*Published in : Metabolomics (2020), vol. 16, article no. 42, pp. 1-11*
*doi.org/10.1007/s11306-020-01662-6*
*Status : Postprint (Author's version)*

# COMBINING RAPID 2D NMR EXPERIMENTS WITH NOVEL PRE-PROCESSING WORKFLOWS AND MIC QUALITY MEASURES FOR METABOLOMICS

Baptiste Féraud[1,2], Estelle Martineau[3,4], Justine Leenders[5], Bernadette Govaerts[1], Pascal de Tullio[5], Patrick Giraudeau[3]

[1]*Institute of Statistics, Biostatistics and Actuarial Sciences (ISBA), Université Catholique de Louvain (UCL), Voie du Roman Pays 20, bte L1.04.01, 1348 Louvain-la-Neuve, Belgium*
[2]*Machine Learning Group, Université Catholique de Louvain (UCL), 1348 Louvain-la-Neuve, Belgium*
[3]*Université de Nantes, CNRS, CEISAM UMR 6230, 44000 Nantes, France*
[4]*Spectromaîtrise, CAPACITES SAS, 26 Bd Vincent Gâche, 44200 Nantes, France*
[5]*Metabolomics Group, Center for Interdisciplinary Research on Medicines (CIRM), Université de Liège (ULG), Liège, Belgium*
*Corresponding author: Baptiste Féraud, baptiste.feraud@uclouvain.be*

**Abstract**

**Introduction** The use of 2D NMR data sources (COSY in this paper) allows to reach general metabolomics results which are at least as good as the results obtained with 1D NMR data, and this with a less advanced and less complex level of preprocessing. But a major issue still exists and can largely slow down a generalized use of 2D data sources in metabolomics: the experiment duration.

**Objective** The goal of this paper is to overcome the experiment duration issue in our recently published MIC strategy by considering faster 2D COSY acquisition techniques: a conventional COSY with a reduced number of transients and the use of the Non-Uniform Sampling (NUS) method. These faster alternatives are all submitted to novel 2D pre-processing workflows and to Metabolomic Informative Content analyses. Eventually, results are compared to those obtained with conventional COSY spectra.

**Methods** To pre-process the 2D data sources, the Global Peak List (GPL) workflow and the Vectorization workflow are used. To compare this data sources and to detect the more informative one(s), MIC (Metabolomic Informative Content) indexes are used, based on clustering and inertia measures of quality.

**Results** Results are discussed according to a multi-factor experimental design (which is unsupervised and based on human urine samples). Descriptive PCA results and MIC indexes are shown, leading to the direct and objective comparison of the different data sets.

**Conclusion** In conclusion, it is demonstrated that conventional COSY spectra recorded with only one transient per increment and COSY spectra recorded with 50% of non-uniform sampling provide very similar MIC results as the initial COSY recorded with four transients, but in a much shorter time. Consequently, using techniques like the reduction of the number of transients or NUS can really open the door to a potential high-throughput use of 2D COSY spectra in metabolomics.

**Keywords**

2D NMR, COSY spectra, 2D pre-processing workflows, Global peak list, Vectorization, Metabolomic informative content (MIC), Non-uniform sampling (NUS)

*Published in : Metabolomics (2020), vol. 16, article no. 42, pp. 1-11*
*doi.org/10.1007/s11306-020-01662-6*
*Status : Postprint (Author's version)*

# 1. Introduction

This paper evaluates, in the context of metabolomics, the potential of combining rapid 2D NMR acquisition methods with the concept of Metabolomics Informative Content (MIC) that we recently developed.

It has been proved in Feraud et al. (2015, 2019 that the use of 2D COSY spectral data allows reaching a higher level of Metabolomic Informative Content (MIC) compared to the traditional use of 1D $^{1}$H-NMR spectra, thus taking advantage of the additional information contained in the 2D cross non-diagonal peaks. These gains in terms of MIC are then confirmed by subsequent better performances when searching for relevant final biomarkers to interpret (see again Feraud et al. (2015, 2019).

