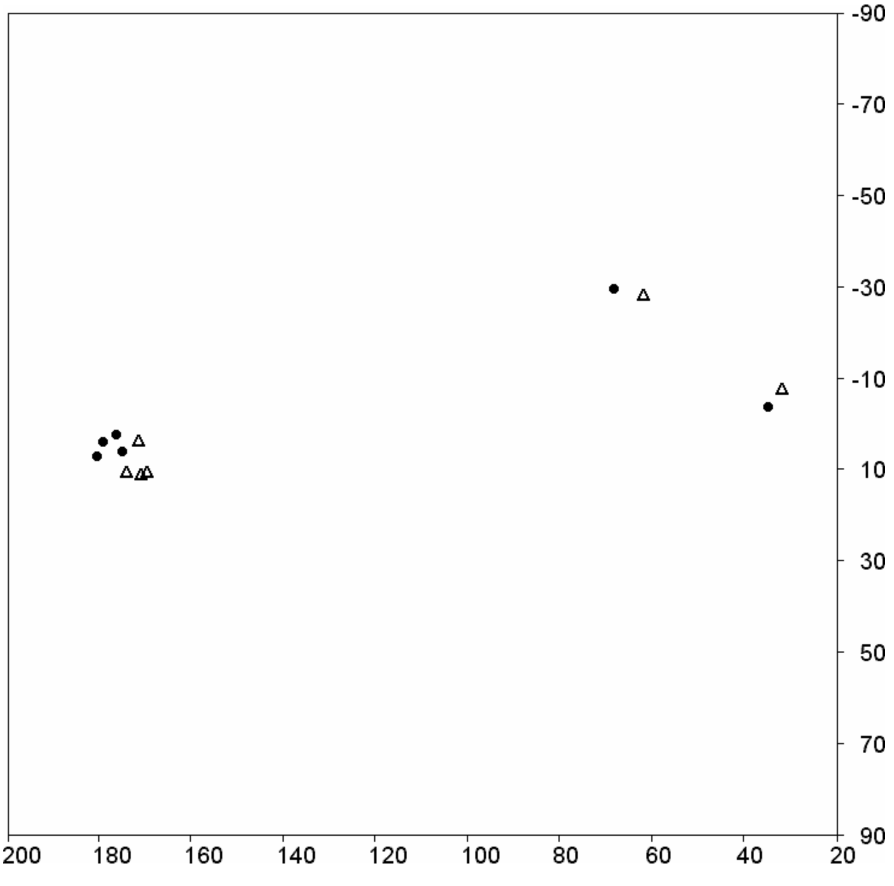


Figure 7.13. Measured (dots) and predicted (triangles) HSQC spectra of Figure 7.10, both rotated clockwise by 45°.

As we see, it has an effect also on the calculated similarity. The similarity between the measured and predicted HSQC spectra is 0.4425 without, and 0.6727 with rotation. To compare the spectra in both cases, the bin method was used with 26^2 bins. As expected, the similarity is higher with rotated spectra. Ideally, it would be higher with related spectra and lower with foreign ones, thus, the overall selectivity of the method would be better. These results will be shown in Section 7.8.

7.7 Test set

The different approaches presented above were investigated with a test set of 289 structures and their corresponding ^{1}H NMR and HSQC spectra. The same test set was used for the comparison of ^{1}H NMR spectra (see Section 5.3).

The HSQC spectra were recorded in the range of -5–185 ppm with 512 digital points in the ^{13}C NMR direction and in the range of -0.5–11.5 ppm with 2048 digital points in the ^{1}H NMR direction. During the tests, the original sizes of the spectra were used, not only the regions where signals are present.

Analogous to the one-dimensional tests, also here, 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). For more information, see Section 5.3.

The overlap between two histograms (in the case of normal and random assignments) was 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 four HSQC signal selection algorithms (Section 7.2) with various parameter values.

Similar to the ^{1}H NMR spectra, also here, tests were made with the 250 proposed (incorrect) structures. Detailed information regarding this is available in Section 5.3.

7.8 Results and discussion

As was presented in the theoretical part of this chapter (Section 7.2), four different algorithms were developed to find the signals in an HSQC spectrum and to assign intensities to these signals. The algorithms can be classified into two main groups based on the type of assigned intensities. The first algorithm (called Method1, see Section 7.2.1) uses uniform intensities of 1 for each signal, whereas with the other three methods, the integrals of the corresponding ^{1}H NMR spectra serve as intensity measure. The first of this type of algorithm (referred to as Method2, see Section 7.2.2.1) uses the integral of the corresponding ^{1}H NMR signal by integrating intensities between the defined edges in order to estimate the intensity of an HSQC signal. The other two methods (Method3 and Method4) estimate the intensities of the

HSQC signals by first dividing the ^{1}H NMR spectrum into groups of signals, i.e., identifying “islands” of signals. In each group, the distance between neighboring signals is closer than a given threshold (in Hz). Both these methods are very similar and differ only in the way how the integrals of the signal groups are distributed among HSQC signals. Method3 (see Section 7.2.2.2) makes an equal distribution of the integrals, while Method4 (see Section 7.2.2.3) is more accurate, calculating the integral of each HSQC signal and possibly merging the closest ones.

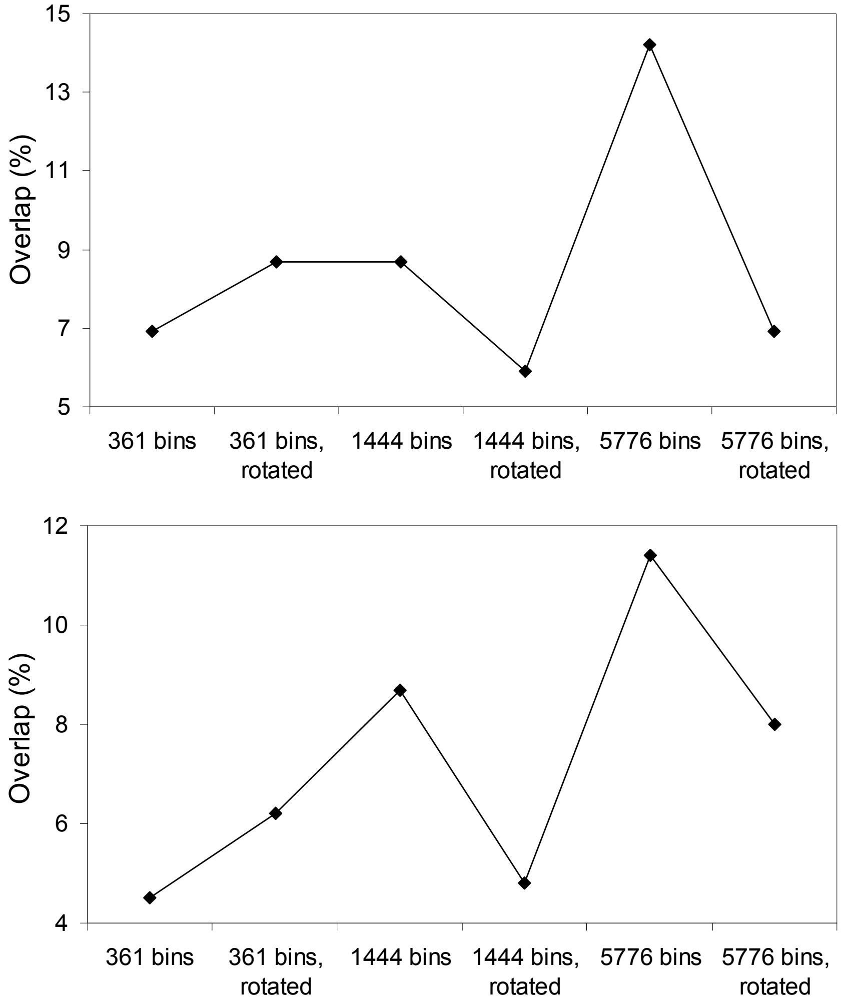


Figure 7.14. Testing the bin method with various parameters (19^2, 38^2, and 76^2 bins, original or rotated spectra): Overlap percentage between the histograms of similarity values, *S*, of measured and calculated HSQC spectra using correct and random structure assignments. For the definition of HSQC signals and intensities, the Method1 (top) and Method2 (bottom) algorithms were used.

In the following, the performances of the four algorithms are compared. The involved performance metric is the overlap percentage between the histograms of similarity values, *S*, of measured and calculated HSQC spectra using correct and random structure assignments. Several tests were made using the bin method with different numbers of maximal bins as well as with normal or rotated HSQC spectra and with correct or incorrect (proposed) structures for predicting HSQC spectra. The first experiments were conducted in order to find the optimal number of maximal bins. Figure 7.14 and Figure 7.15 show the overlap percentage between the histograms of similarity values, S, of measured and calculated HSQC spectra using correct and random structure assignments.

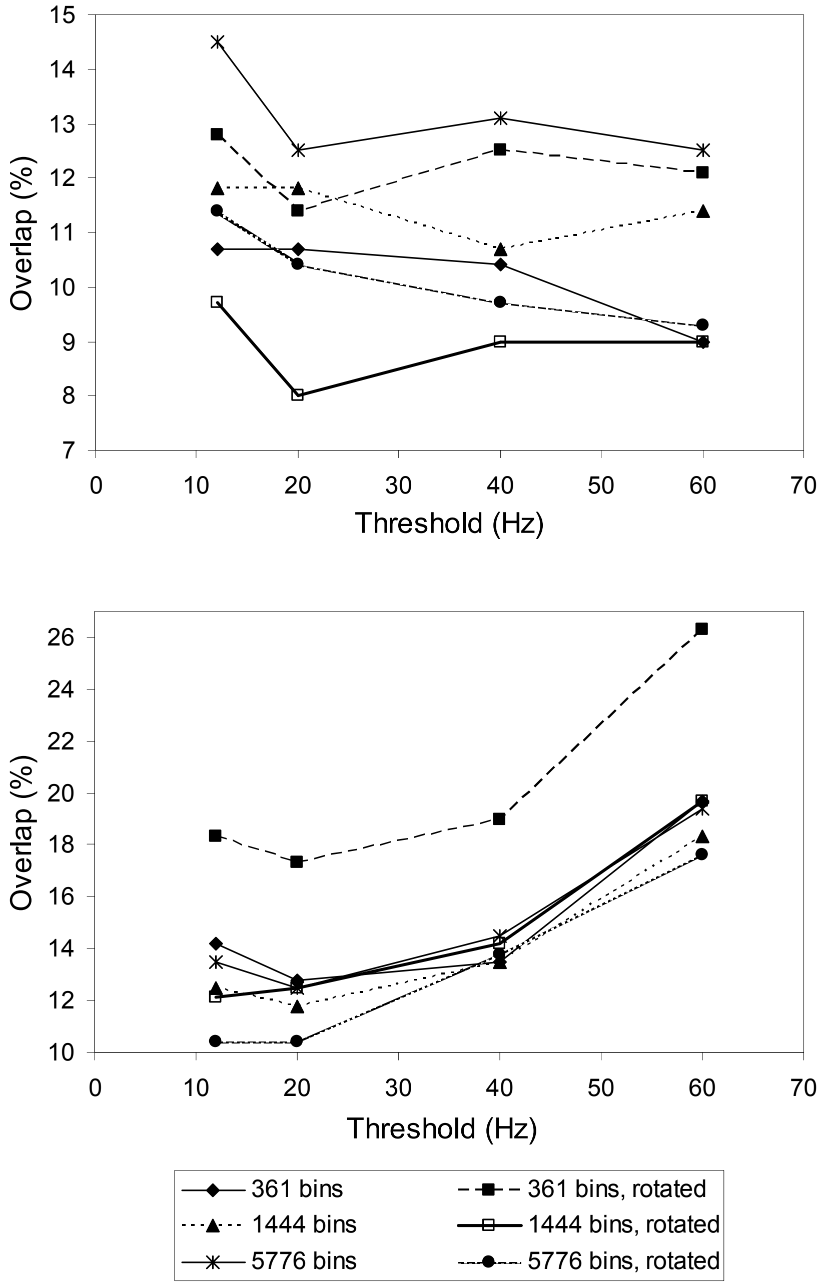


Figure 7.15. Testing the bin method with various parameters (19^2, 38^2, and 76^2 bins, original or rotated spectra): Overlap percentage between the histograms of similarity values, S, of measured and calculated HSQC spectra using correct and random structure assignments. Threshold levels of 12, 20, 40, and 60 Hz were tested. For the definition of HSQC signals and intensities, the Method3 (top) and Method4 (bottom) algorithms were used.

For comparing spectra pairs, the bin method was used with different maximal numbers of bins, i.e., $19^2 = 361$ bins (corresponding to a minimal bin width of 10 ppm in the ^{13}C NMR direction and 0.63 ppm in the ^{1}H NMR direction), $38^2 = 1444$ bins, and $76^2 = 5776$ bins. The spectra were either used in their original form or were rotated clockwise by 45°. In the case of Method3 and Method4 algorithms (Figure 7.15), threshold levels (i.e., distances between neighboring ^{1}H NMR signals within a group of signals) of 12, 20, 40, and 60 Hz were tested. It can be seen that, usually, the lowest overlap is achieved with rotated spectra using a maximal number of 38^2 bins. Only Method4 has a better performance (overlap of 10.4%) with more bins (76^2). It is true that Method2 gives a slightly lower overlap (4.5%) with 19^2 bins, but this is statistically not significant as an overlap of 4.8% is achieved with 38^2 bins. Hence, in the forthcoming tests, these parameter settings (38^2 non-overlapping bins with rotated spectra) will be used with Method2 as well. Method1 and Method3 also perform best with 38^2 bins, showing an overlap of 5.9% and 8.0%, respectively. Finally, it can be stated that the Method3 and Method4 algorithms yield the best results with a threshold level of 20 Hz.

Hence, in the following, the comparisons were made using the same parameter settings for the bin method as defined above for each algorithm. The resulting histograms of similarity values of measured and calculated HSQC spectra are presented in Figure 7.16.

As discussed in Section 7.5.1, overlapping bins may improve the performance of the similarity measure. Thus, tests were conducted using normal and rotated spectra, the bin method with 38^2 and 76^2 bins, and various sizes of overlapping bins, i.e., 10%, 30%, 50%, 70%, and 90% (Figure 7.17). It is observed that in the case of the Method1 algorithm, the lowest overlap of 6.2% for spectra pairs was achieved with rotated spectra and with an overlap of 90% between bins. However, this is higher than that achieved with non-overlapping bins (5.9%) using the same number of bins (1444, i.e, 38^2 bins) with rotated spectra. The same situation is valid for the Method3 algorithm: The lowest overlap is 9.0% (overlap of 90% between bins), while with non-overlapping bins it is only 8.0%. Also, for the Method4 algorithm, the overlap is higher (10.7%) with overlapping bins (overlap of 30% between bins) than with non-overlapping ones (10.4%). In conclusion, it can be stated that overlapping bins do not significantly improve the selectivity of the bin method, therefore, they will not be applied. Moreover, the use of overlapping bins is time-consuming.

