I N S T I T U T D E S T A T I S T I Q U E
B I O S T A T I S T I Q U E E T
S C I E N C E S A C T U A R I E L L E S
(I S B A)

UNIVERSITÉ CATHOLIQUE DE LOUVAIN

D I S C U S S I O N
P A P E R

2014/33

Statistical treatment of 2D-NMR COSY spectra: data preparation, clustering-based repeatability evaluation and comparison with ^{1}H-NMR

FÉRAUD, B., GOVAERTS, B., VERLEYSEN, M. and P. DE TULLIO

Noname manuscript No.
(will be inserted by the editor)

Statistical treatment of 2D-NMR COSY spectra: data preparation, clustering-based repeatability evaluation and comparison with ^{1}H-NMR

Baptiste Féraud · Bernadette Govaerts · Michel Verleysen · Pascal de Tullio

Received: date / Accepted: date

Abstract Compared with the widely used ^{1}H-NMR spectroscopy, two-dimensional NMR experiments provides more sophisticated spectra which should facilitate the identification of relevant spectral zones or biomarkers. This paper focusses on proton ^{1}H-^{1}H-COSY (COrrelation SpectroscopY) spectral data. In spite of longer inherent acquisition times, it is commonly accepted by users (biologists, healthcare professionals) that the introduction of an additional dimension represents a huge qualitative step for investigations in terms of metabolites identification. In other words, it seems natural that more information leads to more predictive power. But, until now, no statistical study clearly proved this assumption. Therefore a fundamental question is "Is this supplementary information relevant?". In order to extend the statistical properties developed for 1D spectroscopy to the challenges raised by 2D spectra, a rigorous study of the repeatability of COSY spectra is needed as a prerequisite. Having introduced new pre-processing concepts, such as the Global Peak List or an ad hoc 2D "bucketing", this paper presents an innovative methodology based on multivariate clustering algorithms to evaluate this

Baptiste Féraud
Institute of Statistics, Biostatistics and Actuarial Sciences (ISBA), Université Catholique de Louvain (UCL), Belgium
Machine Learning Group, UCL, Belgium
Address: Voie du Roman Pays 20, bte L1.04.01, B-1348 Louvain-la-Neuve, Belgium
Tel.: +32 10473053
E-mail: baptiste.feraud@uclouvain.be

Bernadette Govaerts
Institute of Statistics, Biostatistics and Actuarial Sciences (ISBA), Université Catholique de Louvain (UCL), Belgium

Michel Verleysen
Machine Learning Group, Université Catholique de Louvain (UCL), Belgium
SAMM, Université Paris I, Panthéon - Sorbonne, France

Pascal de Tullio
Center for Interdisciplinary Research on Medicines (CIRM), Université de Liège (ULg), Belgium

question. Numerical clustering quality indexes and graphical results are proposed, based both on the spectral presence or absence of peaks and on peak intensities, and through different levels of spectral resolution. The methodology is applied to two real experimental designs: a 4-mixture cell culture media containing various supervised metabolites and a complex human serum based design. The second goal of this paper is to compare clustering performances obtained on COSY and on ^{1}H-NMR spectra, with the aim of understanding to what extent the COSY spectra carry more metabolomic informative content about the signal than 1D ones. It is shown that COSY spectra appear to be statistically repeatable and, in addition, provide better clustering results than corresponding ^{1}H-NMR when using unlabeled information.

Keywords COSY spectra · Repeatability and metabolomic informative content · Peak lists · Multivariate clustering algorithms · Real data

1 Introduction

The common question in medical metabolomics studies [19] consists in searching for metabolites interpretable as biomarkers for a given disease, pathology or toxicity. These biomarkers are usually tracked in partial or complete spectral images, linked with biofluids (serum, urine) or tissues, using adequate univariate or multivariate statistical techniques. It is clear that the sampling preparation, spectral acquisition and data post-processing have a crucial impact on the global quality of the data and on the chances to finally discover relevant biomarkers. Furthermore, an accurate statistical analysis will rarely compensate "dirty" initial data. Many acquisition and post-processing tools are available and it is often difficult to objectively and quantitatively evaluate which ones will provide the more informative data in a given context. This motivated the development of the methodology described in this paper when data are organized in groups.

In this context, proton nuclear magnetic resonance (^{1}H-NMR) spectroscopy generates spectral profiles describing the metabolite composition of collected samples. A comparison of several spectra of metabolites in various specific states permits a preliminary graphical and qualitative investigation of changes in metabolite composition inherent to the presence of a "stressor". However, the complexity of ^{1}H-NMR spectra and the number of spectra (of samples) usually available in metabolomics studies require a semi-automated data analysis. In addition, systematic differences between samples are often hidden behind biological noise and/or behind peak shifts.

To avoid these peak shifts and the usual overlappings of potentially independent signals, the use of two-dimensional homonuclear or heteronuclear experiments has gained attention these last years to identify metabolites. However, these tools are often left behind due to longer, and sometimes prohibitive, acquisition times. In this paper, COSY spectra are investigated [11] [1]. COSY consists in a correlation-based method for determining which signals arise from neighboring protons (usually up to four bonds). Correlations appear when there is spin-spin coupling between protons (i.e. correlation between two or more nearby chemical processes). Several works exist in the literature in order to get around the time-consumming acquisition problem, specifically for multidimensional COSY data (for

example Ultrafast-COSY [4] [21] or PALSY-COSY [26]). Compared to ^{1}H-NMR experiments for instance, the introduction of an additional dimension should allow a better representation of metabolites, a better predictive power and a better biomarker identification. Several works are already based on 2D-COSY or on faster extentions of COSY, such as in [28], [30] or [5]. But, until now, no statistical study clearly proved the superiority of 2D spectra. This lack leads to a central and fundamental question: is this supplementary information really relevant in the context of metabolomics analyses? If it is the case, the interest of using two-dimensional methods would be legitimated, even at the price of potentially superior acquisition times.

In this paper, "Repeatability" is a key notion. By repeatability, the intention is to evaluate the amount of captured information (i.e. the extent to which signals are captured) compared with the noisy part through several spectra. It goes away from the more classic version of repeatability in spectrometry which involves only one spectrum at a time. Generally, a measure can be considered as the sum of a signal (useful information) and noise. In many metabolomics studies, the signal is controlled and often linked with the existence of different mixtures, different samples or different groups of people (people affected by a disease vs. healthy people, people affected by a disease at different levels, etc...). And the noisy part of the information is generally provoked by the characteristics of the design, the methods of acquisition, the temporal dimension (repetitions during time, deterioration of the samples, etc...) and the processing. The purpose would be to measure the predominance of the signal with regard to noise, but it is not obvious in a non-supervised context. A way is to suppose that only the useful information can be "repeatable" and that the noise is independent and identically distributed, with $signal \perp\!\!\!\perp noise$. With this hypothesis, evaluating the repeatability gives an idea if $signal >> noise$ or not. In other words, repeatability may reflect the amount of metabolomic informative content (MIC) for any set of spectra, allows a direct comparison between different spectral tools and can also be seen as an evaluation of predictive ability in this paper.

According to this context and by using peak lists data sets, the first goal is to show that COSY spectra are repeatable, i.e. that COSY spectra allow to capture the main part of the information connected to the signal(s). The second goal is to demonstrate that COSY spectra are "more repeatable" than ^{1}H-NMR corresponding spectra and, by doing so, to demonstrate the utility and the importance of the additional second dimension. To reach these goals and to obtain quantitative responses, a multivariate clustering approach is selected and applied on unlabeled spectra in order to recover the signal. Two clustering algorithms (hierarchical Ward algorithm and K-Means algorithm) are used, as well as multiple combinations between different distance measures, between the use of binary positions vectors (presence or absence of a peak) and of intensity vectors and between different spectral resolutions.