In terms of pre-processing, a main initial intuition has been verified in these previous studies: the use the 2D COSY data sources allows to reach general metabolomics results which are at least as good as the results obtained with 1D data, and this with a less advanced and less complex level of pre-processing.

Among all these assets, a major problem still exists and can largely slow down a generalized use of 2D spectra (COSY or others) in metabolomics: the experiment duration. Typically, the experiment duration of a single COSY spectrum on a biofluid sample such as urine requires from 20 min to an hour (Delikatny et al. 1991) depending on the spectrometer and on the targeted sensitivity and resolution. This drawback can lead to an overload of spectrometer schedules in the case of large sample collections, making conventional 2D NMR not compatible with high-throughput metabolomics.

Numerous approaches have been described in the NMR literature to speed up the duration of 2D NMR experiments (Rouger et al. 2017) and several of them have already been applied to untargeted metabolomics (Marchand et al. 2017) and lipidomics (Marchand et al. 2018). However, these methods often come at a price to pay in terms of resolution and/or sensitivity, and a central question is to determine if they can yield the same level of information as conventional 2D NMR for metabolomics. In this work, the potential of two accelerated 2D NMR approaches (both in the case of COSY) is evaluated together with our recently described concept of Metabolomics Informative Content (MIC). The first one, named NS1, is a simple reduction of the number of transients compared to our previous study (from four to one transient per increment). This requires a careful selection of relevant NMR coherences by the use of magnetic field gradient pulses to avoid phase cycling. The second one, named NUS50, is the use of Non-Uniform Sampling (Barna and Laue 1987, Hoch et al. 2014) at a 50% level, also with one transient per increment.

Practically, NS1 and NUS data sources were obtained and tested in the context of a previously published interlaboratory study Feraud et al. (2019) involving three urine donors. The main interest here is to confront these faster COSY alternatives with ad-hoc 2D pre-processing workflows and with innovative and objective quality measures. By this way, all the 2D data were then pre-processed through the novel Global Peak List and Vectorization 2D workflows, and Metabolomic Informative Content (MIC) measures were calculated on them (Feraud et al. 2015, 2019).

If the faster NUS and NS1 acquisition techniques allow reaching similar results as COSY in Feraud et al. (2019), the experiment duration issue would be solved concerning the use of 2D NMR data sources, thus opening the door to their potential more massive and high throughput use in metabolomics studies. Thereby, in the case of conclusive results and through shorter experimental durations, analyses of larger cohorts of samples may be considered and facilitated.

*Published in : Metabolomics (2020), vol. 16, article no. 42, pp. 1-11*
*doi.org/10.1007/s11306-020-01662-6*
*Status : Postprint (Author's version)*

The paper first provides a detailed description of the considered samples and of the different 2D data sources involved in the study. The NMR methods and parameters are then described in details (Sect. 2).

Section 3 provides reminders on the GPL 2D pre-processing workflow (Sect. 3.1), on the Vectorization workflow (Sect. 3.2) and on the MIC concept and quality measures (Sect. 3.3).

Results and comparisons are detailed and discussed in Sect. 4, including descriptive PCA results (Sect. 4.1) and advanced MIC results (Sect. 4.2). These results allow to measure the impact of the use of faster acquisition techniques with respect to the use of conventional COSY spectral sources.

# 2. Materials

## 2.1 Samples: urine donors

### 2.1.1 Description and motivations

This data set was built in the context of a wider inter-laboratory study about the repeatability of different 1D $^{1}$H-NMR and 2D COSY spectral measures and acquisition protocols (Feraud et al. 2019). The experience involves three factors of variation: three different donors, two urine dilution levels and four different days of acquisition. Eight measures are finally available for each donor.

In this paper, note that the focus is strictly on the group factor (i.e. the donors) as it corresponds to the signal we want to capture and explain in subsequent sections. In this regard, urine dilution and days factors can be considered as additional sources of noise and will not be commented separately in details.

Concretely, the idea is to handle all these data sources with the ad-hoc 2D workflows (presented in Sect. 3) and to visualize which of them succeeds best in capturing the signal or main information (i.e. the donors).