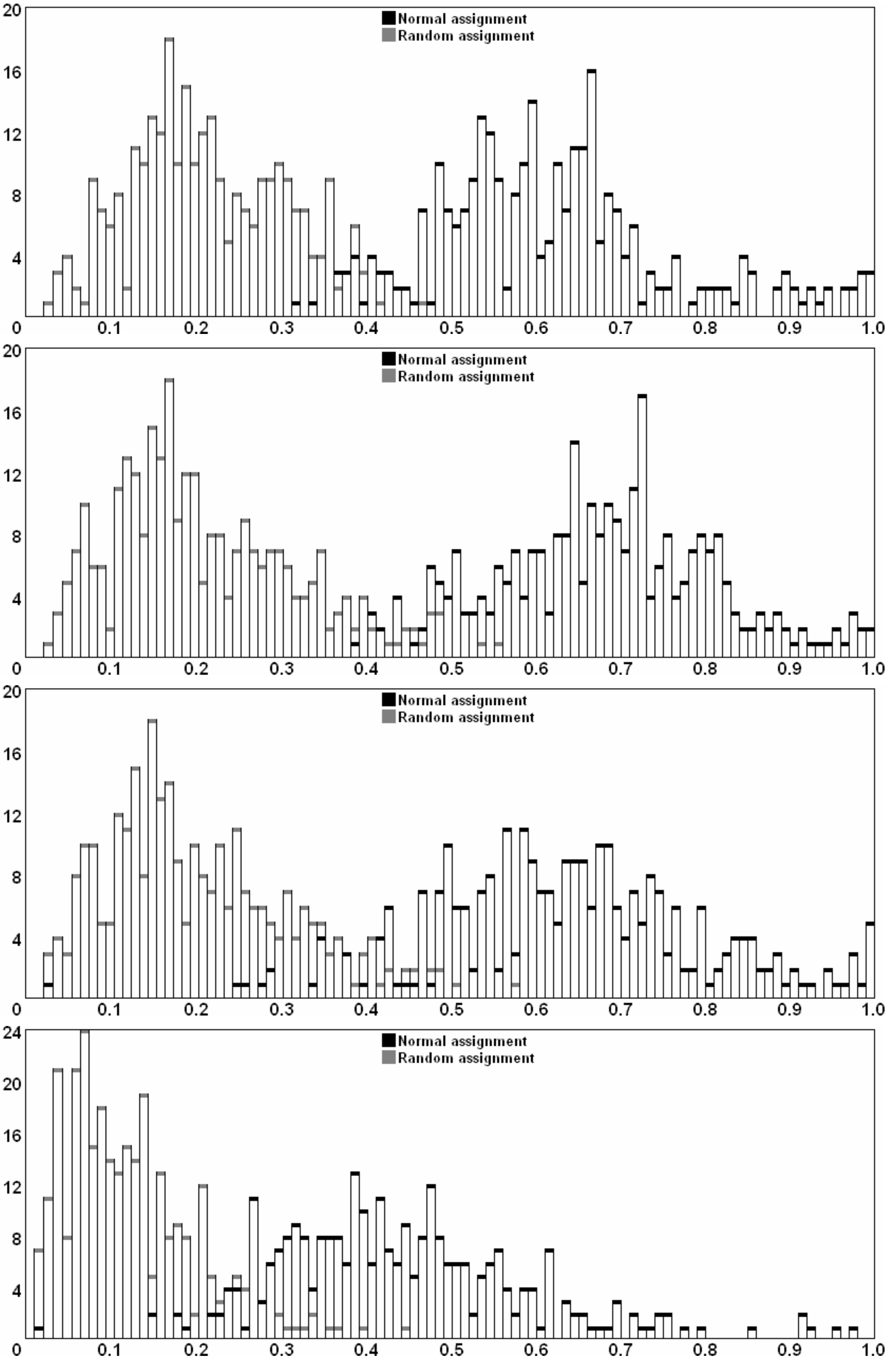


Figure 7.16. Histograms of similarity values of measured and calculated HSQC spectra using normal (black histogram) and random (gray histogram) structure assignments. Overlaps with the four algorithms (Method1–4, from top to bottom) are: 5.9%, 4.8%, 8.0%, and 10.4%. To compare the spectra, the bin method was used with a maximal number of 38^2 or 76^2 bins (Method4). Prior to comparison, the spectra were rotated clockwise by 45°.

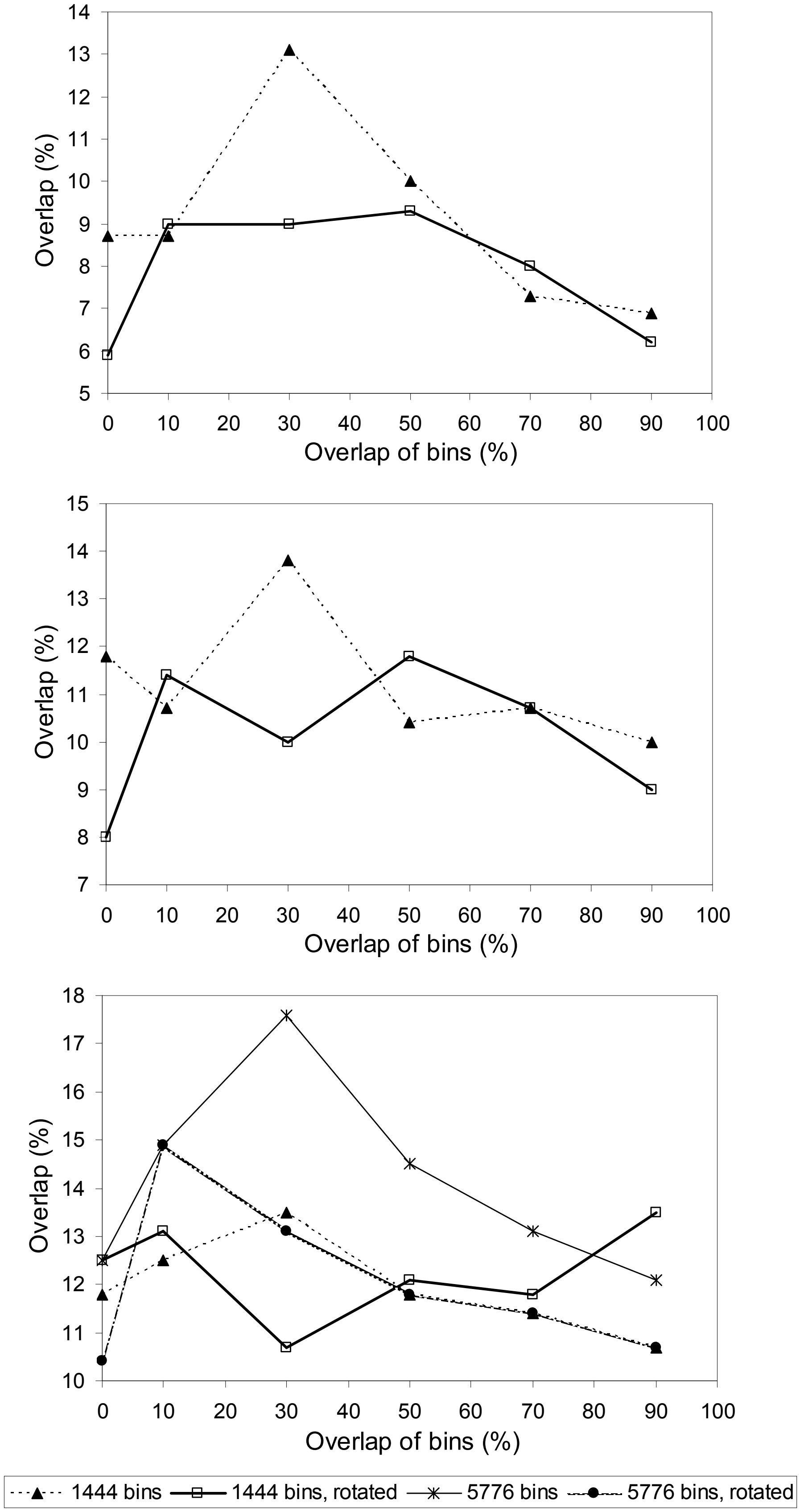


Figure 7.17. Investigating the performance of overlapping bins for Method1, Method3, and Method4 (from top to bottom): Tests were conducted using normal and rotated spectra, the bin method with 38^2 or 76^2 bins, and various sizes of overlapping bins: 10%, 30%, 50%, 70%, and 90%. In the case of Method3 and Method4 algorithms, the tests were made with a threshold level of 20 Hz.

In a next test, experiments involving the proposed (incorrect) structures were performed. Lower similarities are expected when the incorrect structures are used. In a first test, the same procedure was done as before with the correct structures. Each of the 250 measured spectra was compared with two predicted ones (all of them rotated clockwise by 45°): one based on the corresponding incorrect structure (normal assignment) and the other based on a randomly selected incorrect structure from the dataset (random assignment). Figure 7.18 shows the generated histograms for each of the four algorithms (Method1 to Method4 from top to bottom). As expected, here, the overlap is much higher than in the case of the correct structures: 25.2% in comparison with 5.9% (Method1); 22.9% in comparison with 4.8% (Method2); 29.0% in comparison with 8.0% (Method3); and 28.6% in comparison with 10.4% (Method4). These tests were conducted under the same conditions and with the same parameter settings as in the case of the correct structures, thus, it is justified to compare the results. Both experiments show how selective the bin method is, as it is obviously capable of discriminating between correct and incorrect structures. Note that the incorrect structures are the expected, proposed ones. They are proposed by a specialist, therefore, the aim is to have the best possible solutions and, in most cases, they are very similar to the correct ones (in some situations, the only differences are due to the presence of diastereomers and charges, which are not handled by the prediction program).

Figure 7.19 reveals important information about the selectivity of the similarity measure. Each of the 250 measured HSQC spectra was compared with the predicted spectrum (both rotated clockwise by 45°) of the corresponding elucidated structure (black histogram). The average similarity is 0.62, 0.69, 0.64, and 0.44 in case of Method1, Method2, Method3, and Method4 algorithms, respectively. In another test, each of the 250 measured spectrum was compared with the predicted spectrum of the corresponding proposed structure (gray histogram; in some cases, more than one, which is why there are 261 comparisons in total). In this case, the average similarity is 0.48, 0.52, 0.48, and 0.32 for Method1, Method2, Method3, and Method4 algorithms, respectively.

Figure 7.20 and Figure 7.21 present another type of visualization of the above results. Here, 261 measured HSQC spectra are compared with the spectra predicted for the correct and incorrect structures. As expected, in most cases, higher similarities are achieved with correct structures.

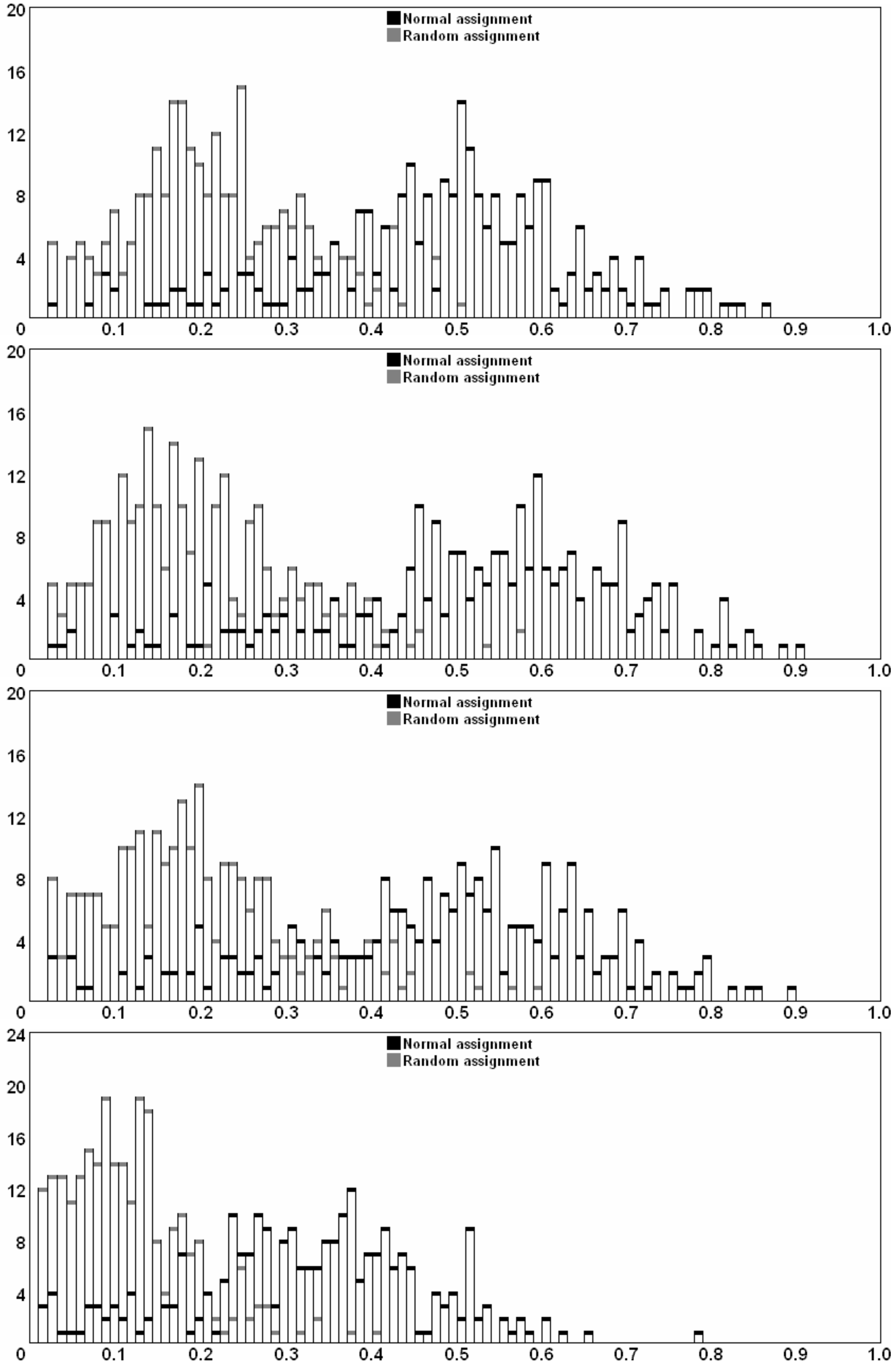


Figure 7.18. Histograms of similarity values of measured and predicted (on the basis of incorrect structures) HSQC spectra using normal (black histogram) and random (gray histogram) structure assignments. Overlaps with the four algorithms (Method1– 4, from top to bottom) are: 25.2%, 22.9%, 29.0%, and 28.6%. For comparing the rotated spectra, the bin method was used with a maximal number of 38^2 or 76^2 bins (Method4).