This work shows that COSY spectra allow to well discriminate groups linked with different signals, proving that they are repeatable in spite of the noisy factors. It also shows that the clustering processes perform better with COSY spectra than with ^{1}H-NMR ones, thus confirming a higher MIC for the two-dimensional spectra.

The paper is organized as follows. Section 2 provides a detailed description of the two experimental designs used to illustrate the methodology and reach the goals. These experimental designs are both involving real data: the first one

implies repeated measurements of 4-mixture cell culture media containing various supervised metabolites, and the second one, a human serum based design with time sampling repetitions and multiple measurement permutations. In each case, COSY spectra and ^{1}H-NMR spectra are collected together for further comparisons. Section 3 provides a description of the different pre-processing steps: construction of a "Global Peak List", construction of binary positions vectors and intensity vectors, symmetrisation, water peak deletion, outlier deletion, etc... In this section, it is also explained how the spectral resolution is controlled (and the size of the data sets) by performing a so-called "bucketing" based on changes in the number of decimals in the global matrix. Clustering algorithms and associated distance measures are also detailed in this section. Section 4 contains the results: first, an analysis of the COSY spectra repeatability based on both positions (presences or absences of peaks) and intensities. Numerical outputs and a dendrogram illustrate these results. The comparisons between 1D and 2D spectra are also discussed in Section 4. Finally, a general conclusion and further works are given in Section 5.

2 Materials: COSY spectra, experimental designs and acquisition parameters

In this section, a short description of the second dimension in COSY spectra will motivate the main goal of this paper. Then, details about the two real experimental designs used to illustrate the methodology are provided. And finally, a technical explanation of ^{1}H-NMR and COSY spectra acquisition is proposed in Section 2.3.

2.1 The COSY second dimension

In a standard ^{1}H-^{1}H-COSY experiment (see Fig. 1), Fourier transform of FID (Free Induction Decay) gives the first dimension and Fourier transform of information with varying evolution times (τ) gives the second one. When the raw spectra are Fourier transformed in the τ dimension, the normal coherence is detected, such signals appear along the diagonal of the resulting 2D spectrum. In addition, the coherence from the frequency of the coupled partner is also detected and results in cross peaks at that position. The diagonal of the resulting 2D spectrum contains the information relative to the one-dimensional corresponding ^{1}H-NMR spectrum [10] (see Fig. 2). Wherever there is a peak on one axis, there is a peak on the diagonal.

However, if resonances are coupled, coherence transfer may lead to crosspeaks (if detectable). In ^{1}H-NMR, another challenge occurs when signals overlap: it is often difficult in that case to determine without error if one peak corresponds to one signal (one metabolite) or to several ones. In 2D, symmetric set of off-diagonal peaks correspond to coupling (see an illustration in Fig. 3). These off-diagonal correlation peaks are, by construction, specific to 2D homonuclear experiments and provide the additional information.

By visualizing 1D and 2D spectra, it can seem obvious that more dimensions mean more information, and probably more predictive power. But, this intuition can be true only if the second dimension contains relevant signal and not only

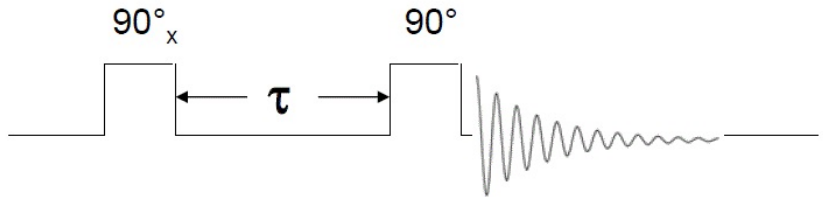

Fig. 1 Two-Dimensional Correlated Spectroscopy (2D-COSY) consists in a first pulse followed by a specific evolution time τ and a second pulse followed by the measurement period. Several variations of τ are needed before obtaining COSY experiments.

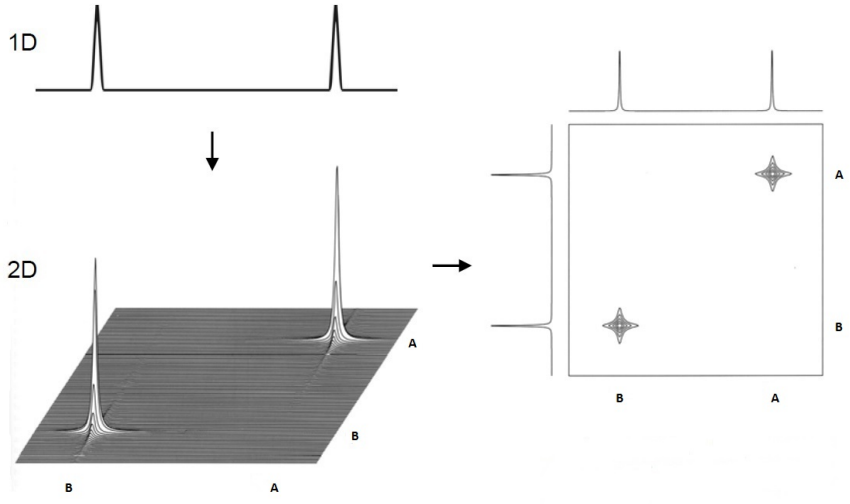

Fig. 2 Correspondance between a 1D spectrum and the diagonal of a 2D COSY spectrum. Final representation via an horizontal slice through the 2D spectrum.

noise. This is what motivates the main objective of this paper: a statistical validation confirming that the additional dimension really provides additional relevant information about the signal is needed.

2.2 Experimental designs

The methodology presented in this paper is applied on two real data sets, obtained from two designed experiments.

2.2.1 First design: cell culture media

The first design is based on four different mixtures (four cell culture media containing various levels of different metabolites like fetal bovine serum, amino acids, vitamins, proteins, etc...): DMEM/F12, MEM, RPMI/1640 et DMEM. 500μl of cell culture media were supplemented with 200μl of deuterated phosphate buffer

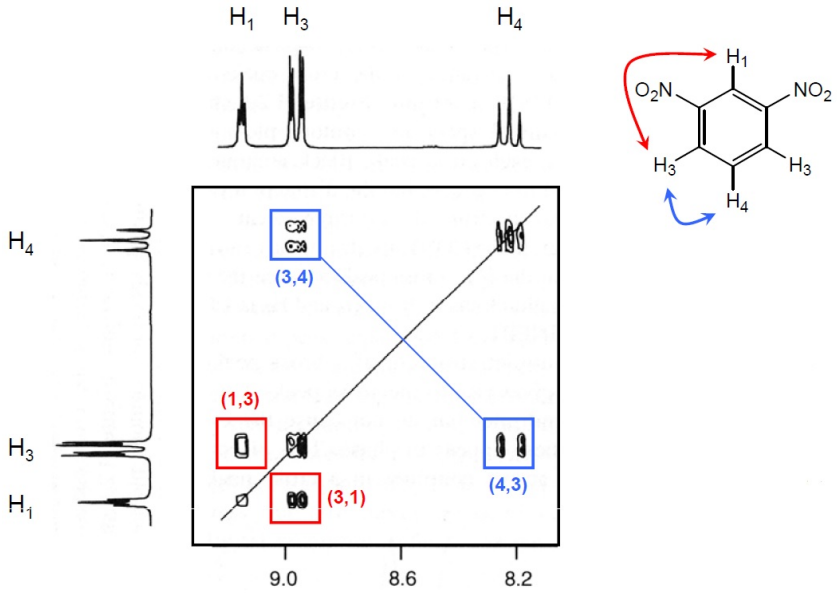

Fig. 3 Basic illustration of coupling and off-diagonal peaks in a COSY spectrum.

containing 1% of sodium azide and 10μl of TMSP (10 mg/ml). The solutions were then transferred into 5mm NMR tubes before NMR measurement. Three samples per mixture were collected, and three repeated measures were performed on each sample. All samples were subject to freezing and defrosting steps, with real risks of degradation and bacterial contamination because of the duration of the 2D analysis process.