### 2.1.2 Urine collection

In order to conduct this experiment and to design the collection of urine samples, the morning urine of three different fasting donors was collected. For each donor, four aliquots of 400 µl and four aliquots of 320 µl were placed at − 80°C. Then, on each consecutive day (four days), six aliquots were thawed (3 donors × 2 quantities) and routinely prepared as follows.

Urine samples of 400 µl were supplemented with 300 µl of deuterated phosphate buffer (DPB, pH 7.4), while 320 µl urine aliquots were supplemented with 380 µl of the same buffer. 10 µl of a 10 mg/ml TMSP solution was then added to all aliquots. The four aliquots of each dilution were put in 5mm NMR tubes for NMR acquisition and analyzed. For each day, the order of measurement was held constant across the six sub-samples. A total of 24 1D and 2D collected signals are finally available.

All spectra are internally labelled as $S_i\_Dj\_E_k$, where $S$ corresponds to the donor label ($i$ = 1,...,3), $D$ is the dilution ( $j$ =0 : no dilution; $j$ =1: 25% diluted) and $E$ is the day of acquisition ($k$= 1,...,4).

*Published in : Metabolomics (2020), vol. 16, article no. 42, pp. 1-11*
*doi.org/10.1007/s11306-020-01662-6*
*Status : Postprint (Author's version)*

## 2.2 NMR acquisition and processing parameters

For the particular acquisitions of interest, all the samples were analyzed on a Bruker Avance-III HD spectrometer operating at a $^{1}H$ frequency of 700.13 MHz, equipped with an inverse $^{1}H/^{13}C/^{15}N/^{2}H$ cryogenically cooled probe and a Z gradient coil. Analyses were performed at 298 K.

COSY experiments were recorded on each sample with 300 increments in the indirect dimension. The pulse sequence includes a water suppression scheme applied during the recovery delay *D*1 and the use of a gradient-based coherence selection which was sufficient to avoid phase-cycling.

First, samples were analyzed under the whole conventional experimental conditions developed in Feraud et al. (2019) (called here COSY NS4, because it involves four transients). Then, in order to significantly reduce the duration of the experiments, the acquisitions were performed with only a single transient (*NS* = 1, called here COSY NS1) instead of four while the other parameters were kept identically between COSY NS1 and COSY NS4. The sensitivity was obviously decreased, but the signal-to-noise ratio remained sufficient to allow subsequent studies.

Finally, Non-Uniform Sampling was applied. In particular, some acquisition parameters were modified or added compared to COSY NS1:

- The percentage of NUS was set to 50% (the resulting data set being called COSY NUS50). This value was previously tested on experimental data in order to obtain a sufficient signal-to-noise ratio with only one transient (*NS* = 1) without losing too much resolution. Considering a fixed number of points $t_1$ (for example equal to 300), experiments without NUS, with 50% of NUS and with 25% of NUS were thereby tested. Spectra and preliminary results based on signal-to-noise ratios are available in Fig. 1. The noisy trails in the first dimension are slightly more important in the 25% NUS spectrum (right of Fig. 1), thus implying lower signal-to-noise ratios for some signals compared to the other experiments. With 50% of NUS, the S/N ratio decreases for signals 3 and 4 with respect to the experiment without NUS and evolves little for the two other signals. This loss in sensibility remains acceptable considering that the experimental time was divided by two.
- Acquisition parameters were slightly different from those applied to the conventional COSY experiments (recovery delay = 3s and 256 $t_1$ increments), but this did not significantly affect the sensitivity and resolution. Note that only 128 increments are finally retained with the 50% NUS. These points were selected by generating a Poisson-gap scheme (Hyberts et al. 2010, Maciejewski et al. 2011) which allows to limit artefacts and to avoid too large intervals between successive $t_1$ values. The distribution of the 128 actual data points is provided as supplementary information.

The overall experiment time was 50 min for conventional experiments with NS4, 13 minutes for NS1 experiments and 8 minutes for NS1 experiments with 50% NUS.