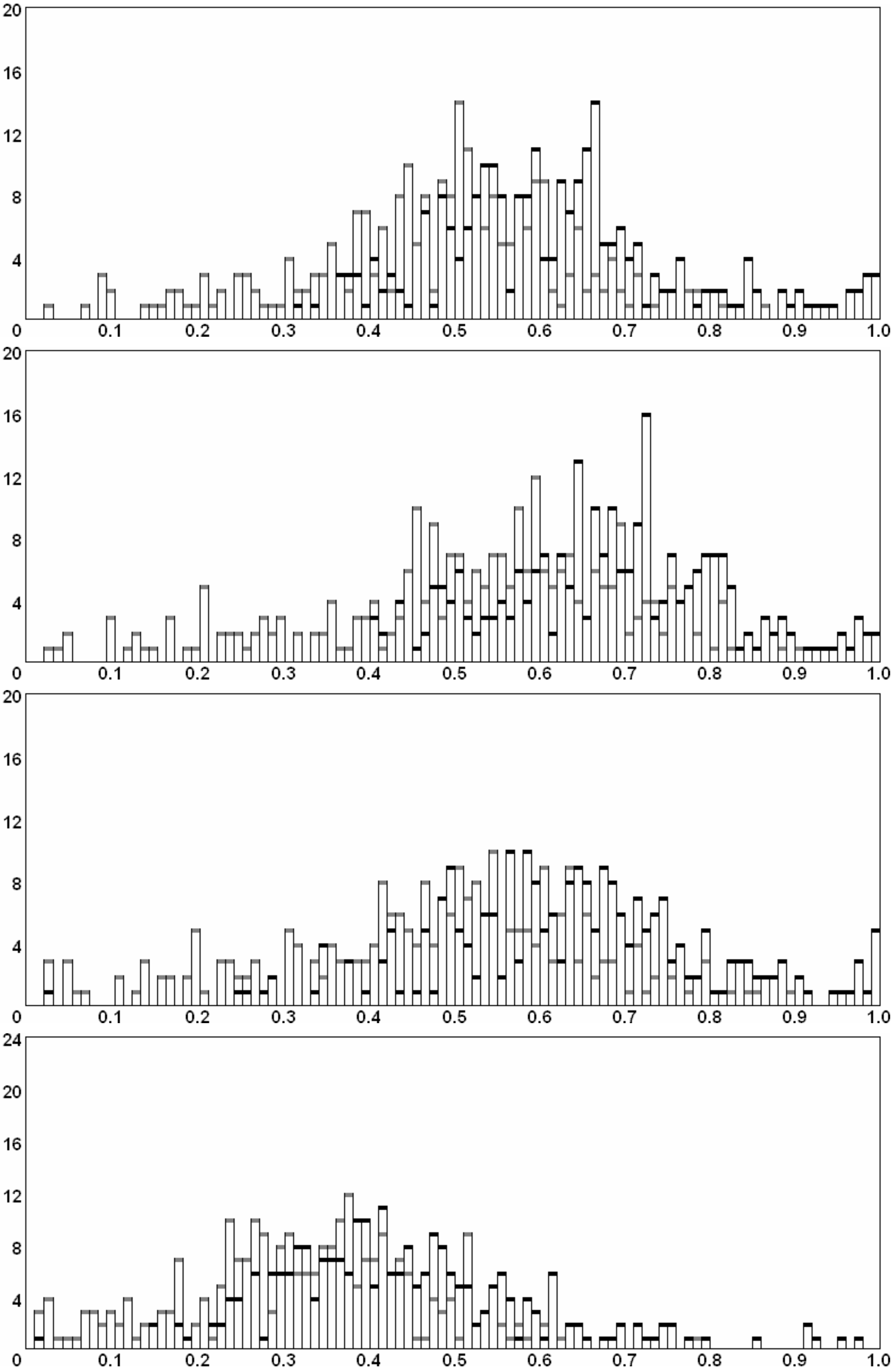
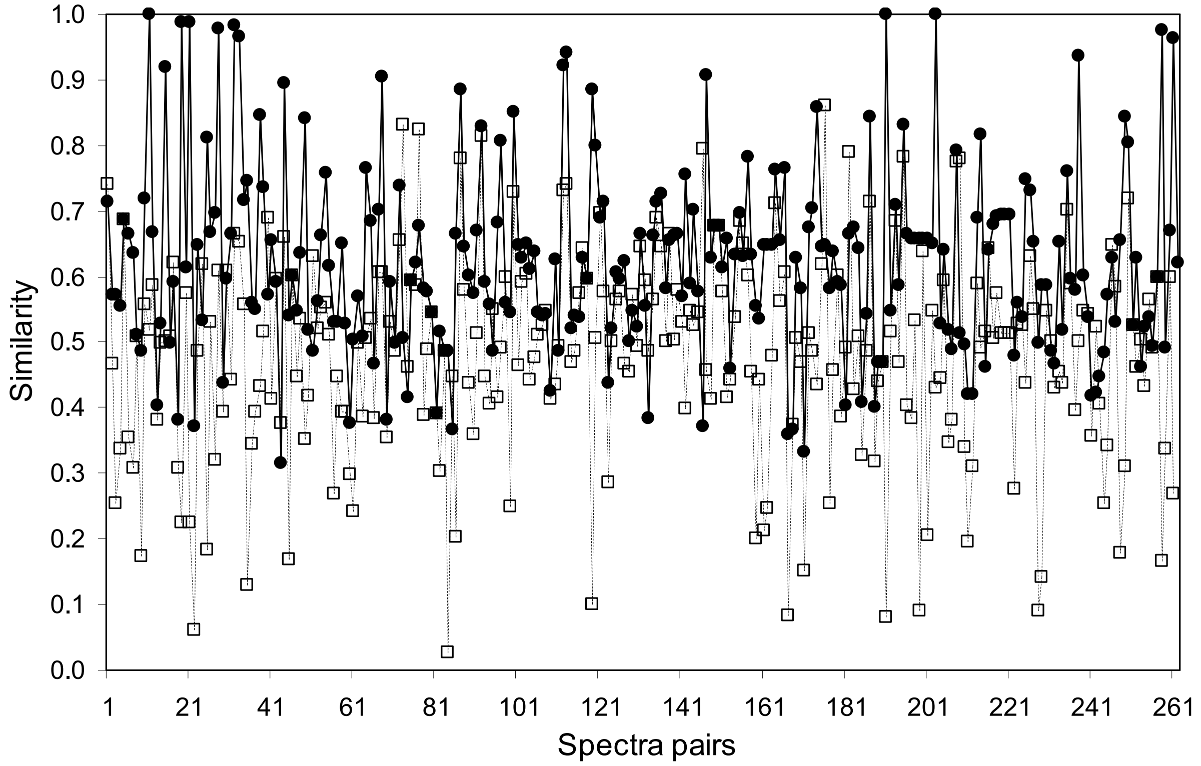


Figure 7.19. Comparing the similarities when correct (black histogram, 250 spectra pairs) and incorrect structures (gray histogram, 261 spectra pairs) are used to predict their HSQC spectra. The predicted spectra are compared with the measured HSQC spectra (both rotated clockwise by 45°) using the bin method with a maximal number of 38^2 or 76^2 bins (Method4). For the definition of HSQC signals and intensities, the four algorithms (Method1– 4) were used (from top to bottom).



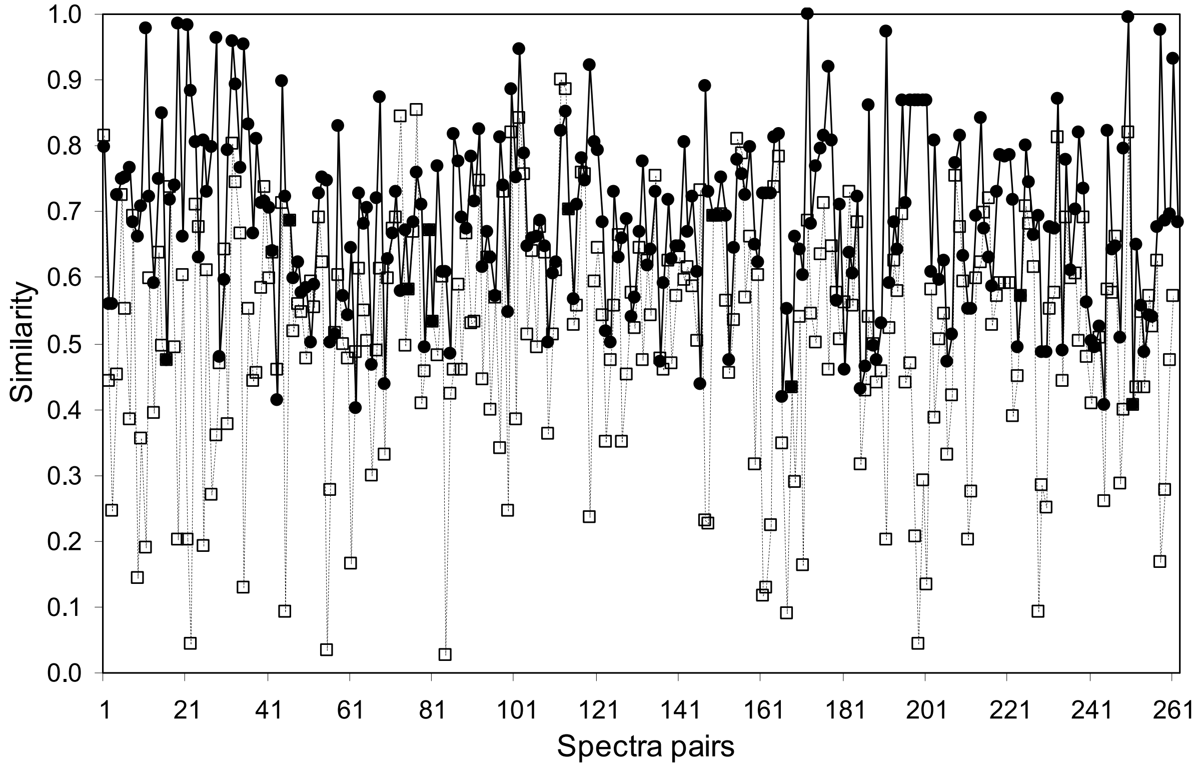
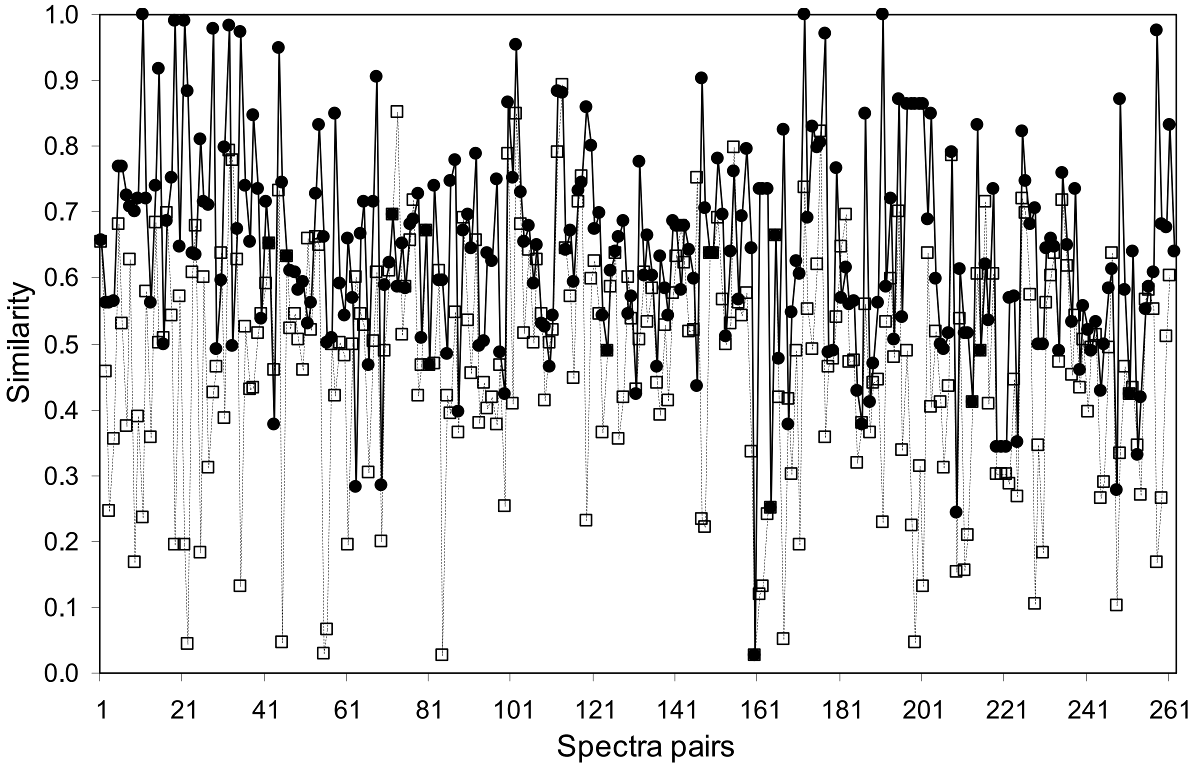


Figure 7.20. Comparing the similarities when correct (solid line) and incorrect structures (dashed line) are used to predict their HSQC spectra. The predicted spectra are compared with the measured ones (both rotated clockwise by 45°) using the bin method with a maximal number of 38^2 bins. The average similarities of 261 comparisons with the algorithms are: 0.62 for correct structures and 0.48 for incorrect ones (Method1); analogously, 0.69 and 0.52 (Method2).