In this design, signal corresponds to the four initial mixtures and noise arises from the sampling, time replicates, risks of degradation and other acquisition and condition parameters. 36 measures are finally available, corresponding to 36 COSY spectra and 36 corresponding peak lists. These peak lists are $(t_i \times 3)$ matrices: with t_i the number of detectable peaks (whose values are above a machine-designed threshold) in sample i ($i = 1, ..., 36$). The three columns in these matrices correspond to the coordinate, or chemical shift, on the first axis (ppm), the coordinate on the second axis (ppm) and the raw intensity of the peak. The 36 corresponding ^{1}H-NMR spectra were also collected. The total number of spectra is called n in the remainder of the paper.

2.2.2 Second design: human serum

The second experimental design is based on human serum. Four blood donors were engaged for the study. For each collected sample, 500 μl of serum were supplemented with 200μl of deuterated phosphate buffer containing 1% of sodium azide and 30μl of TMSP (10 mg/ml). The solutions were then transferred into 5mm NMR tubes before NMR measurement. The design consists in eight days of measurements with replicates within each day and multiple permutations according to a latin hypercube sampling (LHS) method [9] (in order to avoid confusion

between donors and times of analysis). For each day, the four donors samples were analyzed and provided four COSY spectra and four ^{1}H-NMR spectra. Spectral techniques (1D or 2D COSY) have not been applied at the same moment of the day, thus creating different delays before spectral measurement.

Finally, eight measures/spectra/peak lists are obtained per donor, which corresponds to 32 measures in all (32 1D spectra and 32 COSY spectra). Again, peak lists are from particular interest and correspond to $(t_i \times 3)$ matrices with t_i the number of detectable peaks in the sample, or spectra, i $(i = 1, ..., 32)$.

2.3 Spectra acquisition

1D and 2D spectra were recorded at 298K on a Bruker Avance spectrometer operating at 500.13MHz for the proton signal acquisition. The instrument was equipped with a 5mm TCI cryoprobe with a Z-gradient. Due to the nature of the samples, a presaturation sequence was used in all the experiments in order to minimize the water signal. All data were referenced to internal sodium 3-trimethylsilyl-2,2,3,3-d4-propionate (TMSP) at 0.00 ppm chemical shift. According to sample type, the ^{1}H-NMR spectra were acquired using either a 1D NOESY-presat sequence (cell culture mixture) or a CPMG relaxation-editing sequence with presaturation (human sera). Upon the presence of proteins in serum, the use of a sequence with a T2 filter (CPMG) greatly improves the baseline.

The NOESY-presat experiment used a RD-90°-T1-90°-tm-90°-sequence with a relaxation delay of 4s, a mixing time of 100ms and a fixed T1 delay of 20μs. The water suppression pulse was placed during the relaxation delay (RD). The number of transients was typically 32. The acquisition time was fixed to 3.2769001s and a quantity of four dummy scans was chosen.

The CPMG experiment used a RD-90°-(t-180°-t)$_n$-sequence with a relaxation delay (RD) of 2s, a spin echo delay (t) of 400μs and the number of loops (n) equal to 80. The water suppression pulse was placed during the relaxation delay (RD). The number of transients was typically 32. The acquisition time was fixed to 3.982555s and a quantity of four dummy scans was chosen.

The data were processed with the Bruker Topspin 2.1 software with a standard parameter set. The phase and baseline corrections were performed manually over the entire spectral range. Gradient enhanced magnitude COSY experiment with a presaturation during relaxation delay was used for 2D measurements. Spectra were collected with 4096 points in T2 and 300 points in T1 over a sweep width of 16 ppm, with six scans per T1 value. The resulting COSY spectra were processed in Topspin 2.1 using standard methods, with zero-degree shifted sine-squared apodization in both dimensions and zero filling in T1 to yield a transformed 2D dataset of 2048 by 2048 points.

Peak lists were then extracted using ACD/Labs 12.00 (ACD/NMR processor, freeware).

3 Methods: global peak list, pre-processing steps and clustering analysis

In this section, all data manipulations and pre-processing steps, both for standard metabolomics studies and for the particular purpose of this paper, are detailed. Then, a discussion on how to control the data resolution and, consequently, the size of the databases is proposed. Finally, clustering algorithms and related distance measures are also described in detail.

3.1 Global Peak List matrix

When one wants to perform simultaneous data analysis of a set of COSY spectra, a first requirement is to gather them together in a global object. The solution proposed here consists in building a so-called "Global Peak List" (GPL) matrix from the $(t_i \times 3)$ individual peak list matrices available for each 2D spectra (see the middle part of Fig. 4 for an unique individual peak list, the bottom part of Fig. 4 for the GPL). This $(T \times N)$ GPL matrix includes the T pairs of coordinates that appear in at least one of the individual spectra. The $N = 2 + 2n$ columns include first the two coordinates columns, and then, for each individual spectrum, two columns corresponding to the observed intensity measures and to a deduced binary number (1 or 0) indicating if the peak appears or not (i.e. if the corresponding intensity is stricly positive or not). For the first design, the GPL is a (3250×74) matrix $(74 = 2 + 2 * 36)$. For the second one, the GPL is a (6686×66) matrix $(66 = 2 + 2 * 32)$. The GPL matrix can also be viewed as a combination of three matrices: a $(T \times 2)$ matrix of coordinates, a $(T \times n)$ matrix of intensities I and a $(T \times n)$ matrix of positions P (presence or absence).

In this paper, both spectral intensities and peak positions are considered of particular interest. Working on the signals' positions or, in other words, on the simple existence of signals is motivated by a biological justification: a signal, or a particular metabolite, can be observed or not for a particular donor or in a particular media. If a signal is present in detectable quantity, this presence or absence is supposed to be stable, whereas intensities are variable from one measure to another, according to potential uncontrolled factors. Working on positions, or absence/presence, can then be seen as a qualitative approach; working on intensities can be seen as a more quantitative approach as it is directly linked with concentrations.

3.2 Pre-processing steps

Some classical pre-processing steps have been implemented on the Global Peak List matrices:

- **Symmetrisation.** By construction, all COSY spectra have to be symmetrical around the diagonal (see Section 2.1). Consequently, all other points or peaks are artifacts and have to be removed. It mainly concerns negative intensities and typical "crosses" which arise when choosing a wrong baseline and/or when some signals are abnormally too intense [16].