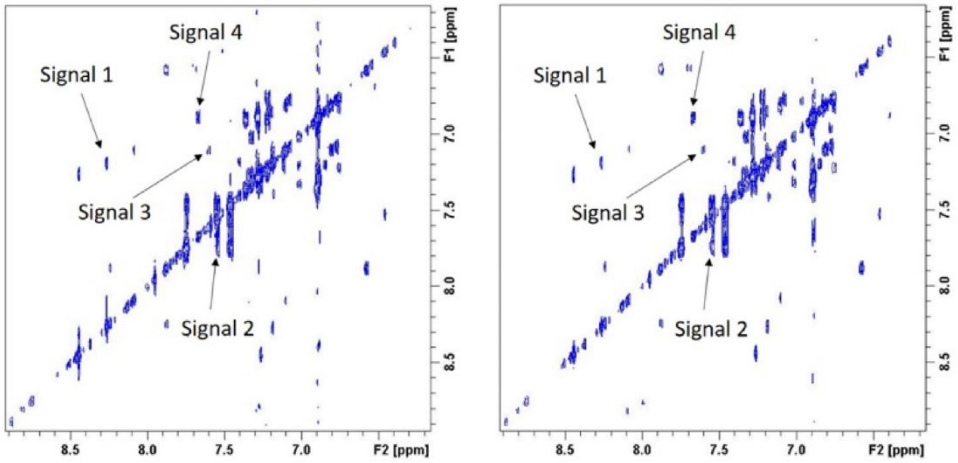
For all the spectra, a multiplication by a cosine apodization function was performed (WDW = SINE and SSB = 0) before Fourier transform in the two dimensions. Baseline correction was performed in the two dimensions (ASBG = 0) and spectra were automatically calibrated according to $\delta$ = 0 ppm.

For NUS spectra, the obtained two-dimensional FID cannot be directly processed as for a conventional 2D NMR experiment after data acquisition. Indeed, the final 2D spectrum is reconstructed through the use of a non-linear data processing method. This spectral reconstruction was performed via the "Compressed

*Published in : Metabolomics (2020), vol. 16, article no. 42, pp. 1-11*
*doi.org/10.1007/s11306-020-01662-6*
*Status : Postprint (Author's version)*

Sensing" (CS, Kazimierczuk and Orekhov 2011) algorithm which has been developed specifically for 2D NMR experiments.

All the spectra were pre-processed according to the data pipeline fully described in Feraud et al. (2019) and summarised in Sect. 3.




| | without NUS (texp = 13 min) | | 50% de NUS (texp = 8 min 30 s) | | 25% de NUS (texp = 4 min 43 s) | |
|---|---|---|---|---|---|---|
| | S/N (col) | S/N (row) | S/N (col) | S/N (row) | S/N (col) | S/N (row) |
| Signal 1 | 59 | 65 | 45 | 77 | 57 | 70 |
| Signal 2 | 35 | 62 | 31 | 47 | 20 | 34 |
| Signal 3 | 30 | 25 | 21 | 17 | 6 | 5 |
| Signal 4 | 125 | 90 | 72 | 79 | 57 | 61 |

**Fig. 1** Preliminary analyses of NUS amounts (with NS = 1): experimental COSY spectra acquired with 50% NUS (left) and 25% NUS (right) and corresponding measured signal-to-noise ratios

# 3. Methods: application of the 2D workflows (GPL and Vectorization) and of the MIC on the faster COSY data sources

In order to build global objects which can contain all the information from the 24 individual spectra, the Global Peak List (GPL) and the Vectorization 2D workflows are performed on the pools of COSY NS4, COSY NS1 and COSY NUS50 spectra.

## 3.1 The 2D GPL pre-processing workflow

The GPL pre-processing workflow is fully detailed in Feraud et al. (2015). First, each of the $n$ initial individual 2D spectra involved in an experiment is converted into an individual peak list, i.e. a ($t_i$ × 3) matrix which contains the two ppm coordinates and the concentration intensity of $t_i$ existing peaks. After some manipulations, the $n$ peak lists are merged, resulting in a ($T \times M_{GPL}$) Global Peak List (GPL) matrix. This matrix includes the $T$ pairs of coordinates that appear in at least one of the individual spectra and all the corresponding intensities. Note that the number of rows $T$ is not known in advance.