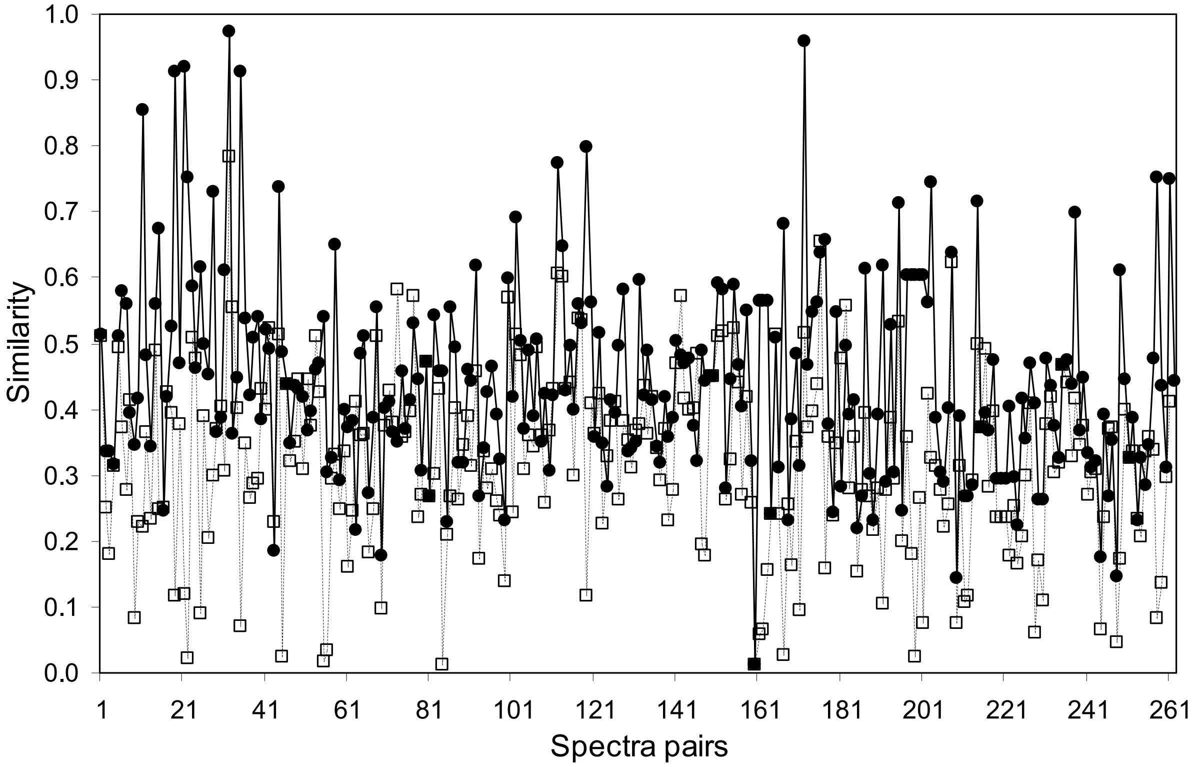


Figure 7.21. Comparing the similarities when correct (solid line) and incorrect structures (dashed line) are used to predict their HSQC spectra. The predicted spectra are compared with the measured ones (both rotated clockwise by 45°) using the bin method with a maximal number of 38^2 or 76^2 bins (with Method4 algorithm). The average similarities of 261 comparisons with the algorithms are: 0.64 for correct structures and 0.48 for incorrect ones (Method3); analogously, 0.44 and 0.32 (Method4).

7.9 Conclusions

As a conclusion, it can be stated that the four methods are able to discriminate between correct and incorrect structures by comparing their predicted spectra with the corresponding measured ones. When correct structures are used, the overlap between spectra pairs using normal and random assignments is as low as 7.3% (average of the four methods). On the other hand, when incorrect structures are used, in average, the overlap between the two cases was found to be as high as 26.4%. For the first three signal definition algorithms, the best results were achieved with a maximal number of 38^2 bins, while for Method4, the maximal number of bins was 76^2. Prior to the comparisons, both measured and predicted HSQC spectra were rotated clockwise by 45°. Concerning the computation times, the Method1 algorithm is the fastest as no ^{1}H NMR spectra are required.

Paradoxically, the Method4 algorithm yielded the worst results, although it was expected to achieve the best performance. In addition, it requires a maximal number of 76^2 bins, which increases the computing time considerably.

Finally, it can be added that the best performance (overlap of 4.8%) is achieved with the Method2 algorithm using non-uniform intensities for HSQC signals. Nevertheless, the overlap of 5.9% generated with the Method1 algorithm involving uniform intensities is low too. Since this method is faster, therefore, it is also the optimal one.

It is interesting to compare these results with those achieved using only the ^{1}H NMR spectra. Note that in Chapter 5, the experiments were conducted with ^{1}H NMR spectra of the same test set, thus, comparison of the results is justified. There, the best performance (lowest overlap of 5.9%) was achieved when the X–H signals (i.e., hydrogens not bonded to carbons) were removed. However, with the original ^{1}H NMR spectra, the overlap was as high as 11.4%. Hence, the use of HSQC spectra is appropriate.

8 Similarity of IR spectra

In contrast to ^{1}H NMR spectra, infrared (IR) spectra have been compared in various studies (see Sections 3.1.4 and 3.3). In most cases, the correlation coefficient between two spectra has been used as a similarity measure.[27, 31, 33, 38] As shown in Chapter 4 (especially in Figure 4.6 and Figure 4.8), the correlation coefficient is unable to discriminate between large and small differences in signal positions as soon as that difference is bigger than about twice the line width. In contrast to ^{1}H NMR, the line width of IR absorption bands in condensed phases is much larger (typically 10–20 cm^{-1} for a total spectral range of ca. 3000 cm^{-1}, while it is typically 0.3 Hz for a total range of 10000 Hz). This difference explains why the correlation coefficient may be successfully used to access IR but not ^{1}H NMR spectral similarities.

In this chapter, spectral similarities obtained by the two methods are compared on a small set of related structures and spectra. It is shown that the performance obtained with the bin method is comparable to (or in some cases better than) that achieved with the correlation coefficient.

8.1 Data set

For testing purposes, the Chemical Concept[127] database was used. The IR spectra have a JCAMP-DX format,[128-130] while the corresponding molecular structures are stored in molfiles.[131, 132] The original spectra were recorded in the range of 400–4000 cm^{-1}, but for the following experiments, only the 650–1800 cm^{-1} region was used (597 points) since the signals in the rest of the spectra contain less relevant information or strongly depend on experimental conditions.

From the available approximately 10500 structures and IR spectra, 2000 were selected at random. Each of the 2000 IR spectra was compared with every other one using the bin method, after which they were classified in 30 clusters with the Ward's method (described in Section 3.2.2.2). A cluster of 116 structures was selected (see Chapter 10, Appendix, Table 10.1) for further investigations.

8.2 Clustering

First, the similarities were calculated for each spectra pair. This yielded a total of 6670 different comparisons (i.e., $n(n-1)/2$, with $n=116$). To compare the spectra, the bin method was used with 500 bins and the correlation coefficient. Then, the pair-wise distances were calculated between spectra by simply subtracting the obtained similarity value from 1. The distribution of pair-wise distances is shown in Figure 8.1.

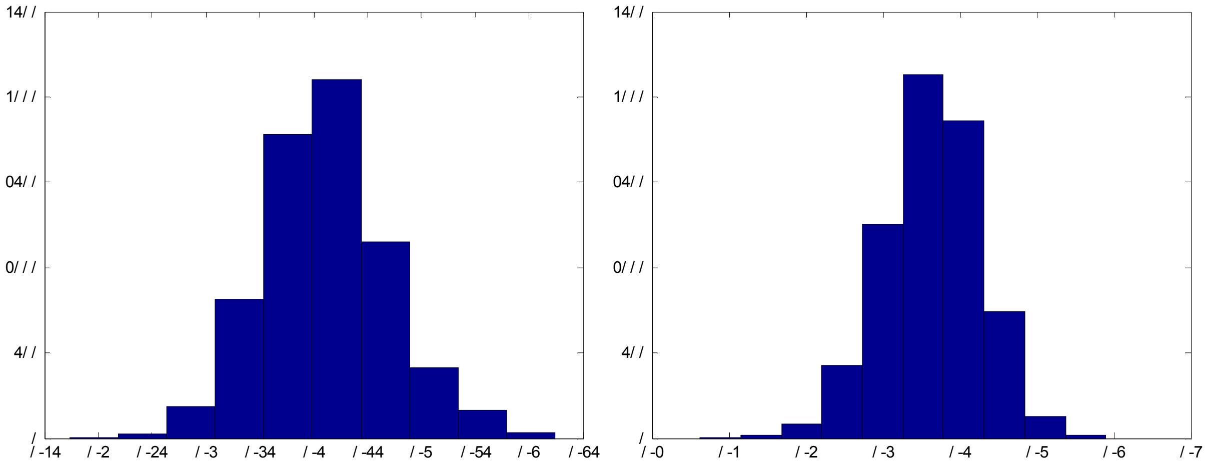


Figure 8.1. Distribution of the pair-wise distances of the 6670 comparisons for the 116 IR spectra using the bin method (left) and the correlation coefficient (right).

In the next step, the 116 spectra were clustered based on the distances using the Ward's method (see Section 3.2.2.2) with both similarity measures. In order to visualize the clusters, dendrograms (see Figure 8.2 and Figure 8.3) with ten leaf nodes were generated using the MATLAB 7.3 software package.[144] The contents of the leaf nodes are presented in Table 8.1 and Table 8.2.

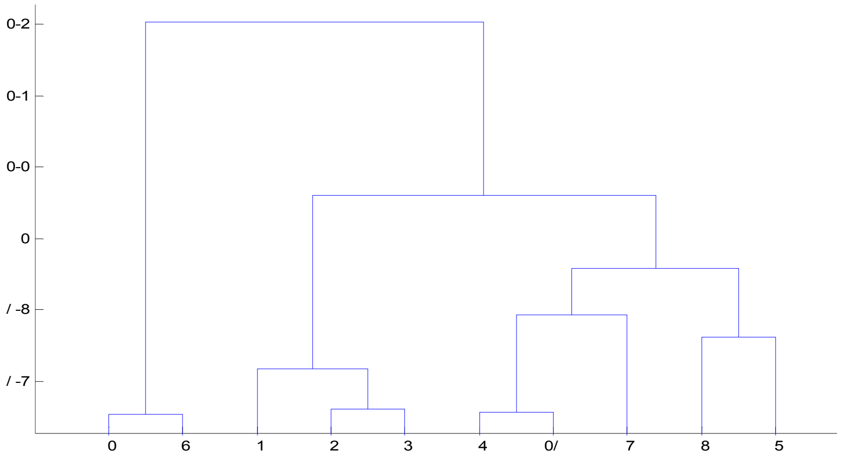


Figure 8.2. Dendrogram of the pair-wise distances using Ward's linkage with ten leaf nodes. The spectra were compared with the bin method.

Table 8.1. Content of the leaf nodes of the dendrogram shown in Figure 8.2. For comparison of the spectra, the bin method was used.

Leaf node	Content of leaf node
1	1 6 8 18 25 33 34 35 40 42 45 53 62 69 74 77 80 85 98 100 102 105 106
2	2 24 72 75 91 94 96 110 112
3	3 14 15 27 108
4	4 99
5	5 16 17 22 28 29 39 43 50 70 81 82 83 86 90 92 104 113 115
6	11 49 55 57 58 59 61 64 65 66 68
7	7 20 21 23 26 31 32 36 41 51 54 56 60 76 109
8	12 30 37 71 93 95 101 107 111
9	9 10 19 44 46 47 48 52 78 79 84 88 89 97 103 114 116
10	13 38 63 67 73 87

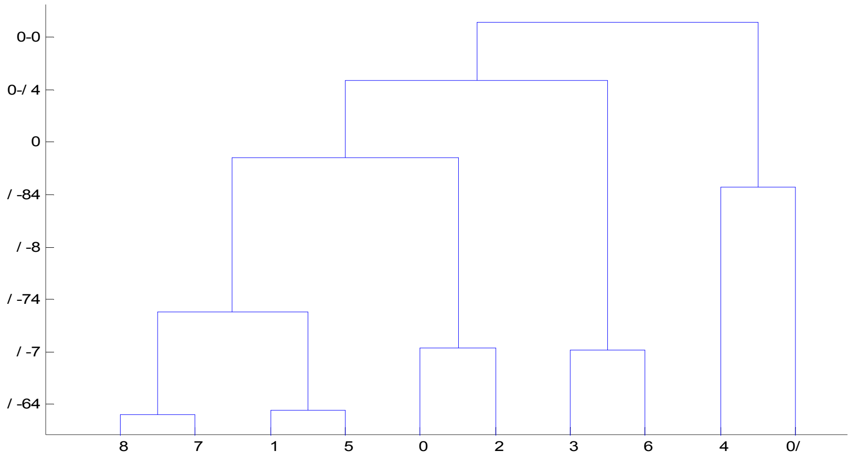


Figure 8.3. Dendrogram of the pair-wise distances using Ward's linkage with ten leaf nodes. The spectra were compared using the correlation coefficient.

Table 8.2. Content of the leaf nodes of the dendrogram shown in Figure 8.3. To compare the spectra, the correlation coefficient was used.

Leaf node	Content of leaf node
1	1 18 35 37 40 50 52 53 55 69 70 80 89 98 105 115 116
2	2 3 43 47 84 88 90 92 97 104 106 113 114
3	13 38 42 63 67 73 87
4	4 15 27 99 108
5	5 8 16 17 20 21 22 25 28 30 31 56 82 83 107
6	6 24 33 39 74 100
7	7 14 23 26 32 34 45 85 86 93 101 102 111
8	19 29 46 48 51 54 60 62 71 75 81 91 94 95 110
9	9 12 36 41 44 76 96 103 109 112
10	10 11 49 57 58 59 61 64 65 66 68 72 77 78 79

8.3 Results and discussion

In order to make a further comparison of the two similarity measures, the spectra pairs are ordered by descending similarities. In the following, several spectra pairs are investigated, presenting the rankings (order numbers after sorting the spectra) obtained by both similarity measures. Figure 8.4 presents the comparison of spectra no. 57 and no. 65. With both methods, the spectra were found to be very similar. The rankings are 2 and 4 using the bin method and the correlation coefficient, respectively. It is seen that the two structures are highly similar. In this case, both methods are capable of detecting the similarities between the two spectra.

As another example, Figure 8.5 shows two similar spectra, but the corresponding structures are less similar. Here, the bin method is able to detect the similarity between the spectra better than the correlation coefficient, since the rank is 10 in comparison with 48 when the correlation coefficient is used.

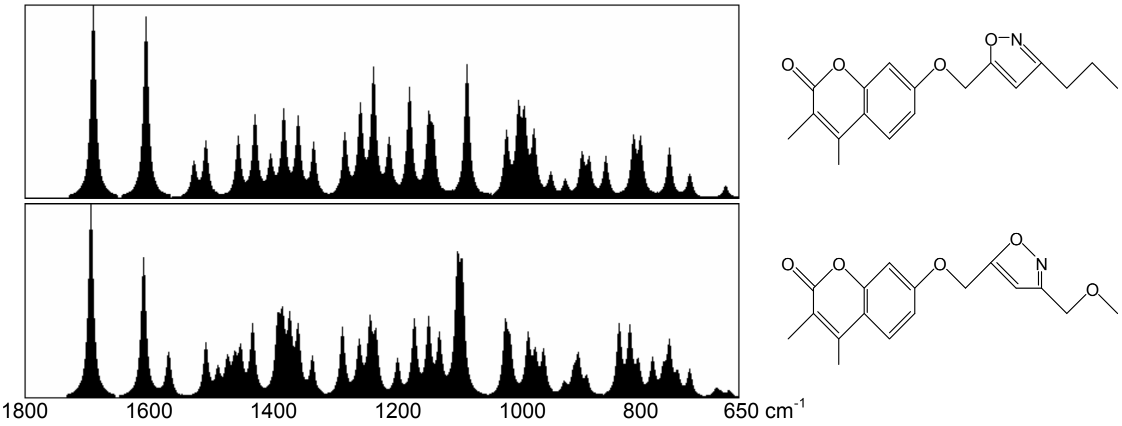


Figure 8.4. Comparing spectrum no. 57 (top) with no. 65 (bottom). The spectra pair has a ranking of 2 with the bin method and 4 with the correlation coefficient. Indeed, both spectra and structures are very similar.