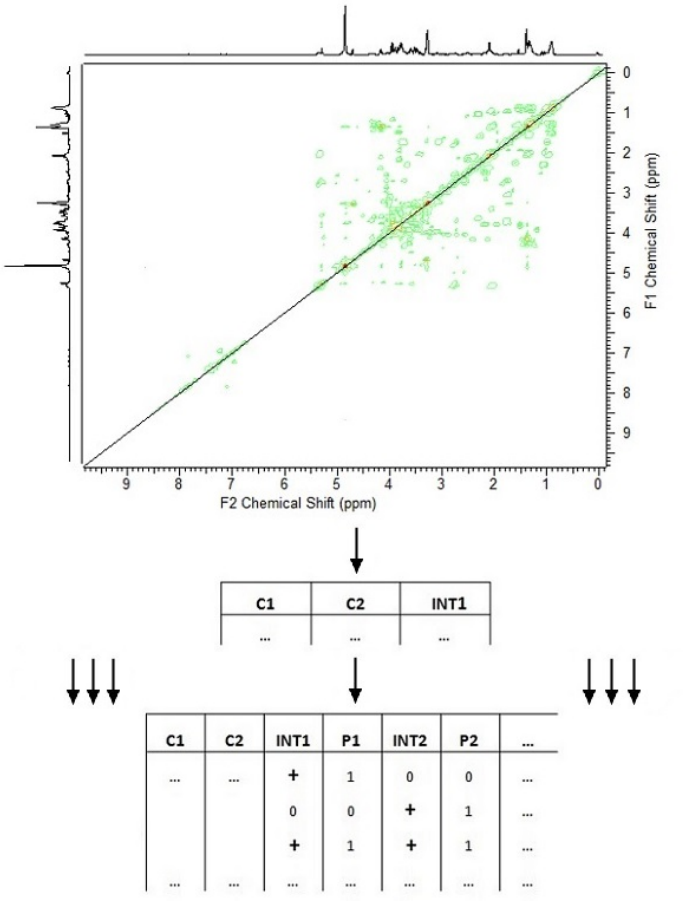

Fig. 4 From an individual COSY spectrum and peak list to the GPL matrix. See text for details about intensities (INT1, INT2, ...) and positions (P1, P2, ...) columns.

- **Water zone deletion.** Water is not of interest for metabolomics investigations. In COSY spectra, water peaks are concentrated in a square area between 4.5 and 5.5 ppm. Intensities and positions inside this area were simply removed by assigning them a zero value. More advanced methods also exist, see for example [15].
- **Normalization of the intensities.** Vectors of recoded intensities are simply obtained by applying a constant sum (= 1) normalization after water zone deletion.
- **Outlier deletion.** Some spectra can contain unexpected extreme signal values due, for example, to exceptional external factors. These outliers should be identified before data analysis and removed because they could be too influent. On both position and recoded intensities vectors, Principal Component Analysis (PCA) were applied for graphical identification of outliers, and Mahalanobis distances were calculated between raw spectra and between group centered spectra. In the context of the first experimental design, three spectral

outliers were found and ignored for upcoming analysis (because simultaneously suspected via positions and intensities). In the same way, three spectra were suspected to be outliers for the second design (see Fig. 5 for an illustration).

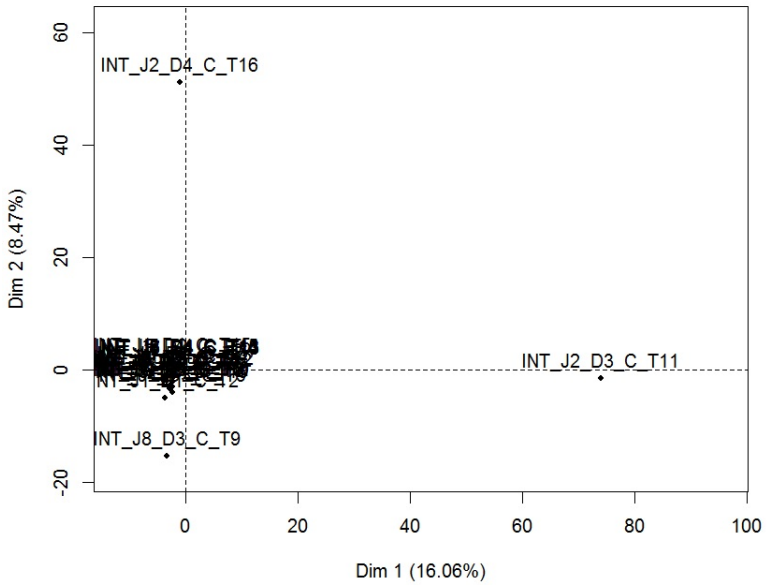

Fig. 5 Graphical identification of outliers using PCA's score plot. Here, three spectral outliers are detected based on intensities vectors: Day 2 (J2) at Time 16 (T16) for donor 4 (D4), Day 2 (J2) at Time 11 (T11) and Day 8 (J8) at Time 9 (T9) for donor 3 (D3).

3.3 Control of data resolution

In ^{1}H-NMR spectroscopy, bucketing tools are common and widely used to control the spectral resolution and/or to overcome the misalignments problem. Classical and more advanced bucketing methods have already shown their usefulness for ^{1}H-NMR spectra [22] [24]. In this paper, a bucketing step adapted to 2D-COSY is proposed in order to control the size of the database and, consequently, the resolution level of the two-dimensional spectra. Practically, a variation of the number of decimals of the coordinates is proposed. The intensities belonging to a bucket are then aggregated. For example, if the couples of coordinates [3.286; 4.194], [3.281; 4.189] and [3.278; 4.191] provide positive intensities INT1, INT2 and INT3 respectively, the couple [3.28; 4.19] provides an intensity equal to INT1+INT2+INT3 when adjusting the number of allowed decimals from 3 to 2. In this paper, three, two and one decimal cases are tested for analysis. Using this method, the width of the COSY peaks is adjusted and the resolution and size of the databases are also adjusted simultaneously. Furthermore, intermediate resolutions can of course be computed in the same way.

	One decimal	Two decimals	Three decimals
First design	(909×74)	(2348×74)	(3250×74)
Second design	(1106×66)	(4172×66)	(6686×66)

Table 1 Dimensions of the GPL matrices according to different resolutions

For positions, the aggregation process leads to another dummy variable and results from a simple function: a peak is considered in a bucket if at least one peak is present at the lower resolution level. For example, 1 and 1 lead to 1, 1 and 0 lead to 1 and 0 and 0 lead to 0 (absence everywhere). Finally, for the two experimental designs, the GPL matrices have the dimensions described in Table 1 after "bucketing" and before outlier deletion.

3.4 Clustering algorithms and distance measures

This section introduces a methodology able to quantify and compare with adequate indexes the "repeatability" (in a sense of advantageous signal capture compared to noise) of different sets of spectra. This methodology is applied in Section 4 to the two designs (each organized in four groups: mixtures or blood donors) in order to compare the repeatability of 1D and 2D COSY spectra.

For 2D experiments, several combinations between intensities and position vectors, and between the use of one, two or three decimal(s) GPL matrices is tested. Using all this information, an intuitive way to evaluate the repeatability of COSY spectra consists in non-supervised multivariate clustering (blind, with no a priori labeled information). The key idea is quite simple: if one manages to well separate and recover the four initial mixtures starting from the 36 unlabeled spectral measures of the first design, and/or if one manages to well separate and recover the four blood donors starting from the 32 unlabeled spectral measures of the second design, the goal is reached. It would mean that the signal, specific to each group of spectra, is sufficiently captured, despite the noise due to the time repetitions, the sampling methods, the freezing and defrosting periods of the samples, the risks of bacterial contamination, the potential environmental changes, etc... In other words, the objective is to verify that the within-cluster (intra) variance is minimized and that the between-cluster (inter) variance is maximized during data acquisition. Clustering is a natural and accurate tool to check this problem.