*Published in : Metabolomics (2020), vol. 16, article no. 42, pp. 1-11*
*doi.org/10.1007/s11306-020-01662-6*
*Status : Postprint (Author's version)*

The individual peak lists can be obtained from initial spectra with, for example, the ACD/Labs free software (ACD/NMR processor). It implies the choice of a threshold to determine when an intensity level begins to be relevant and can not be associated with pure noise only. It was shown that this threshold has to be maintained at a low level, typically between 0.02 and 0.05 in ACD/Labs.

A list of pre-processing steps can be applied on the individual initial peak lists in order to increase the impact of the informative sources and to remove potential unnecessary artefacts. These steps include, for instance, the symmetrisation of homonuclear 2D spectra with respect to the diagonal, or the removal of negative intensities. With biological samples, one major problem is the strong residual water signal, even with a previous application of some solvent signal suppression techniques. As a result, a water zone deletion is very useful to avoid over-representations (it mainly concerns the water zone, but may also concern urea, maleic acid or lipoproteins). A normalization of the intensities (using Constant Sum = 1) and a further log-transformation can also be of crucial interest and added in order to stabilize distribution variances.

Finally, a dimension reduction step is also applied. In $^{1}$H-NMR spectroscopy, bucketing tools are common and widely used to control the spectral resolution and/or to overcome the misalignments problems. In this 2D workflow, a soft bucketing step adapted to 2D COSY is used in order to control the resolution level of the two-dimensional spectra. Practically, a variation of the number of decimals of the coordinates is simply 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. Using this method, the width of the COSY peaks is adjusted and the size of the resulting database is adjusted simultaneously. Furthermore, intermediate resolutions can be computed in a similar way.

## 3.2 The 2D Vectorization pre-processing workflow

The Vectorization approach is fully described in Feraud et al. (2019). Unlike the GPL approach, the vectorization one can be directly applied on raw Bruker text files. The choice of the intensity threshold when using ACD/Labs is then not anymore an issue.

The principle is quite straightforward: each individual initial ($m \times p$) 2D spectral matrix is first bucketed twice (by rows and by columns) and summarized into a ($M_V \times M_V$) object. The high dimensionality of the data and small residual peak shifts can indeed impede the future multivariate data analyses Liland (2011) and bucketing reduces such problems by integrating the $p$ original spectral intensities into $M_V$ predefined intervals, or buckets, with $M_V < p$. For convenience with standards, $M_V$ is often chosen equal to 256 or 512 here (or subsequent $2^k$ multiples according to the initial resolution) in order to control the dimension reduction and to avoid truncating extremities.

In this step, the optimal trade-off lies between keeping the spectral information and removing the peak shifts as well as decreasing the total number of variables. Among possible binning methods, The R package PepsNMR's (Martin et al. 2017) bucketing function proposes two integration options, either trapezoidal or rectangular, with equally sized buckets and is generalized to cut the original axis at any chosen location. For 2D COSY, rectangular bucketing is chosen for its more intuitive aspect linked with a kind of pixelation of a spectral grid.

*Published in : Metabolomics (2020), vol. 16, article no. 42, pp. 1-11*
*doi.org/10.1007/s11306-020-01662-6*
*Status : Postprint (Author's version)*

These bucketed 2D objects are then vectorized, transformed into a ($1 \times M_V^2$) row vector. Finally, these row vectors are stacked to form a global matrix of size ($n \times M_V^2$) containing the whole information from all the initial spectra.

Before this bucketing step, the water zone is set to zero and the normalization (Constant Sum = 1) step is performed. A subsequent log-transformation could also be considered for the same reason as the GPL methodology, but only if zero values are treated or removed before.

### 3.3 The MIC concept

Once these global objects are built, their quality in terms of informative content (with respect to their ability to allow to retrieve the main signal, which correspond to the urine donors and to the *y* response vector) has to be evaluated via the Metabolomic Informative Content (MIC) indexes. The MIC concept is fully detailed in Feraud et al. (2015) and involves a pool of clustering quality measures and criteria.