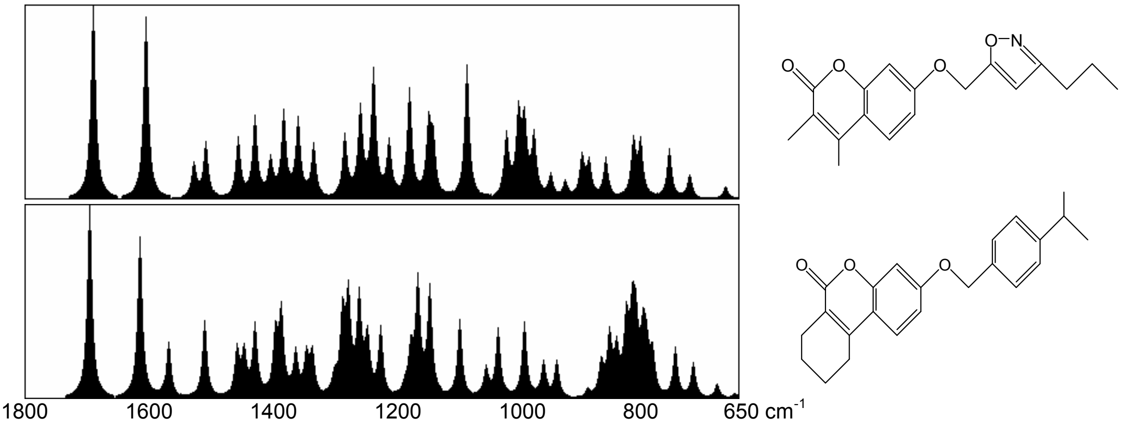


Figure 8.5. Comparison of spectra no. 57 (top) and no. 61 (bottom). The spectra pair has a ranking of 10 and 48 using the bin method and the correlation coefficient, respectively.

Finally, two spectra pairs are shown, where the similarities are high (low ranking number) with the bin method and low (high ranking number) with the correlation coefficient, and vice versa. In the case of the spectra pair in Figure 8.6, the ranking is 4 with the bin method and 112 with the correlation coefficient. Nevertheless, the structures are fairly similar.

It is clear that in the example of Figure 8.7, both the spectra and the structures are rather different. Nevertheless, the correlation coefficient assigns a high similarity (rank 17), while with the bin method, the spectra pair achieves a rank number of 345, which means a significantly lower similarity. In general, both methods lead to similar results. It can be stated that the bin method can better detect the similarities (or dissimilarities) in IR spectra than the correlation coefficient.

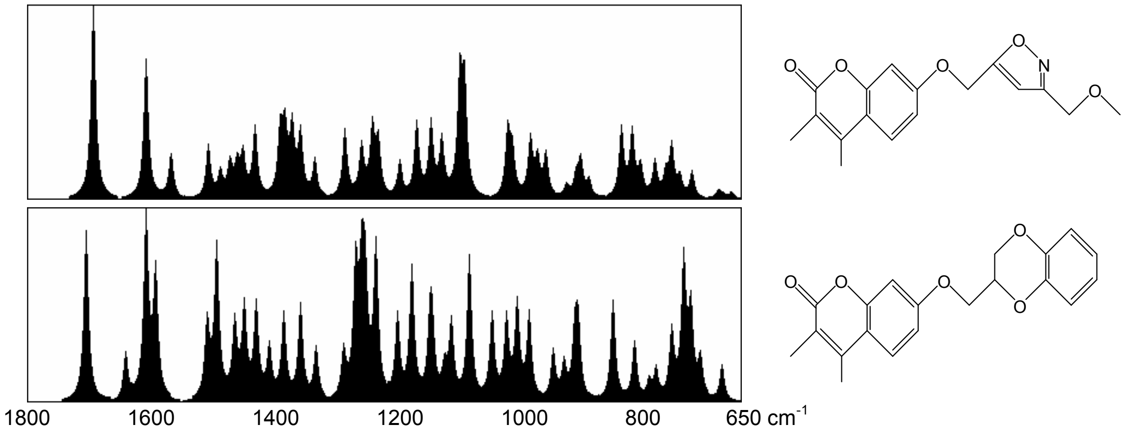


Figure 8.6. Comparing spectrum no. 65 (top) with no. 68 (bottom). The spectra pair has a ranking of 4 (bin method) and 112 (correlation coefficient).

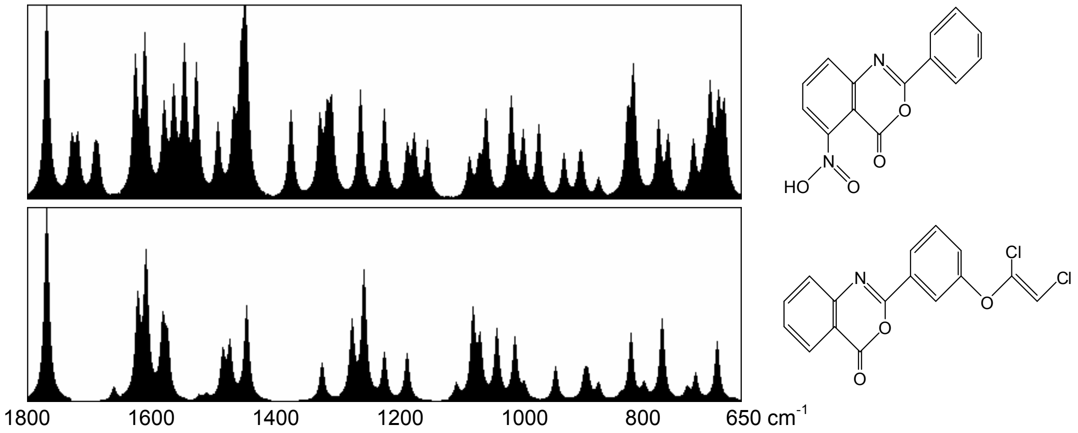


Figure 8.7. Comparison of spectra no. 3 (top) and no. 27 (bottom). The spectra pair has a ranking of 345 and 17 using the bin method and the correlation coefficient, respectively.

Based on preliminary results obtained with IR spectra of polymorphous crystals, it can be stated that the performance of the bin method is competitive with that of the commercial software, TQ Analyst 5.0.[145] Note that in contrast to the previous test set with 116 IR spectra, we here had the opportunity to compare the results achieved with this commercial software.

8.4 Conclusions

It has been shown that the bin method, due to its generality, can be successfully applied also to the comparison of IR spectra. Since IR spectra have broad lines, the correlation coefficient can be efficiently used for spectra comparison so that the application of the bin method is not as crucial as in the case of ^{1}H NMR or HSQC spectra, and the improvements in the selectivity are also minor. However, in some situations, the bin method can better detect the similarities (or dissimilarities) in IR spectra than the correlation coefficient.

9 Conclusions

The similarity of related ^{1}H NMR spectra was successfully detected by a novel method based on dividing the spectra into $n = 1$ 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.

In Chapter 4, it was shown that the bin method is an efficient, general similarity measure that can also successfully compare ^{1}H NMR spectra. This is confirmed by the results in Chapter 5, where it has been applied to a further real-life data set containing 289 ^{1}H NMR spectra. By excluding the X–H signals (i.e., hydrogens not bonded to carbon) from both measured and predicted spectra, the selectivity of the criterion is increased. The overlap between normal and random assignments is found to be almost halved, namely 5.9% instead of 11.4% in case the X–H signals in the spectra are retained.

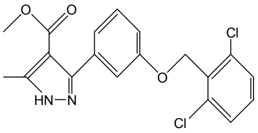
In Chapter 6, the test set was a combinatorial library, with very similar chemical structures (generated from the same educt structures) and spectra. Also under this condition, the bin method is able to cope with the problem. Since the library is complex, several reference spectra could be generated, with which the measured spectra are compared. Gross errors in the test set are easily detected with all the described procedures. However, none of them alone is capable of simultaneously achieving high values of true positives and true negatives. The combination of the various methods provides the most promising approach.

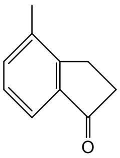
Due to its generality, the bin method was easily adapted and applied to the comparison of two-dimensional HSQC spectra. The same data set was used as in Chapter 5, since besides the ^{1}H NMR also the HSQC spectra of each of the 289 compounds were available. The best performance (overlap of 4.8%) was achieved using non-uniform intensities for HSQC signals. This is a better result than that achieved with ^{1}H NMR spectra when the X–H signals (i.e., hydrogens not bonded to carbons) were removed from the spectra (5.9%). Hence, the use of HSQC spectra is appropriate.

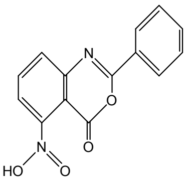
It was shown that the bin method, due to its generality, can be successfully applied also to the comparison of IR spectra. Since these spectra have broad lines, the correlation coefficient can be used efficiently for spectra comparison, hence, the application of the bin method is not as crucial as in the case of ^{1}H NMR or HSQC spectra and the ensuing improvements in the selectivity are also minor. Nonetheless, in some situations, similarities (or dissimilarities) in IR spectra can be detected better with the bin method than by using the correlation coefficient. Although, so far, it has only been tested with one-dimensional ^{1}H NMR, two-dimensional HSQC, and IR spectra, the application of the bin method with spectra of more dimensions including image analysis is straightforward.

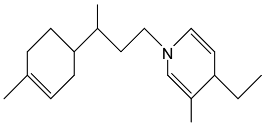
10 Appendix

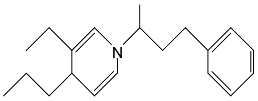
Table 10.1. List of the 116 structures that are clustered in Chapter 8 on the basis of the similarities between their IR spectra.