Two well-known clustering algorithms are considered: the hierarchical Ward algorithm and the K-Means algorithm (for intensities only in this second case).

- **The Ward's Algorithm** [27] [17] is a commonly used procedure for forming hierarchical groups of mutually exclusive subsets. It is particularly useful for large-scale studies (ideally with more than 100 objects) when a precise optimal solution for a specified number of groups is not practical. Given n objects, this procedure reduces them to $n-1$ mutually exclusive sets by considering the union of all possible $n(n-1)/2$ pairs and selecting the union having the minimal dissimilarity measure or aggregation index. Initially, each object is assigned to its own cluster and then the algorithm proceeds iteratively, at each stage joining the two most similar clusters, continuing until there is just a single cluster. At each stage distances between clusters are recomputed by the

Lance-Williams dissimilarity update formula [17] [18]. In this work, the *hclust* R (http://www.R-project.org) function is used, which performs hierarchical clustering and proposes a set of dissimilarity measures.

– **K-means clustering** [14] is a vector quantization method, originally from signal processing, that is popular for cluster analysis in machine learning. K-means clustering aims at partitioning the n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells. The most common algorithms use an iterative refinement and alternates between assignment steps and update steps [13]. They can only be applied on intensities here because means can not be properly defined for binary variables. The *kmeans* R function is used, allowing the joint use of the Hartigan-Wong, Lloyd and McQueen algorithms [6] [12].

Of course, both algorithms need the choice of an appropriate similarity measure, or distance, to quantify the neighborhood relation between two objects.

Clustering on binary positions vectors is first considered here. Specific similarity measures such as Ochiai, Jaccard, Dice or Russel-Rao are needed to capture the binary specificity [20]. To avoid redundancy, only two of them are used in this paper: the Jaccard and Ochiai ones. Given p binary (0=absent; 1=present) attributes, like the positions in this paper, these similarity measures between any two objects X and Y of a library are built from a general contingency table, counting the number of common attributes (see Table 2).

	$Y=1$	$Y=0$
$X=1$	a	b
$X=0$	c	d

Table 2 Contingency table for two binary attributes

On this basis, the Jaccard and Ochiai similarity measures have both a range of 0 to 1 and are defined as follows:

$$Jaccard(X,Y) = \frac{a}{a+b+c} = \frac{|X \cap Y|}{|X \cup Y|} \tag{1}$$

$$Ochiai(X,Y) = \sqrt{\frac{a}{a+b}}\sqrt{\frac{a}{a+c}} = \frac{|X \cap Y|}{\sqrt{|X|.|Y|}}. \tag{2}$$

Obviously, two spectra with many common peaks are supposed to provide high measures. The Jaccard measure can be interpreted as the size of the intersection divided by the size of the union of the sample sets; and the Ochiai measure as a geometric mean of the probabilities that if one object has the attribute, the other object has it too. Because product increases weaker than sum when only one of the terms grows, Ochiai will be really high only if both of the two proportions (probabilities) are high. It implies that the objects must share a great part of their attributes to be considered similar by Ochiai. Notice that the Ochiai coefficient can be considered as being identical to the cosine similarity index [7].

Clustering algorithms on recoded intensities are also implemented and use the classic euclidean distance as similarity measure. In Cartesian coordinates, if $X = (x_1, x_2, ..., x_p)$ and $Y = (y_1, y_2, ..., y_p)$ are two objects in $\mathbb{R}^p$, then the euclidean distance from X to Y, or from Y to X, is given by:

$$d(X,Y) = d(Y,X) = \sqrt{\sum_{i=1}^{p}(y_i - x_i)^2} = ||Y - X||_2 . \quad (3)$$

In this paper, the number of clusters is set to $K = 4$ in order to correspond to the initial number of mixtures (first design) and to the initial number of blood donors (second design). Multiple combinations between Ward/K-Means algorithms, between intensities/positions vectors, between one/two/three decimals for the resolution of the GPL matrices (and between Ochiai/Jaccard measures for the binary part of the clustering) were experimented; results are discussed in Section 4.

3.5 Numerical indexes to evaluate the quality of the clustering results

To allow an objective and numerical evaluation of the quality of the clustering results, four indexes have been used: Dunn, Davies-Bouldin, Rand and Adjusted Rand indexes. The two first evaluate the homogeneity of clusters regardless of the initial groups content; the two last measure the correctness of the non-supervised clustering process according to these initial groups.

The **Dunn index (DI)** [3] corresponds to the ratio between the smallest distance between observations not in the same cluster and the largest intra-cluster distance. Let $\mathbf{C} = (C_1, ..., C_K)$ be a particular clustering partition of n objects into K disjoint clusters. Let Δ_m be the maximum distance between observations in the cluster C_m. The Dunn index is computed as:

$$DI_{\mathbf{C}} = \min_{C_k, C_l \in \mathbf{C}, C_k \neq C_l} \left\{ \frac{\min_{i \in C_k, j \in C_l} dist(i,j)}{\max_{C_m \in \mathbf{C}} \Delta_m} \right\} . \quad (4)$$

In this work, the evaluation of inter and intra-cluster distances is based on euclidean, Ochiai or Jaccard metrics according to the nature of the data.

The **Davies-Bouldin index (DBI)** [2] is an internal evaluation scheme, where the validation of how well the clustering has been done is evaluated using quantities and features inherent to the dataset. Let C_i be a cluster. Let x_j be a p dimensional feature vector of the objects assigned to cluster C_i, A_i the medoid associated with C_i and $|C_i|$ the cardinality of C_i. Medoids are preferred compared with centroids in order to be able to work on binary positions. With these preliminary definitions, let Γ_i be the euclidean measure of variation within this cluster:

$$\Gamma_i = \sqrt{\frac{1}{|C_i|}\sum_{j=1}^{|C_i|} ||x_j - A_i||_2} .$$

The euclidean example is detailed here but note that many other distance metrics can be used. Again, in the case of binary variables, the euclidean distance is replaced by the Ochiai and Jaccard ones. Let also $\gamma(C_i, C_j)$ be a measure of separation between cluster C_i and cluster C_j, and $a_{i,l}$ the l^{th} element of A_i, $(l = 1, ..., L)$. One can write:

$$\gamma(C_i, C_j) = d(A_i, A_j) = ||A_i - A_j||_2 = \sqrt{\sum_{l=1}^{L} |a_{i,l} - a_{j,l}|^2} \ .$$

On this basis, the Davies-Bouldin index is defined as follows:

$$DBI_{\mathbf{C}} \equiv \frac{1}{K} \sum_{i=1}^{K} \max_{j:i \neq j} \left\{ \frac{\Gamma_i + \Gamma_j}{\gamma(C_i, C_j)} \right\}. \tag{5}$$

Finally, the **Rand index (RI)** and the **Adjusted Rand index (ARI)** [8] are measures of the similarity between two data clusterings, used to evaluate the quality of the classification. From a mathematical point of view, the Rand index is related to the accuracy, but is applicable even when class labels are not directly used. Given again a set of n objects $X = \{x_1, \ldots, x_n\}$ and two partitions of X to compare, $C^1 = \{C_1^1, \ldots, C_{K_1}^1\}$, a partition into K_1 subsets, and $C^2 = \{C_1^2, \ldots, C_{K_2}^2\}$, a partition into K_2 subsets, let us define the following notations:

- m_a as the number of pairs of elements in X that are in the same set in C^1 and in the same set in C^2,
- m_b as the number of pairs of elements in X that are in the same set in C^1 and in different sets in C^2,
- m_c as the number of pairs of elements in X that are in different sets in C^1 and in the same set in C^2,
- m_d as the number of pairs of elements in X that are in different sets in C^1 and in different sets in C^2.