Repetitions of the sample measures have to be ideally planned during the data acquisition when signal/noise studies are of major interest. Moreover, the presence of groups to recover into the data is very important and helpful by taking the role of the above-mentioned signal. When such data are available, quality criteria to compare distinct pre-processing or acquisition strategies can be derived from unsupervised as well as supervised chemometric tools. In this context, the MIC concept mainly includes complementary inertia measures and clustering analysis quality measures.

First, the inertia analysis decomposes the total variance into two complementary parts: the variance between the groups (maximized in a good partition) and the variance within the groups of observations (minimized in a good partition). Second, the unsupervised clustering results, obtained via the Ward and K-means algorithms, are summarized into some criteria. The (adjusted) Rand indexes measure the true class recovery efficiency and should be maximized, with a maximal value of 1, while the Dunn and Davies-Bouldin indexes measure the clustering homogeneity and they have to be respectively maximized and minimized (formula details can be found in Feraud et al. 2015).

According to these measures, all the candidate data sources can then be ranked. A priori, the spectral pool which provides the best MIC performances on average would be the pool that facilitates in the best way the capture of the relevant information (to discern the useful signal relative to the noise).

## 4. Results and discussion

Conventional 2D COSY spectral data are chosen as a reference data source (Bax and Freeman 1981; Xi et al. 2006; Marchand et al. 2017) because it is one of the most sensitive 2D NMR experiments (compared to heteronuclear 2D NMR). It provides a significant dispersion of overlapped resonances while being more sensitive than heteronuclear experiments, and contains less redundancy than TOCSY spectra (Dalvit and Bovermann 1995).

The duration of a two-dimensional NMR experiment is both directly proportional to the number of time increments in the indirect dimension $t_1$, allowing this dimension to be correctly sampled, and the number of transients *NS* which is required to obtain a sufficient sensitivity. Therefore, these two parameters must be high enough to get the desired resolution and accuracy, leading to a significant increase of the total

*Published in : Metabolomics (2020), vol. 16, article no. 42, pp. 1-11*
*doi.org/10.1007/s11306-020-01662-6*
*Status : Postprint (Author's version)*

duration of 2D NMR experiments. Moreover, longer experiments become more sensitive to spectrometer instabilities, which are responsible for the appearance of $t_1$-noise in the 2D spectra.

Thus, experiments involving a reduced number of time increments $t_1$ and/or a reduced number of transients *NS* are intuitive options to decrease the experimental duration. In order not to lose too much resolution and sensitivity, a solution has been proposed in Barna and Laue (1987) to overcome this problem by introducing the concept of Non-Uniform Sampling. The principle of NUS consists in reducing the duration of 2D NMR experiments by only acquiring a fraction of the time increments $t_1$ in the indirect dimension. While NUS can be used to increase the resolution with a constant experiment time (Le Guennec et al. 2015), here it was applied to reduce the experiment time while keeping the resolution constant in the indirect dimension. As discussed in Sect. 2, a 50% NUS scheme was found optimum, leading to an experiment time decreased by two compared to the conventional experiment.

Let us remember, from Feraud et al. (2019), that the experimental design involves three different urine donors, two urine dilution levels and four days of acquisition. Eight measures are finally available for each donor. A total of 24 COSY NS4, 24 COSY NS1 and 24 COSY NUS50 individual spectra were then collected and taken into consideration for metabolomics purpose. As mentioned earlier, the first objective is to blindly retrieve the main signal of the experimental design which corresponds to the group membership (i.e. the urine donors) of the samples. PCA (Principal Component Analysis) and MIC results are successively shown in this section.

## 4.1 Descriptive PCA results

For descriptive statistics, Fig. 2 shows PCA results for six cases as input matrices: log-transformed GPL (with one decimal for the bucketing step and a ACD/Labs significance threshold of 0.02) and Vectorization for COSY NS4 data, COSY NS1 data and for COSY NUS50 data. One can see that these PCA results generally indicate a good a priori separability between the urine donors, and that this separability is not affected by the reduction of the experiment duration.