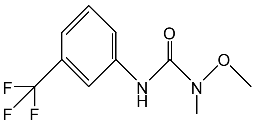
1


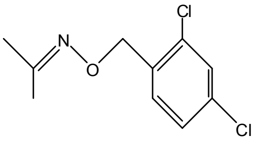
2


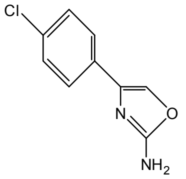
3


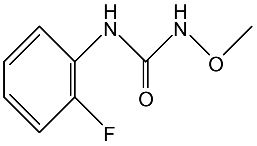
4


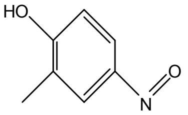
5


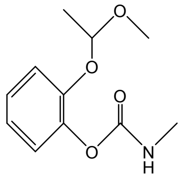
6


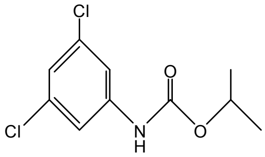
7


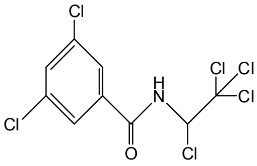
8


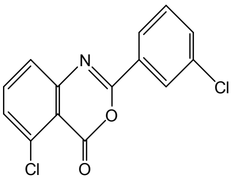
9


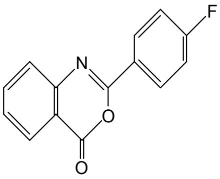
10


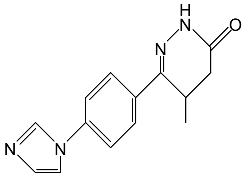
11


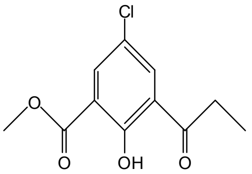
12


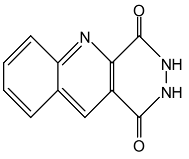
13


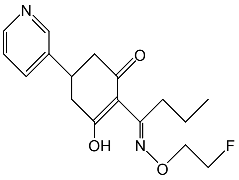
14


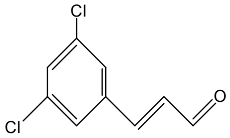
15


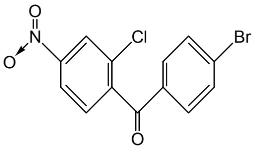
16


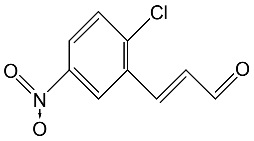
17


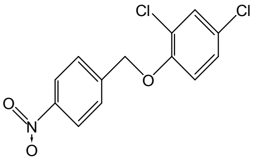
18


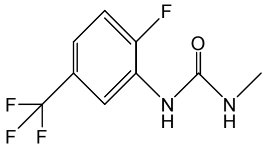
19


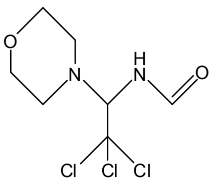
20


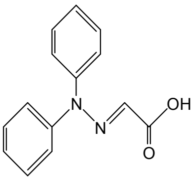
21


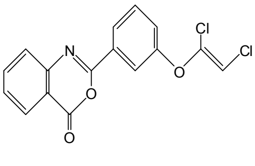
22


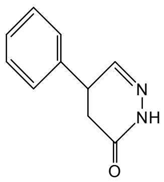
23


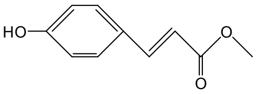
24


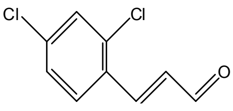
25


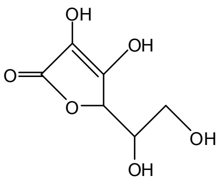
26


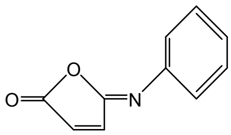
27


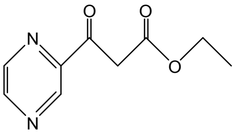
28


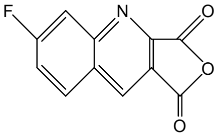
29


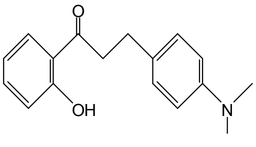
30


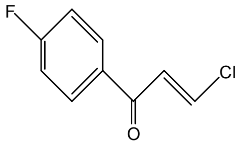
31


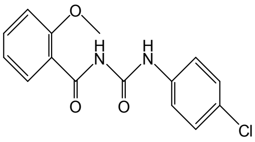
32


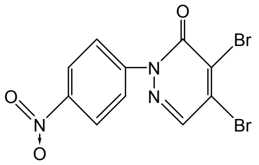
33


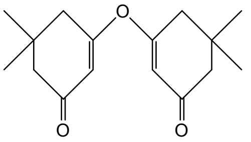
34


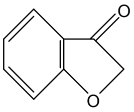
35


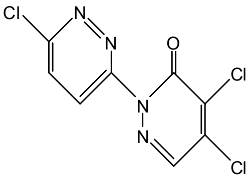
36


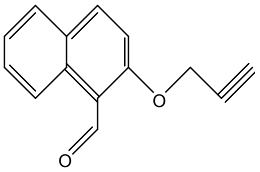
37


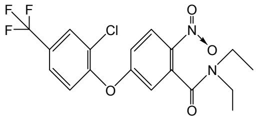
38


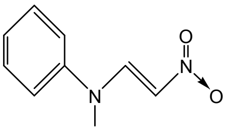
39


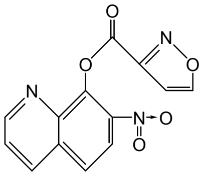
40


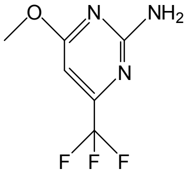
41


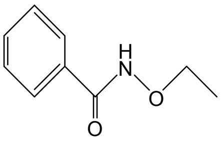
42


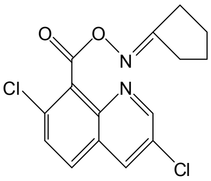
43


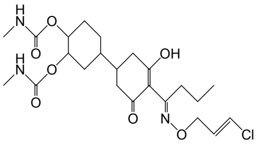
44


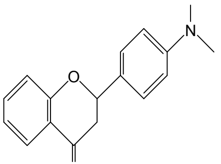
45


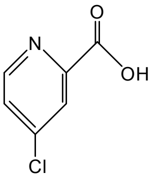
46


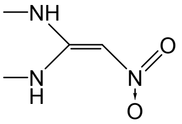
47


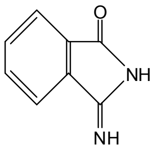
48


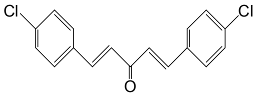
49


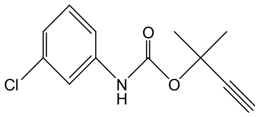
50


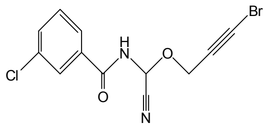
51


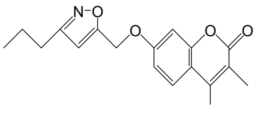
52


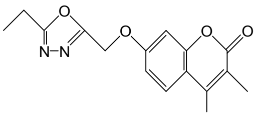
53


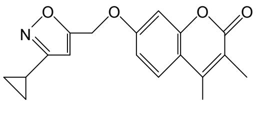
54


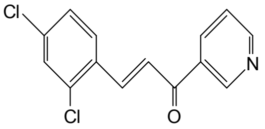
55


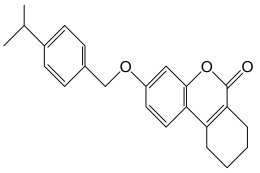
56


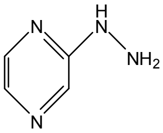
57


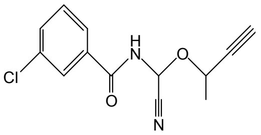
58


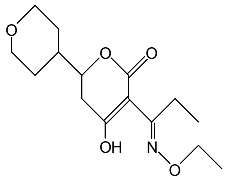
59


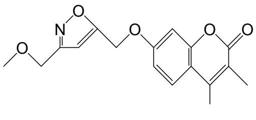
60


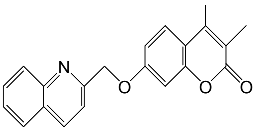
61


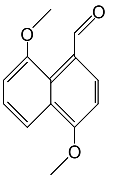
62


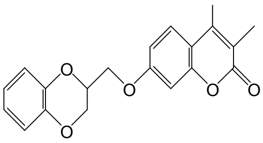
63


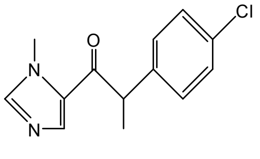
64


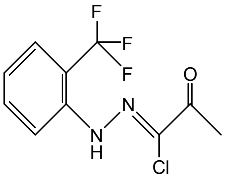
65


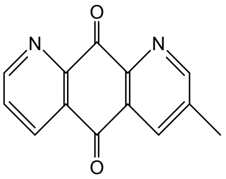
66


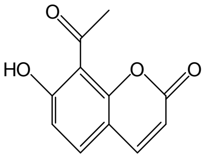
67


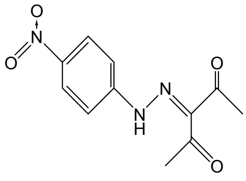
68


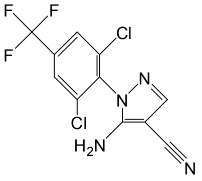
69


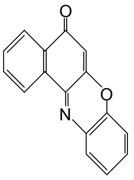
70


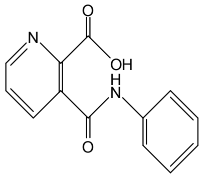
71


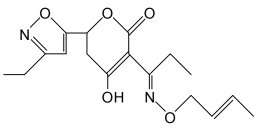
72


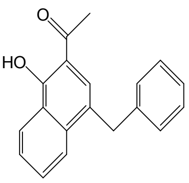
73


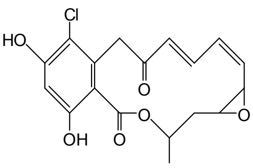
74


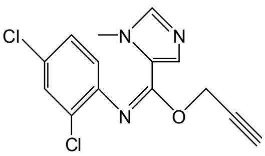
75


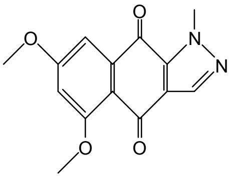
76


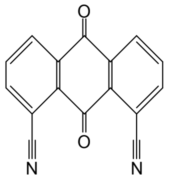
77


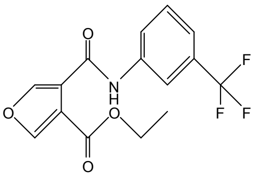
78


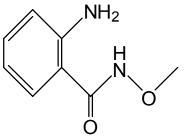
79


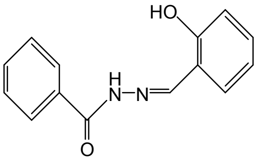
80


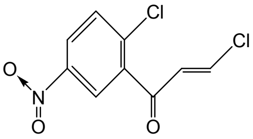
81


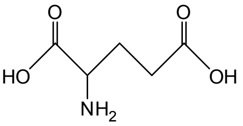
82


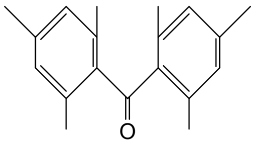
83


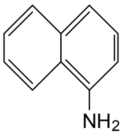
84


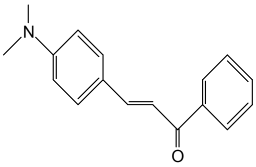
85


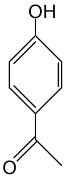
86


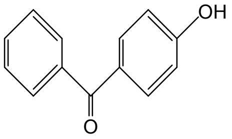
87


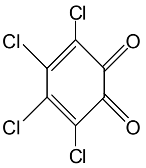
88


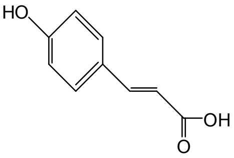
89


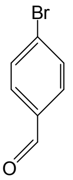
90


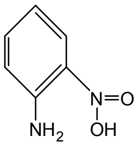
91


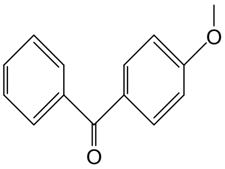
92


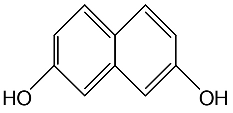
93


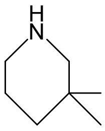
94


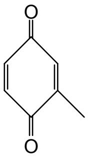
95


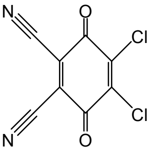
96


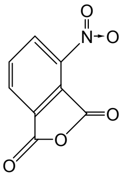
97


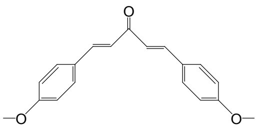
98


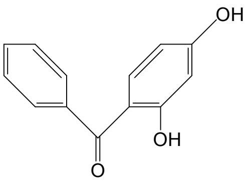
99


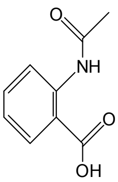
100


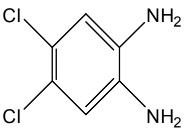
101


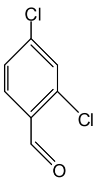
102


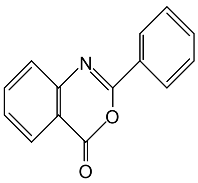
103


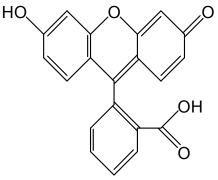
104


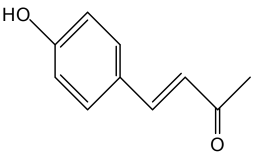
105


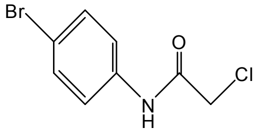
106


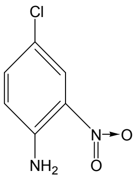
107


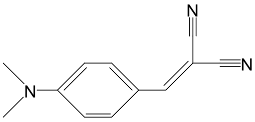
108


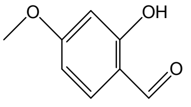
109


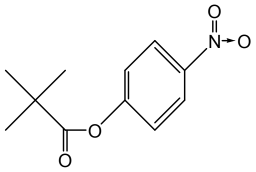
110


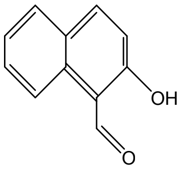
111


112


113


114


115


116


References

1. Jain, A. K.; Murty, M. N.; Flynn, P. J., Data clustering: A review. *ACM Comput. Surv.* **1999**, *31*, 264–323.
2. Linusson, A.; Wold, S.; Nordén, B., Fuzzy clustering of 627 alcohols, guided by a strategy for cluster analysis of chemical compounds for combinatorial chemistry. *Chemom. Intell. Lab. Syst.* **1998**, *44*, 213–227.
3. Krooshof, P. W. T.; Tran, T. N.; Postma, G. J.; Melssen, W. J.; Buydens, L. M. C., Effects of including spatial information in clustering multivariate image data. *TrAC, Trends Anal. Chem.* **2006**, *25*, 1067–1080.
4. Filippini, D.; Lundström, I., Preferential color substances and optimized illuminations for computer screen photo-assisted classification. *Anal. Chim. Acta* **2006**, *557*, 393–398.
5. Sahgal, N.; Monk, B.; Wasil, M.; Magan, N., Trichophyton species: use of volatile fingerprints for rapid identification and discrimination. *Br. J. Dermatol.* **2006**, *155*, 1209–1216.
6. Wong, H. S.; Buenfeld, N. R., Euclidean Distance Mapping for computing microstructural gradients at interfaces in composite materials. *Cem. Concr. Res.* **2006**, *36*, 1091–1097.
7. Wua, Y. S.; van Vliet, L. J.; Frijlink, H. W.; van der Voort Maarschalk, K., The determination of relative path length as a measure for tortuosity in compacts using image analysis. *Eur. J. Pharm. Sci.* **2006**, *28*, 433–440.
8. Li, J.; Hibbert, D. B., Comparison of spectra using a Bayesian approach. An argument using oil spills as an example. *Anal. Chem.* **2005**, *77*, 639–644.
9. Estrada, E., Point scattering: A new geometric invariant with applications from (nano)clusters to biomolecules. *J. Comput. Chem.* **2007**, *28*, 767–777.
10. Pande, M. B. S.; Nagabhushan, P.; Hegde, M. L.; Rao, T. S. S.; Rao, K. S. J., An algorithmic approach to understand trace elemental homeostasis in serum samples of Parkinson disease. *Comput. Biol. Med.* **2005**, *35*, 475–493.
11. Hillenbrand, J. M.; Houde, R. A., A narrow band pattern-matching model of vowel perception. *J. Acoust. Soc. Am.* **2003**, *113*, 1044–1055.
12. Mao, J.; Jain, A. K., A self-organizing network for hyperellipsoidal clustering (HEC). *IEEE Trans. Neural Netw.* **1996**, *7*, 16–29.
13. Nayak, G. S.; Kamath, S.; Pai, K. M.; Sarkar, A.; Ray, S.; Kurien, J.; D'Almeida, L.; Krishnanand, B. R.; Santhosh, C.; Kartha, V. B.; Mahato, K. K., Principal component analysis and artificial neural network analysis of oral tissue fluorescence spectra: classification of normal premalignant and malignant pathological conditions. *Biopolymers* **2006**, *82*, 152–166.
14. Massart, D. L.; Vandeginste, B. G. M.; Buydens, L. M. C.; de Jong, S.; Lewi, P. J.; Smeyers-Verbeke, J., *Handbook of Chemometrics and Qualimetrics: Part A*. Elsevier Science B.V.: Amsterdam, 1997.
15. Das, P.; Datta, S., Exploring the effects of chemical composition in hot rolled steel product using Mahalanobis distance scale under Mahalanobis-Taguchi system. *Comput. Mater. Sci.* **2007**, *38*, 671–677.
16. Taguchi, G.; Chowdhury, S.; Wu, Y., *The Mahalanobis-Taguchi System*. 1st ed.; McGraw-Hill: New York, 2000.