The Rand index is defined as:

$$RI = \frac{m_a + m_d}{m_a + m_b + m_c + m_d} \tag{6}$$

Intuitively, $m_a + m_d$ can be considered as the number of agreements between C^1 and C^2, and $m_b + m_c$ as the number of disagreements. The Rand index has a value between 0 and 1, with 0 indicating that the two data clusters do not agree on any pair of points and 1 indicating that the data clusters are exactly the same.

The Adjusted Rand index (ARI) is the corrected-for-chance version of the Rand index. Though the Rand Index may only yield a value between 0 and 1, the Adjusted Rand Index can yield negative values if the index is less than the expected index. Following the same notations m_a, m_b, m_c and m_d, it is defined as [23]:

$$ARI = \frac{\binom{n}{2}(m_a + m_d) - [(m_a + m_b)(m_a + m_c) + (m_b + m_d)(m_c + m_d)]}{\binom{n}{2}^2 - [(m_a + m_b)(m_a + m_c) + (m_b + m_d)(m_c + m_d)]} \tag{7}$$

Note that DI, RI and ARI have to be as high as possible to represent good clustering partitions. DBI has to be as small as possible.

4 Results and discussion

In this section, clustering methods and related quality indexes are used to assess the repeatability of different versions of COSY spectra. The first upcoming subsection proposes numerical outputs based on the indexes defined in Section 3.5 for both designs and for combinations of experiments described in Section 3.3 and Section 3.4. Then, in Section 4.2, comparisons between clustering results based on 2D COSY spectra and on corresponding ^{1}H-NMR spectra are shown. This section will allow to demonstrate on the two designs that, indeed, COSY spectra include additive relevant information for spectral classification as compared to ^{1}H-NMR ones (providing consequently a better MIC). By doing this, a statistically-proved response to the initial key question (does supplementary information include relevant and crucial information?) is provided.

4.1 Repeatability of COSY spectra

The numerical results are available in Table 3 for the first experimental design, and in Table 4 for the second one. Note that only a part of all the considered combinations are shown to avoid too much redundancy and to improve clarity.

Signal	Dec.	Algorithm	Distance	T	DI	DBI	RI	ARI
Positions	1	Ward	Jaccard	909	0.712	1.466	**0.914**	**0.811**
Positions	1	Ward	Ochiai	909	0.712	1.466	**0.914**	**0.811**
Positions	2	Ward	Jaccard	2348	**0.857**	1.688	0.757	0.420
Positions	2	Ward	Ochiai	2348	**0.827**	1.688	0.757	0.420
Positions	3	Ward	Jaccard	3250	**0.893**	1.722	0.601	0.189
Positions	3	Ward	Ochiai	3250	**0.892**	1.722	0.601	0.189
Intensities	1	Ward	Euclidean	909	0.298	**0.541**	**0.927**	**0.853**
Intensities	1	K-means	Euclidean	909	0.298	**0.541**	**0.927**	**0.853**
Intensities	2	Ward	Euclidean	2348	0.239	1.249	0.669	0.609
Intensities	2	K-means	Euclidean	2348	0.311	1.356	0.657	0.501
Intensities	3	Ward	Euclidean	3250	0.439	1.540	0.588	0.425
Intensities	3	K-means	Euclidean	3250	0.434	1.614	0.597	0.439

Table 3 First design: numerical clustering performances according to different versions of COSY spectra. The column "Dec." denotes the number of decimal(s) for the data in the GPL matrix.

Globally, the results are very promising with many RI (and ARI) greater than 0.7, particularly with bucketed data when one or two decimals in the GPL matrices are used. Unlabeled individuals are well grouped together, with no prior knowledge, and this according to the presence of numerous potential noise factors (sampling, time replicates, degradation, changes in temperature, etc...).

Considering first the position-based partitions, particular groups are already well isolated in the majority of cases (for instance, the third cell culture mixture

Signal	Dec.	Algorithm	Distance	T	DI	DBI	RI	ARI
Positions	1	Ward	Jaccard	1106	0.796	1.569	**0.937**	**0.825**
Positions	1	Ward	Ochiai	1106	0.722	1.569	**0.937**	**0.825**
Positions	2	Ward	Jaccard	4172	**0.945**	1.800	0.772	0.401
Positions	2	Ward	Ochiai	4172	**0.899**	1.800	0.772	0.401
Positions	3	Ward	Jaccard	6686	**0.981**	1.792	0.650	0.171
Positions	3	Ward	Ochiai	6686	**0.963**	1.792	0.650	0.171
Intensities	1	Ward	Euclidean	1106	0.419	**0.643**	**0.932**	**0.804**
Intensities	1	K-means	Euclidean	1106	0.419	**0.643**	**0.932**	**0.804**
Intensities	2	Ward	Euclidean	4172	0.689	1.592	0.789	0.422
Intensities	2	K-means	Euclidean	4172	0.706	1.742	0.735	0.288
Intensities	3	Ward	Euclidean	6686	0.778	1.668	0.578	0.055
Intensities	3	K-means	Euclidean	6686	0.765	1.886	0.647	0.135

Table 4 Second design: numerical clustering performances according to different versions of COSY spectra. The column "Dec." denotes the number of decimal(s) for the data in the GPL matrix.

in the first design was known to be significantly different from the others). An example is given in Fig. 6.

With the recoded intensities-based partitions, the four mixtures are generally well recovered by the algorithms in spite of the sampling procedure and time repetitions. For the first design, the best clustering result is obtained with the one-decimal GPL matrix, thus underlying the importance of the "bucketing" step: Fig. 7 shows that there is only one error during the blind clustering process and RI and ARI indexes are maximized (RI=0.927 and ARI=0.853 in Table 3). The conclusion is the same for the second design concerning the blood donors (RI=0.932 and ARI=0.804 in Table 4).

More generally, Dunn and Davies-Bouldin indexes are subject to significative variations between experiments and are not very satisfactory for some combinations (several DI less than 0.5 and DBI greater than 1.5 in Table 3 and Table 4). This means that obtained clusters are not always compact and well separated. However, to judge if the clusters conform to the reality, i.e. if members in a cluster correspond to members of initial groups, Rand and Adjusted Rand indexes prevail because they are built on the degree of agreement and disagreement between groups (here between the reality and each of the clustering partitions).

In conclusion, and based on the two experimental designs, the clustering results show that COSY spectra appear to be statistically repeatable and contain informative signal which helps to succeed in distinguishing groups of different spectra. In other words, COSY spectra are enough repeatable so that the signal connected to the initial groups is distinguished from the noise. Working on positions gives satisfactory results but not better ones than working on intensities. Moreover, these results show that the 2D "bucketing" step is important and probably necessary to solve misalignment problems (as it is the case in 1D). This additional information can be profitable for further robust statistical analysis, as biomarker discovery.