It is also interesting to visualize the effects of the ACD/ Labs significance threshold and of the number of retained decimals during the bucketing step on PCA results based on the GPL matrices (Fig. 3). When the number of decimal(s) increases and/or the significance threshold increases, the first and second urine donors tend to become more and more confused.

Consequently, GPL using large buckets (with one decimal) and a conservative significance threshold should be favored to keep a sufficient amount of information according to the signal. This will be confirmed via the use of the MIC indexes in Sect. 4.2.

## 4.2 MIC results

In this section, the MIC results obtained with conventional COSY NS4 spectra presented in Feraud et al. (2019) are supplemented by new MIC results on the faster COSY NS1 and COSY NUS50 spectral data sources.

All these results are shown in Table 1. Note that all GPL matrices are log-transformed before MIC calculations.

One can see in Table 1 that the entire results are very close between the two faster COSY experiments considered in this paper and the conventional one. Indeed, GPL matrices with larger buckets (with 1 or 2

*Published in : Metabolomics (2020), vol. 16, article no. 42, pp. 1-11*
*doi.org/10.1007/s11306-020-01662-6*
*Status : Postprint (Author's version)*

decimal(s) for the coordinates during the soft bucketing step) allow to reach perfect clustering results in almost all the cases, with Rand and adjusted-Rand indexes equal to 1. Matrices obtained via the 2D Vectorization workflow provide quite similar clustering performances and are also characterized by particularly good inertia measures (in bold in Table 1).


**Fig. 2** First two PCA factors for GPL and Vectorization of COSY NS4 data (top), COSY NS1 data (middle) and COSY NUS50 data (bottom)

*Published in : Metabolomics (2020), vol. 16, article no. 42, pp. 1-11*
*doi.org/10.1007/s11306-020-01662-6*
*Status : Postprint (Author's version)*

Thus, faster COSY NS1 and COSY NUS50 pools of spectra are very comparable with COSY NS4 spectra in terms of Metabolomic Informative Content (MIC). No significant loss of information about the signal is observed when using faster acquisition methods for 2D COSY spectra (while the experiment duration was reduced by a factor 4 and 8 for COSY NS1 and COSY NUS50, respectively).

Note that COSY NS1 and COSY NUS50 also allow to consider GPL matrices of smaller sizes (see the last columns of Table 1).