17. Asuero, A. G.; Sayago, A.; González, A. G., The correlation coefficient: An overview. *Crit. Rev. Anal. Chem.* **2006**, *36*, 41–59.
18. Weisstein, E. W., Correlation coefficient. In *MathWorld - A Wolfram Web Resource*, Wolfram Research, Inc.: Champaign, IL, USA, http://mathworld.wolfram.com/CorrelationCoefficient.html, 2007.
19. Pearson, K.; Lee, A., On the laws of inheritance in man: I. Inheritance of physical characters. *Biometrika* **1903**, *2*, 357–462.
20. Pearson, K., On the laws of inheritance in man: II. On the inheritance of the mental and moral characters in man, and its comparison with the inheritance of the physical characters. *Biometrika* **1904**, *3*, 131–190.
21. Galton, F., Co-relations and their measurement, chiefly from anthropometric data. *Proc. Royal Soc. London* **1888**, *45*, 135–145.
22. Abdi, H., Multiple correlation coefficient. In *Encyclopedia of Measurement and Statistics*, 2nd ed.; Salkind, N. J., Ed.; Sage Publications Inc.: Thousand Oaks, CA, USA, 2007.
23. Draper, N.; Smith, H., *Applied Regression Analysis*. 2nd ed.; John Wiley & Sons, Inc.: New York, 1981.
24. Prokhorov, A. V., Partial correlation coefficient. In *Encyclopaedia of Mathematics*, Hazewinkel, M., Ed.; Springer-Verlag: Berlin, http://eom.springer.de/P/p071610.htm, 2002.
25. Rummel, R. J., *Understanding Correlation*; Department of Political Science, University of Hawaii: Honolulu, USA, http://www.hawaii.edu/powerkills/UC.HTM, 1976.
26. Sato, T.; Yamanishi, Y.; Horimoto, K.; Kanehisa, M.; Toh, H., Partial correlation coefficient between distance matrices as a new indicator of protein-protein interactions. *Bioinformatics* **2006**, *22*, 2488–2492.
27. Tanabe, K.; Saeki, S., Computer retrieval of infrared spectra by a correlation coefficient method. *Anal. Chem.* **1975**, *47*, 118–122.
28. Baumann, K.; Clerc, J. T., Computer-assisted IR spectra prediction – linked similarity searches for structures and spectra. *Anal. Chim. Acta* **1997**, *348*, 327–343.
29. Weisstein, E. W., Spearman rank correlation coefficient. In *MathWorld - A Wolfram Web Resource*, Wolfram Research, Inc.: Champaign, IL, USA, http://mathworld.wolfram.com/SpearmanRankCorrelationCoefficient.html, 2007.
30. Kendall, M. G., A new measure of rank correlation. *Biometrika* **1938**, *30*, 81–93.
31. Varmuza, K.; Karlovits, M.; Demuth, W., Spectral similarity versus structural similarity: infrared spectroscopy. *Anal. Chim. Acta* **2003**, *490*, 313–324.
32. Demuth, W.; Karlovits, M.; Varmuza, K., Spectral similarity versus structural similarity: mass spectrometry. *Anal. Chim. Acta* **2004**, *516*, 75–85.
33. Holzgang-Schneiter, H.-E., *Automatische Spektreninterpretation: Überprüfung der Zusammengehörigkeit von Infrarotspektrum und Konstitution*. ETH Zürich: Diss. Nr. 15556, Zürich, 2004.
34. Sievert, H.-J. P.; Drouen, A. C. J. H., Spectral matching and peak purity. In *Diode Array Detection in HPLC*, Huber, L.; George, S. A., Eds.; Marcel Dekker, Inc.: New York, 1993; pp 51–126.
35. Sonnergaard, J. M., On the misinterpretation of the correlation coefficient in pharmaceutical sciences. *Int. J. Pharm.* **2006**, *321*, 12–17.
36. Prince, J. T.; Marcotte, E. M., Chromatographic alignment of ESI-LC-MS proteomics data sets by ordered bijective interpolated warping. *Anal. Chem.* **2006**, *78*, 6140–6152.

37. Šašić, S.; Muszynski, A.; Ozaki, Y., A new possibility of the generalized two-dimensional correlation spectroscopy. 1. Sample-sample correlation spectroscopy. *J. Phys. Chem. A* **2000**, *104*, 6380–6387.

38. Šašić, S.; Sato, H.; Shimoyama, M.; Ozaki, Y., Two-dimensional (2D) correlation coefficient analyses of heavily overlapped near-infrared spectra. *Analyst* **2005**, *130*, 652–658.

39. Skvortsova, Y.; Wang, G.; Geng, M. L., Statistical two-dimensional correlation coefficient mapping of simulated tissue phantom data: Boundary determination in tissue classification for cancer diagnosis. *J. Mol. Struct.* **2006**, *799*, 239–246.

40. Fisher, R. A., Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika* **1915**, *10*, 507–521.

41. Wang, G.; Karnes, J.; Bunker, C. E.; Geng, M. L., Two-dimensional correlation coefficient mapping in gas chromatography: Jet fuel classification for environmental analysis. *J. Mol. Struct.* **2006**, *799*, 247–252.

42. Liu, Y.; Meng, Q.; Chen, R.; Wang, J.; Jiang, S.; Hu, Y., A new method to evaluate the similarity of chromatographic fingerprints: Weighted Pearson product-moment correlation coefficient. *J. Chromatogr. Sci.* **2004**, *42*, 545–550.

43. Stein, S. E.; Scott, D. R., Optimization and testing of mass spectral library search algorithms for compound identification. *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 859–866.

44. Lappi, S. E.; Franzen, S., Eigenvector mapping: a method for discerning solvent effects on vibrational spectra. *Spectrochim. Acta Part A* **2004**, *60*, 357–370.

45. Tabb, D. L.; MacCoss, M. J.; Wu, C. C.; Anderson, S. D.; John R. Yates, I., Similarity among tandem mass spectra from proteomic experiments: detection, significance, and utility. *Anal. Chem.* **2003**, *75*, 2470–2477.

46. Frewen, B. E.; Merrihew, G. E.; Wu, C. C.; Noble, W. S.; MacCoss, M. J., Analysis of peptide MS/MS spectra from large-scale proteomics experiments using spectrum libraries. *Anal. Chem.* **2006**, *78*, 5678–5684.

47. Sadygov, R. G.; Martin Maroto, F.; Hühmer, A. F. R., ChromAlign: A two-step algorithmic procedure for time alignment of three-dimensional LC-MS chromatographic surfaces. *Anal. Chem.* **2006**, *78*, 8207–8217.

48. *NIST Standard Reference Database 1A*. National Institute of Standards and Technology: Gaithersburg, MD, USA, http://www.nist.gov/srd/nist1a.htm, 2007.

49. McLafferty, F. W.; Hertel, R. H.; Villwock, R. D., Probability based matching of mass spectra. *Org. Mass Spectrom.* **1974**, *9*, 690–702.

50. Hertz, H. S.; Hites, R. A.; Biemann, K., Identification of mass spectra by computer-searching a file of known spectra. *Anal. Chem.* **1971**, *43*, 681–691.

51. Tanimoto, T. T., *An elementary mathematical theory of classification and prediction*; IBM Corporation, Watson Research Center: Kingston, NY, USA, 17th November, 1958; pp 30–39.

52. Rogers, D. J.; Tanimoto, T. T., A computer program for classifying plants. *Science* **1960**, *132*, 1115–1118.

53. Jaccard, P., Distribution de la flore alpine dans le bassin des Dranses et dans quelques régions voisines. *Bull. Soc. Vaud. Sci. Nat.* **1901**, *37*, 241–272.

54. Jaccard, P., The distribution of the flora in the alpine zone. *The New Phytologist* **1912**, *11*, 37–50.

55. Willett, P.; Winterman, V., Implementation of nearest-neighbor searching in an online chemical structure search system. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 36–41.

56. Clerc, J.-T.; Terkovics, A. L., Versatile topological structure descriptor for quantitative structure/property studies. *Anal. Chim. Acta* **1990**, *235*, 93–102.

57. Willett, P., Similarity-based approaches to virtual screening. *Biochem. Soc. Trans.* **2003**, *31*, 603–606.

58. Willett, P.; Barnard, J. B.; Downs, G. M., Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996.

59. Willett, P., Similarity-based virtual screening using 2D fingerprints. *Drug Discov. Today* **2006**, *11*, 1046–1053.

60. Fechner, U.; Franke, L.; Renner, S.; Schneider, P.; Schneider, G., Comparison of correlation vector methods for ligand-based similarity searching. *J. Comput.-Aided Mol. Des.* **2003**, *17*, 687–698.

61. Webb, A. R., *Statistical Pattern Recognition*. 2nd ed.; John Wiley and Sons Ltd.: Malvern, UK, 2002.

62. Fukunaga, K., *Introduction to Statistical Pattern Recognition*. 2nd ed.; Academic Press, Inc.: Boston, MA, USA, 1990.

63. Massart, D. L.; Vandeginste, B. G. M.; Buydens, L. M. C.; de Jong, S.; Lewi, P. J.; Smeyers-Verbeke, J., *Handbook of Chemometrics and Qualimetrics: Part B*. Elsevier Science B.V.: Amsterdam, 1998.

64. Lavine, B. K., Clustering and classification of analytical data. In *Encyclopedia of Analytical Chemistry: Instrumentation and Applications*, Meyers, R. A., Ed.; John Wiley & Sons, Inc.: Chichester, UK, 2000; pp 9689–9710.

65. Halkidi, M.; Batistakis, Y.; Vazirgiannis, M., On clustering validation techniques. *J. Intell. Inf. Syst.* **2001**, *17*, 107–145.

66. Ros, F.; Pintore, M.; Chrétien, J. R., Molecular descriptor selection combining genetic algorithms and fuzzy logic: application to database mining procedures. *Chemom. Intell. Lab. Syst.* **2002**, *63*, 15–26.

67. Teppola, P.; Mujunen, S.-P.; Minkkinen, P., Adaptive Fuzzy C-Means clustering in process monitoring. *Chemom. Intell. Lab. Syst.* **1999**, *45*, 23–38.

68. Noordam, J. C.; van den Broek, W. H. A. M., Multivariate image segmentation based on geometrically guided fuzzy C-means clustering. *J. Chemom.* **2002**, *16*, 1–11.

69. Smoliński, A.; Walczak, B.; Einax, J. W., Hierarchical clustering extended with visual complements of environmental data set. *Chemom. Intell. Lab. Syst.* **2002**, *64*, 45–54.

70. Liang, J.; Kachalo, S., Computational analysis of microarray gene expression profiles: clustering, classification, and beyond. *Chemom. Intell. Lab. Syst.* **2002**, *62*, 199–216.

71. Bondarenko, I.; Malderen, H. V.; Treiger, B.; Espen, P. V.; Grieken, R. V., Hierarchical cluster analysis with stopping rules built on Akaike's information criterion for aerosol particle classification based on electron probe X-ray microanalysis. *Chemom. Intell. Lab. Syst.* **1994**, *22*, 87–95.

72. Daszykowsk, M.; Walczak, B.; Massart, D. L., Looking for natural patterns in data. Part 1. Density-based approach. *Chemom. Intell. Lab. Syst.* **2001**, *56*, 83–92.

73. Ward, J. H., Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.* **1963**, *58*, 236–244.

74. El-Haudouchi, A.; Willett, P., Hierarchic document clustering using Ward's method. In *Proceedings of the 9th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM Press: Palazzo dei Congressi, Pisa, Italy, 1986; pp 149–156.

75. Linkage Function. In *Statistical Glossary*, http://www.statistics.com/resources/glossary/l/linkage.php, 2007.
76. Tran, T. N.; Wehrens, R.; Buydens, L. M. C., Clustering multispectral images: a tutorial. *Chemom. Intell. Lab. Syst.* **2005**, *77*, 3–17.
77. Cheng, Y., Mean shift, mode seeking, and clustering. *IEEE Trans. Pattern Anal. Mach. Intell.* **1995**, *17*, 790–799.
78. Teppola, P.; Mujunen, S.-P.; Minkkinen, P., A combined approach of partial least squares and fuzzy c-means clustering for the monitoring of an activated-sludge waste-water treatment plant. *Chemom. Intell. Lab. Syst.* **1998**, *41*, 95–103.
79. Zürcher, M.; Clerc, J. T.; Farkas, M.; Pretsch, E., General theory of similarity measures for library search systems. *Anal. Chim. Acta* **1988**, *206*, 161–172.
80. Farkas, M.; Bendl, J.; Welti, D. H.; Pretsch, E.; Dütsch, S.; Portmann, P., Similarity search for a ^{1}H-NMR spectroscopic data base. *Anal. Chim. Acta* **1988**, *206*, 173–187.
81. Martinsen, D. P.; Song, B.-H., Computer applications in mass spectral interpretation: A recent review. *Mass Spectrom. Rev.* **1985**, *4*, 461–490.
82. Hansen, M. E.; Smedsgaard, J., A new matching algorithm for high resolution mass spectra. *J. Am. Soc. Mass Spectrom.* **2004**, *15*, 1173–1180.
83. Ellison, S. L. R.; Gregory, S. L., Predicting chance infrared spectroscopic matching frequencies. *Anal. Chim. Acta* **1998**, *370*, 181–190.
84. Weisstein, E. W., Combination. In *MathWorld - A Wolfram Web Resource*, Wolfram Research, Inc.: Champaign, IL, USA, http://mathworld.wolfram.com/Combination.html, 2007.
85. Nielsen, N.-P. V.; Carstensen, J. M.; Smedsgaard, J., Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping. *J. Chromatogr. A* **1998**, *805*, 17–35.
86. Johnson, K. J.; Wright, B. W.; Jarman, K. H.; Synovec, R. E., High-speed peak matching algorithm for retention time alignment of gas chromatographic data for chemometric analysis. *J. Chromatogr. A* **2003**, *996*, 141–155.
87. Pravdova, V.; Walczak, B.; Massart, D. L., A comparison of two algorithms for warping of analytical signals. *Anal. Chim. Acta* **2002**, *456*, 77–92.
88. Forshed, J.; Schuppe-Koistinen, I.; Jacobsson, S. P., Peak alignment of NMR signals by means of a genetic algorithm. *Anal. Chim. Acta* **2003**, *487*, 189–199.
89. Møller, S. F.; Jørgensen, B. M., Peak alignment and robust principal component analysis of gas chromatograms of fatty acid methyl esters and volatiles. *J. Chromatogr. Sci.* **2007**, *45*, 169–176.
90. Walczak, B.; Wu, W., Fuzzy warping of chromatograms. *Chemom. Intell. Lab. Syst.* **2005**, *77*, 173–180.
91. Wu, W.; Daszykowski, M.; Walczak, B.; Sweatman, B. C.; Connor, S. C.; Haselden, J. N.; Crowther, D. J.; Gill, R. W.; Lutz, M. W., Peak alignment of urine NMR spectra using fuzzy warping. *J. Chem. Inf. Model.* **2006**, *46*, 863–875.
92. Dunn, W. B.; Ellis, D. I., Metabolomics: Current analytical platforms and methodologies. *TrAC, Trends Anal. Chem.* **2005**, *24*, 285–294.
93. Lindon, J. C.; Holmes, E.; Nicholson, J. K., So what's the deal with metabonomics? *Anal. Chem.* **2003**, *75*, 385A–391A.
94. Craig, A.; Cloarec, O.; Holmes, E.; Nicholson, J. K.; Lindon, J. C., Scaling and normalization effects in nmr spectroscopic metabonomic data sets. *Anal. Chem.* **2006**, *78*, 2262–2267.

95. Cloarec, O.; Dumas, M. E.; Trygg, J.; Craig, A.; Barton, R. H.; Lindon, J. C.; Nicholson, J. K.; Holmes, E., Evaluation of the orthogonal projection on latent structure model limitations caused by chemical shift variability and improved visualization of biomarker changes in ^{1}H NMR spectroscopic metabonomic studies. *Anal. Chem.* **2005**, *77*, 517–526.

96. Bollard, M. E.; Stanley, E. G.; Lindon, J. C.; Nicholson, J. K.; Holmes, E., NMR-based metabonomic approaches for evaluating physiological influences on biofluid composition. *NMR Biomed.* **2005**, *18*, 143–162.

97. Holmes, E.; Nicholson, J. K.; Nicholls, A. W.; Lindon, J. C.; Connor, S. C.; Polley, S.; Connelly, J., The identification of novel biomarkers of renal toxicity using automatic data reduction techniques and PCA of proton NMR spectra of urine. *Chemom. Intell. Lab. Syst.* **1998**, *44*, 245–255.