4.2 Comparisons with corresponding ^{1}H-NMR spectra

Besides the convincing repeatability of COSY spectra, this section intends to demonstrate that the additional information contained in 2D spectra (compared

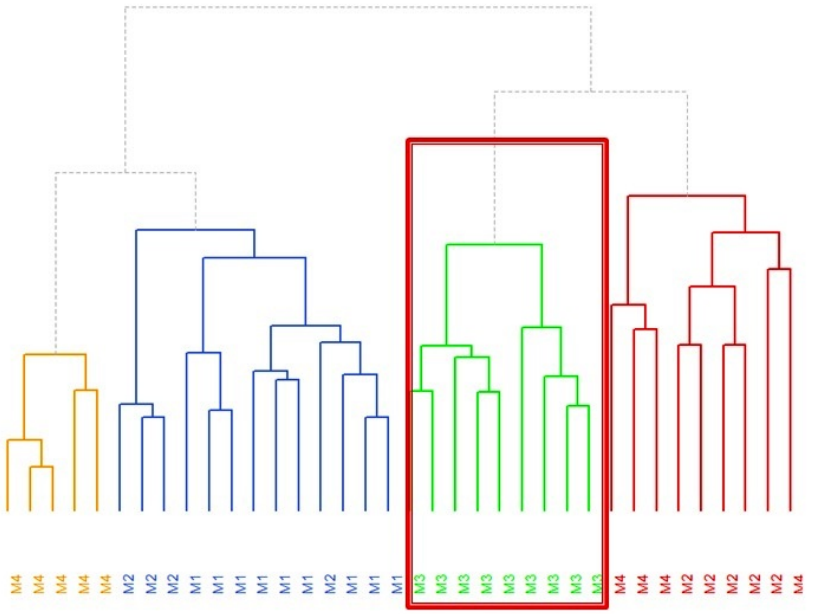

Fig. 6 Illustrative dendrogram of the clustering on positions (first design, Ward algorithm with Ochiai distance, dec = 2, before outlier deletion). M1 to M4 denote the four different initial mixtures. Different colours are attributed to different clusters.

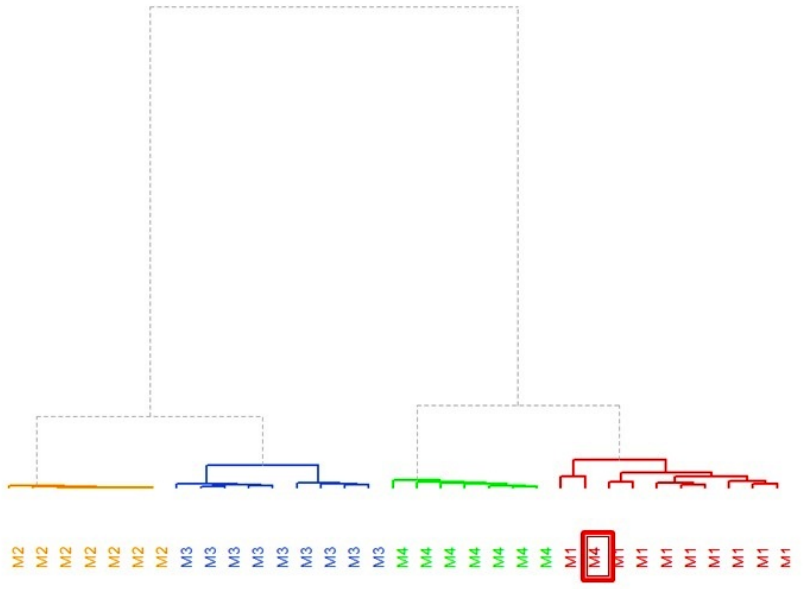

Fig. 7 Illustrative dendrogram of the clustering on intensities (first design, Ward algorithm, dec = 1, after outlier deletion). M1 to M4 denote the four different initial mixtures. Different colours are attributed to different clusters.

to 1D) is relevant and crucial by improving the quality of the clustering results. It is always difficult to directly compare objects of different dimensions. To be able to compare ^{1}H-NMR and COSY spectra, some pre-processing steps are necessary to upgrade the one-dimensional spectra before performing clustering.

4.2.1 Upgrade of ^{1}H-NMR spectra

Some very powerful and complete tools are available to pre-process, transform and interpret ^{1}H-NMR data, for example in [29] [25]. Here, the goal is not to focus on sophisticated pre-processing methods for ^{1}H-NMR data, but just to upgrade them in order to obtain comparable pre-processing levels for both 1D and COSY data. To do so, one-dimensional spectra were subject to the following steps after phase and baseline correction:

- elimination of negative intensities,
- deletion of the water peaks between 4.5 and 5.5 ppm,
- application of the constant sum transformation,
- "bucketing" by using the same number of decimals (for the ppm coordinates) than for the COSY experiments,
- outlier deletion.

In a general way, it is quite intuitive that 2D spectra require less pre-processing to obtain an acceptable vizualisation of the signal and of potential biomarkers. Finally, note that the reasoning based on the positions makes no sense when working with 1D spectra. Consequently, clustering processes were only performed on 1D normalized intensities (raw intensities, "bucketed" with two decimals and with only one decimal).

4.2.2 Comparisons

The clustering results obtained with ^{1}H-NMR spectra are available in Table 5 for the first design and in Table 6 for the more complex second one.

Signal	Dec.	Algorithm	Distance	T	DI	DBI	RI	ARI
Intensities	1	Ward	Euclidean	201	**0.927**	**0.120**	**0.908**	**0.807**
Intensities	1	K-means	Euclidean	201	0.560	**0.463**	**0.835**	0.667
Intensities	2	Ward	Euclidean	2001	0.712	**0.426**	0.717	0.516
Intensities	2	K-means	Euclidean	2001	0.601	0.739	0.732	0.476
Raw intensities		Ward	Euclidean	199967	0.362	1.048	0.606	0.544
Raw intensities		K-means	Euclidean	199967	0.284	1.182	0.410	0.196

Table 5 First design: numerical clustering performances for 1D data.

If one compares these results with Rand and Adjusted Rand indexes in Table 3 and Table 4 respectively, it appears clearly that these indexes are mainly higher when the clustering algorithms are performed on COSY spectra. Resulting partitions of the spectra are better for COSY: they are more in agreement with the real initial groups. Thus, this proves, for the two experiments, that the additional spectral dimension does provide relevant information to improve the clustering results and, consequently, to improve the repeatability and the spectral informative content (MIC).

Signal	Dec.	Algorithm	Distance	T	DI	DBI	RI	ARI
Intensities	1	Ward	Euclidean	207	**0.981**	**0.688**	0.704	0.233
Intensities	1	K-means	Euclidean	207	0.334	**0.663**	0.702	0.230
Intensities	2	Ward	Euclidean	2054	**0.901**	0.986	0.704	0.233
Intensities	2	K-means	Euclidean	2054	0.635	1.119	**0.740**	**0.290**
Raw intensities		Ward	Euclidean	205034	0.670	1.074	0.588	0.017
Raw intensities		K-means	Euclidean	205034	0.447	1.321	0.675	0.145

Table 6 Second design: numerical clustering performances for 1D data.

5 Conclusion and further works

In this article, an advanced clustering approach is proposed to evaluate and quantify the "repeatability" (as defined in the introduction) of metabolomic NMR spectral data, and is applied on both ^{1}H-NMR and COSY spectra, using both qualitative input (binary positions, linked with the absence or presence of a peak or metabolite) or quantitative inputs (intensities). More precisely, the choice of this clustering approach is to be related to the presence of initial groups in the signal. In the two real experimental designs detailed in Section 2.2, four different cell culture mixtures and four different blood donors were respectively involved. The elements of these groups were "mixed" via several noisy factors: sampling, time replicates, delays, deterioration, risk of bacterial contamination, etc... The goal of the clustering processes was to blindly recover these initial groups using all the individual unlabeled spectra. And the final quality of these processes informs us about the quantity of captured signal and can be finally viewed as a measure of repeatability.

This work first shows some pre-processing steps for the data analyst in order to handle 2D COSY data, including the construction of the Global Peak List (GPL) matrix and a bucketing step which allows to control the data resolution.

It then demonstrates that COSY spectra provide very satisfying clustering results: in other words, in spite of the multiple noise factors, the informative part of the signal can be well discovered and, finally, the final clusters are mainly in concordance with the initial groups. The clustering methodology also highlights that 2D "bucketing", by aggregating data according to reduced numbers of decimals for the ppm coordinates, helps to improve repeatability when using COSY data.

Furthermore, this paper also provides a comparison between COSY and ^{1}H-NMR spectra, based on the quality of respective clustering results. It is demonstrated that COSY spectra provide more metabolomic informative content (MIC) about the groups than corresponding upgraded ^{1}H-NMR spectra, thus proving the importance and the relevance of the additional dimension of COSY to go further in NMR-based metabolomics studies.

These promising results have to be confirmed on faster versions of COSY experiments, in order to deal with more competitive acquisition times. The methodology can also be applied to heteronuclear two-dimensional spectroscopy tools, like HSQC spectra. Ultimately, once the repeatability of a given 2D tool or technology is verified, the research of discriminating zones or biomarkers, which represents the main objective of most metabolomics studies, can benefit from the advantageous use of 2D-NMR spectra instead of the more traditional ^{1}H-NMR ones.

Acknowledgements The authors are grateful to the Laboratoire de Pharmacognosie et de Chimie Pharmaceutique, ULg, for providing data. Support from the IAP Research Network P7/06 of the Belgian State (Belgian Science Policy) is also gratefully acknowledged.

References

1. Akitt J. W., Mann B. E., NMR and Chemistry. Cheltenham, UK: Stanley Thornes. p. 287 (2000).
2. Davies D.L., Bouldin D.W., A cluster separation measure, IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-1 (2), pp. 224-227 (1979).
3. Dunn J.C., A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters, Journal of Cybernetics, 3 (3), pp. 32-57 (1973).
4. Giraudeau P., Remaud G., Akoka S., Evaluation of Ultrafast 2D NMR for Quantitative Analysis, Anal. Chem., 81 (1), pp. 479-484 (2009).
5. Haitao L., Shuchun L., Liangdong G., Yonggang Z. and al., New Furanones from the Plant Endophytic Fungus Pestalotiopsis besseyi, Molecules 2012, 17(12), pp. 14015-1402 (2012).
6. Hartigan J. A., Wong M. A., A K-means clustering algorithm, Applied Statistics 28, pp. 100-108 (1979).
7. Holliday J.D., Hu C.Y., Willett P., Grouping of coefficients for the calculation of Inter-Molecular Similarity and Dissimilarity using 2D fragment bit-strings, Combinatorial Chemistry and High Throughput Screening, 5, Number 2, pp. 155-166 (2002).
8. Hubert L., Arabie P., Comparing partitions, Journal of Classification 2 (1), pp.193-218 (1985).
9. Iman R. L., Latin hypercube sampling, John Wiley and Sons, Ltd (2008).
10. Jacobsen N. E., NMR spectroscopy explained: simplified theory, applications and examples for organic chemistry and structural biology. John Wiley and Sons, Ltd (2007).
11. Keeler J., Understanding NMR Spectroscopy (2nd ed.), John Wiley and Sons, pp. 190-191 (2010).
12. Lloyd S. P., Least squares quantization in PCM, Technical Note, Bell Laboratories, IEEE Transactions on Information Theory 28, pp. 128-137 (1957, 1982).
13. MacKay D., An Example Inference Task: Clustering, Information Theory, Inference and Learning Algorithms, Cambridge University Press, pp. 284-292 (2003).
14. MacQueen J. B., Some Methods for classification and Analysis of Multivariate Observations, Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, 1, University of California Press, pp. 281-297 (1967).
15. Mao X., Ye C., Phase-shift presaturation for water peak suppression in biomolecular NMR experiments, Science in China. Series C, Life sciences, 40 (4), pp. 345-350 (1997).
16. Marion D., Bax A., Baseline distortion in real-fourier-transform NMR spectra, Journal of Magnetic Resonance (1969), 79(2), pp. 352-356 (1988).
17. Murtagh F., Legendre P., Ward's hierarchical clustering method: clustering criterion and agglomerative algorithm, arXiv preprint arXiv:1111.6285 (2011).
18. Murtagh F., Contreras P., Algorithms for hierarchical clustering: an overview, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2(1), pp. 86-97 (2012).
19. Nicholson J., Connelly J., Lindon J.C., Holmes E., Metabonomics: a generic platform for the study of drug toxicity and gene function, Nature Reviews Drug Discovery, 1, pp. 153-161 (2002).
20. Plasse M., Niang N., Saporta G., Villeminot A., Leblond L., Combined use of association rules mining and clustering methods to find relevant links between binary rare attributes in a large data set, Computational Statistics and Data Analysis, 52(1), 596-613 (2007).
21. Queiroz Junior L. H. K.; Ferreira A. G., Giraudeau P., Optimization and practical implementation of ultrafast 2D NMR experiments. Qum. Nova [online], vol.36, n.4, pp. 577-581 (2013).
22. Rousseau R., Statistical contribution to the analysis of metabonomic data in ^{1}H-NMR spectroscopy, PhD Thesis, UCL, http://hdl.handle.net/2078.1/75532 (2011).
23. Santos J. M., Embrechts M., On the use of the adjusted rand index as a metric for evaluating supervised classification, Artificial Neural Networks, ICANN 2009, Springer Berlin Heidelberg, pp. 175-184 (2009).
24. Sousa S.A., Magalhaes A., Castro Ferreira M.M., Optimized bucketing for NMR spectra: Three case studies, Chemometrics and Intelligent Laboratory Systems, 122, pp. 93-102 (2013).

25. Vanwinsberghe J., Bubble: development of a matlab tool for automated ^{1}H-NMR data processing in metabonomics, Master's thesis, Université de Strasbourg (2005).
26. Vega-Vazquez M., Cobas J.C., Martin-Pastor M., Fast multidimensional localized parallel NMR spectroscopy for the analysis of samples, Magn Reson Chem, 48(10): pp. 749-52 (2010).
27. Ward J.H., Hierarchical Grouping to optimize an objective function, Journal of American Statistical Association, 58(301), pp.236-244 (1963).
28. Xi Y., deRopp J.S., Viant M., Woodruff D., Yu P., Automated screening for metabolites in complex mixtures using 2D COSY NMR spectroscopy, Metabolomics, Vol. 2, No. 4, pp. 221-233 (2007).
29. Xia J., Wishart D., MetPA: a web-based metabolomics tool for pathway analysis and visualization, Bioinformatics, 26(18), pp.2342-2344 (2010).
30. Yun K., Sunghyouk P., Jongheon S., Dong-Chan O., Application of ^{1}3C-labeling and ^{1}3C-^{1}3C COSY NMR experiments in the structure determination of a microbial natural product, Archive of Pharmacal Research, DOI 10.1007/s12272-013-0254-8 (2013).