98. *Chemometrics applied to a metabonomic study of mouse urine*. Vol. InfoMetrix 38-04/04; Application note. Infometrix Inc.: Bothell, WA, USA, http://www.infometrix.com/apps/38-0404_MetaboChemoAN.pdf, 2004.

99. Bürgin Schaller, R.; Pretsch, E., A computer program for the automatic estimation of ^{1}H NMR chemical shifts. *Anal. Chim. Acta* **1994**, *290*, 295–302.

100. Bürgin Schaller, R.; Munk, M. E.; Pretsch, E., Spectra estimation for computer-aided structure determination. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 239–243.

101. Kalelkar, S.; Dow, E. R.; Grimes, J.; Clapham, M.; Hu, H., Automated analysis of proton NMR spectra from combinatorial rapid parallel synthesis using self-organizing maps. *J. Comb. Chem.* **2002**, *4*, 622–629.

102. Karfunkel, H. R.; Rohde, B.; Leusen, F. J. J.; Gdanitz, R. J.; Rihs, G., Continuous similarity measure between nonoverlapping X-ray powder diagrams of different crystal modifications. *J. Comput. Chem.* **1993**, *14*, 1125–1135.

103. Stephenson, D. S.; Binsch, G., Automated analysis of high-resolution NMR spectra I. Principles and computational strategy. *J. Magn. Reson.* **1980**, *37*, 395–407.

104. de Gelder, R.; Wehrens, R.; Hageman, J. A., A generalized expression for the similarity of spectra: application to powder diffraction pattern classification. *J. Comput. Chem.* **2001**, *22*, 273–289.

105. Dods, J.; Gruner, D.; Brumer, P., A genetic algorithm approach to fitting polyatomic spectra via geometry shifts. *Chem. Phys. Lett.* **1996**, *261*, 612–619.

106. Lawton, S. L.; Bartell, L. S., Application of the overlap integral in X-ray diffraction powder pattern recognition. *Powder Diffr.* **1994**, *9*, 124–135.

107. Pearson, K., On the theory of contingency and its relation to association and normal correlation. In *Draper's Company Research Memoirs, Biometric Series, No. 1*, Cambridge University Press: Cambridge, UK, 1904.

108. Pearson, K., On the theory of contingency. *J. Am. Stat. Assoc.* **1930**, *25*, 320–327.

109. Chernoff, H.; Lehmann, E. L., The use of maximum likelihood estimates in χ^2 tests for goodness of fit. *Ann. Math. Statist.* **1954**, *25*, 579–586.

110. Filliben, J. J.; Heckert, A., Exploratory Data Analysis. In *NIST/SEMATECH e-Handbook of Statistical Methods*, Croarkin, C.; Tobias, P., Eds.; Gaithersburg, MD, USA, http://www.itl.nist.gov/div898/handbook/eda/section3/eda35f.htm, 2006.

111. Weisstein, E. W., Fisher's exact test. In *MathWorld - A Wolfram Web Resource*, Wolfram Research, Inc.: Champaign, IL, USA, http://mathworld.wolfram.com/FishersExactTest.html, 2007.

112. Sawae, R.; Sakata, T.; Takarabe, K.; Tei, M.; Mizuno, Y.; Mori, Y.; Manmoto, Y., Quantum random walks on the set of contingency tables. *Int. J. Quantum Chem.* **2002**, *90*, 1321–1325.

113. Héberger, K.; Rajkó, R., Generalization of pair correlation method (PCM) for non-parametric variable selection. *J. Chemom.* **2002**, *16*, 436–443.

114. Weber, K. A.; Perry, R. G., Groundwater abstraction impacts on spring flow and base flow in the Hillsborough River Basin, Florida, USA. *Hydrogeology J.* **2006**, *14*, 1252–1264.

115. Lancaster, H. O., *An Introduction to Medical Statistics*. John Wiley and Sons, Inc.: New York, 1974.

116. Holub, O.; Ferreira, S. T., Quantitative histogram analysis of images. *Comput. Phys. Commun.* **2006**, *175*, 620–623.

117. Niesner, R.; Gericke, K.-H., Quantitative determination of the single-molecule detection regime in fluorescence fluctuation microscopy by means of photon counting histogram analysis. *J. Chem. Phys.* **2006**, *124*, 134704-1–134704-8.

118. Peters, B., Using the histogram test to quantify reaction coordinate error. *J. Chem. Phys.* **2006**, *125*, 241101-1–241101-4.

119. Chodera, J. D.; Swope, W. C.; Pitera, J. W.; Seok, C.; Dill, K. A., Use of the weighted histogram analysis method for the analysis of simulated and parallel tempering simulations. *J. Chem. Theory Comput.* **2007**, *3*, 26–41.

120. Zhang, W.; Yeo, B. S.; Schmid, T.; Zenobi, R., Single molecule tip-enhanced Raman spectroscopy with silver tips. *J. Phys. Chem. C* **2007**, *111*, 1733–1738.

121. O'Halloran, R. L.; Holmes, J. H.; Altes, T. A.; Salerno, M.; Fain, S. B., The effects of SNR on ADC measurements in diffusion-weighted hyperpolarized He-3 MRI. *J. Magn. Reson.* **2007**, *185*, 42–49.

122. Bodis, L.; Ross, A.; Pretsch, E., Novel similarity measure for comparison of spectra. In *Abstr. Paper Am. Chem. Soc. Natl. Meet. 231*, 48-CINF, American Chemical Society: Atlanta, GA, USA, 2006.

123. Bodis, L.; Ross, A.; Pretsch, E., A novel spectra similarity measure. *Chemom. Intell. Lab. Syst.* **2007**, *85*, 1–8.

124. Ross, A.; Schlotterbeck, G.; Senn, H.; von Kienlin, M., Application of chemical shift imaging for simultaneous and fast acquisition of NMR spectra on multiple samples. *Angew. Chem. Int. Ed.* **2001**, *40*, 3243–3245.

125. Macnaughtan, M. A.; Hou, T.; Xu, J.; Raftery, D., High-throughput nuclear magnetic resonance analysis using a multiple coil flow probe. *Anal. Chem.* **2003**, *75*, 5116–5123.

126. Raftery, D., High-throughput NMR spectroscopy. *Anal. Bioanal. Chem.* **2004**, *378*, 1403–1404.

127. *SpecInfo*. Chemical Concepts GmbH: P.O. Box 100202, D-69442 Weinheim, Germany, 1998.

128. McDonald, R. S.; Paul A. Wilks, J., JCAMP-DX: A standard form for exchange of infrared spectra in computer readable form. *Appl. Spectrosc.* **1988**, *42*, 151–162.

129. Davies, A. N.; Lampen, P., JCAMP-DX for NMR. *Appl. Spectrosc.* **1993**, *47*, 1093–1099.

130. Lampen, P.; Lambert, J.; Lancashire, R. J.; McDonald, R. S.; McIntyre, P. S.; Rutledge, D. N.; Fröhlich, T.; Davies, A. N., An extension to the JCAMP-DX standard file format, JCAMP-DX V.5.01. *Pure Appl. Chem.* **1999**, *71*, 1549–1556.

131. Dalby, A.; Nourse, J. G.; Hounshell, W. D.; Gushurst, A. K. I.; Grier, D. L.; Leland, B. A.; Laufer, J., Description of several chemical structure file formats used by computer programs developed at molecular design limited. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 244–255.

132. *MDL CTfile Formats*. MDL Information Systems, Inc.: San Leandro, CA, USA, http://www.mdli.com/downloads/public/ctfile/ctfile.jsp, 2005.

133. *NMRPrediction*, Porta Nova Software GmbH: Imfeldstr. 37, CH-8037 Zürich, Switzerland, http://www.upstream.ch/products/nmr.html.

134. *Borland Delphi*, CodeGear: Scotts Valley, CA, USA, http://www.codegear.com/products/delphi.

135. Pretsch, E.; Bühlmann, P.; Affolter, C., *Structure Determination of Organic Compounds*. 3rd ed.; Springer-Verlag: Berlin, 2000.

136. Golotvin, S. S.; Vodopianov, E.; Lefebvre, B. A.; Williams, A. J.; Spitzer, T. D., Automated structure verification based on ^{1}H NMR prediction. *Magn. Reson. Chem.* **2006**, *44*, 524–538.

137. Schröder, H.; Neidig, P.; Rossé, G., High-throughput structure verification of a substituted 4-phenylbenzopyran library by using 2D NMR techniques. *Angew. Chem. Int. Ed.* **2000**, *39*, 3816–3819.

138. *SPSS*, SPSS Inc.: Chicago, IL, USA, http://www.spss.com/spss.

139. Cavanagh, J.; Fairbrother, W. J.; Palmer, A. G., III; Rance, M.; Skelton, N. J., *Protein NMR Spectroscopy: Principles and Practice*. 2nd ed.; Elsevier Academic Press: Amsterdam, 2007.

140. Griffiths, L.; Horton, R., Towards the automatic analysis of NMR spectra: Part 6. Confirmation of chemical structure employing both ^{1}H and ^{13}C NMR spectra. *Magn. Reson. Chem.* **2006**, *44*, 139–145.

141. Golotvin, S. S.; Vodopianov, E.; Pol, R.; Lefebvre, B. A.; Williams, A. J.; Spitzer, T. D., *Automated evaluation of a chemical structure with only 1D ^{1}H and 2D ^{1}H-^{13}C HSQC*. Poster presented at Experimental Nuclear Magnetic Resonance Conference (ENC): Pacific Grove, CA, USA, 2006.

142. *MestReC*, Mestrelab Research: Santiago de Compostela, Spain, http://www.mestrec.com/producto.php?id=3.

143. *Bruker TopSpin*, Bruker BioSpin GmbH: D-76287 Rheinstetten, Germany, http://www.bruker-biospin.com/topspin.html.

144. *MATLAB*, The MathWorks: Natick, MA, USA, http://www.mathworks.com/products/matlab.

145. *TQ Analyst*, Thermo Fisher Scientific, Inc.: Waltham, MA, USA, http://www.thermo.com/com/cda/product/detail/0,1055,1000001344641,00.html.

Curriculum vitae

Personal details

Date of birth	March 23, 1980
Place of birth	Cluj/Klausenburg, Romania
Marital status	single

Education

07/2004 – 07/2007	**ETH Zurich, Switzerland – Dr. sc. ETH Zurich** PhD student and research assistant at the Department of Chemistry and Applied Biosciences Research project financed by **F. Hoffmann-La Roche AG, Basel, Switzerland** Doctoral thesis: *Quantification of Spectral Similarity: Towards Automatic Spectra Verification* Several attended chemistry and computer science lectures and taken exams
11/2003 – 04/2004	**University of Paderborn, Germany** Graduate student at the Faculty of Computer Science, Electrical Engineering and Mathematics Research project: *Mixed Integer Linear Programming for Airline Optimization*, in cooperation with Lufthansa Systems
10/2003 – 06/2004	**Babes-Bolyai University, Cluj, Romania – Master in Computer Science** Master studies in Computer Science, Intelligent Systems (in English) Master thesis: *Financial Time Series Forecasting Using Artificial Neural Networks* Average grade of the semesters of study: 10 out of 10 (top 2%)
10/2002 – 06/2003	**Technical University of Munich, Germany** Undergraduate student at the Faculty of Computer Science
08/2001	**Eotvos Lorand University, Budapest, Hungary** Summer course at the Faculty of Sciences, Department of Computer Science
10/1999 – 06/2001	**Babes-Bolyai University, Cluj, Romania** Faculty of Business, specialization in Management
10/1999 – 03/2003	**Babes-Bolyai University, Cluj, Romania – Teacher-training Diploma** Department of Teachers Training: attended teaching modules (7 semesters) for psychology, pedagogy, computer science teaching methodology and teaching practice
10/1999 – 06/2003	**Babes-Bolyai University, Cluj, Romania – Diploma in Computer Science** Undergraduate student at the Faculty of Mathematics and Computer Science Diploma work: *Genetic Algorithms in Game Theory* Average grade of the years of study: 9.91 out of 10 (top 2%)
09/1995 – 06/1999	**Bathory Istvan High School, Cluj, Romania – High School Leaving Diploma** Average grade of the eight exams: 9.17 out of 10

Professional experience

07/2003 – 09/2003	**ETH Zurich, Switzerland** Academic guest Research project: *Automatic Structure-Spectra Compatibility Tests*
09/2001	**Richter Gedeon Ltd., Targu-Mures, Romania** Practical training: Marketing research
07/2001	**ComGenex Inc., Budapest, Hungary** Practical training: Web programming

Conferences, workshops, talks, and publications

June 8 – 10, 2007	**McKinsey DIVE Workshop, Nice, France** Three-day workshop, teamwork on solving a real-case study
February 19 – 21, 2007	**Didactic Workshop, Didaktikzentrum, ETH Zurich** Teaching and learning in higher education; motivating and activating students; rhetoric and presentation skills; video analysis
September 2 – 9, 2006	**Cortona Week Workshop, Cortona, Tuscany, Italy** Interdisciplinary, transdisciplinary residential week and conference organized by the ETH Zurich
11/2004 – 04/2006	**F. Hoffmann-La Roche AG, Basel, Switzerland** Several talks given and technical reports submitted
March 26 – 30, 2006	**231st ACS National Meeting, Atlanta, GA, USA** Poster presented, title: *Novel Similarity Measure for Comparing Spectra* *Winner of the* ***CINF-IO Informatics Scholarship for Scientific Excellence***
August 28 – 31, 2005	**Conferentia Chemometrica, Hajdúszoboszló, Hungary** Talk given, title: *Novel Similarity Criteria for Comparison of Spectra*
Publications	Bodis, L.; Ross, A.; Pretsch, E., A novel spectra similarity measure. *Chemom. Intell. Lab. Syst.* **2007**, *85*, 1–8. Bodis, L.; Ross, A.; Pretsch, E., Novel similarity measure for comparison of spectra. *Abstr. Paper Am. Chem. Soc. Natl. Meet.* **2006**, *231*, 48-CINF. Bodis, J.; Bodis, L.; Lercher, J. A., Mass transfer rates during three phase catalytic reductive amination over supported noble metals. *Studia-Chem.* **2005**, *50*, 229–236.

Skills

Programming languages	Pascal/Delphi, C/C++/Visual C++, Visual Basic, Java, Assembly, Matlab, LISP, Prolog, UML Windows programming (API, MFC, ActiveX DLL/Control) UNIX/Linux programming (shell, ANSI C, C++, IPC, Client/Server) Web programming (HTML, CGI, JSP, JavaScript, Web Services, XML) Databases (ADO, SQL, MS SQL Server, Oracle, Fox Pro)
Languages	**Hungarian**: mother tongue **Romanian**: native fluency, High School Leaving diploma **English**: fluent, university and master studies **German**: fluent, university studies in Germany and Switzerland

Zürich, July 2007 Lorant Bodis