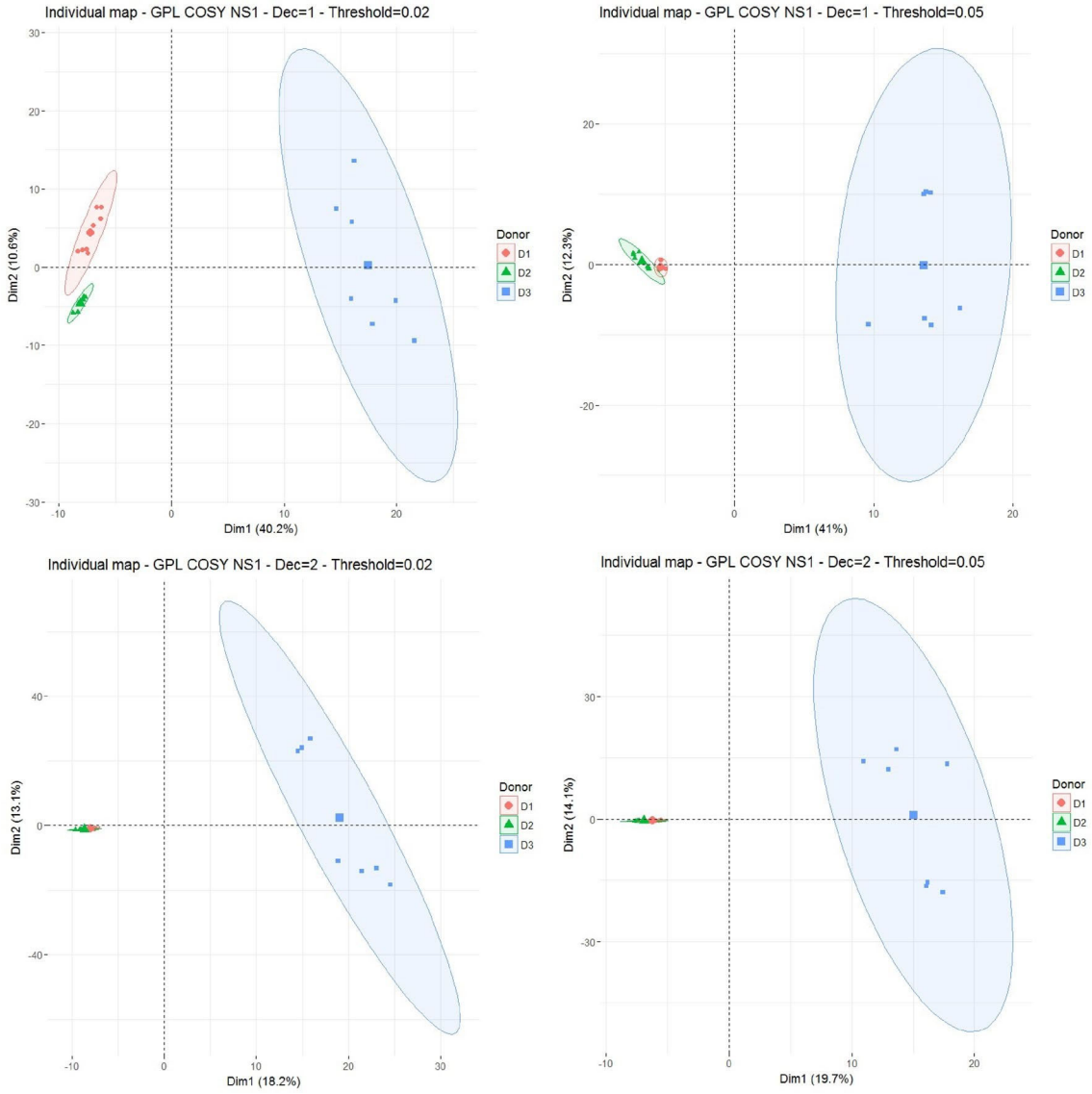



**Fig. 3** The GPL parameters (the number of decimals for bucketing and the ACD/Labs significance threshold) effects on PCA results (illustrated on the COSY NS1 spectral data)

# 5. Conclusions

It has been proved in Feraud et al. (2015, 2019) that 2D COSY spectra allow to capture a better level of Metabolomic Informative Content compared to the traditional use of 1D $^{1}$H-NMR spectra. A more massive

use of COSY in metabolomics (or -omics in general) studies is nevertheless slowed down by important and often non-acceptable experiment durations to handle a whole experimental multi-factors design.

However, the experiment duration can be significantly reduced by imposing a smaller number of transients (*NS*) during the conventional COSY experiment and/or by considering Non-Uniform Sampling (NUS) techniques which allow to consider a smaller number of time increments $t_1$.

In this paper, COSY NS1 spectra, using only one transient (*NS* = 1) and a standard number of increments, and COSY NUS50 spectra, involving a NUS percentage of 50% along with the use of only one transient again, are tested and compared with COSY NS4 spectra in order to retrieve different urine donors (the signal) in a multi-factor design.

It is demonstrated in Table 1 that the two faster COSY acquisition techniques provide very similar MIC results compared to the initial COSY. Thus, the proved advantage of the COSY data source over the 1D $^1$H-NMR one is still verified. Consequently, using techniques like the reduction of the number of transients or the Non-Uniform Sampling can really open the door to a potential more massive and high-throughput use of 2D COSY spectra in metabolomics studies and can facilitate the analyses of larger cohorts of samples.

This paper is also innovative in the sense that it is the first time that 2D pre-processing workflows (GPL and Vectorization) and MIC quality measures are applied on faster acquisition methods.

Future investigations could consider the use of alternative sampling schemes or data reconstructions schemes to improve the quality of non-uniform sampled 2D spectra (Robson et al. 2019), as well as alternative time-saving strategies. Among these strategies, ultrafast 2D NMR (as developed in Frydman et al. 2002) may be considered, but probably only with more concentrated samples because of an important loss in sensitivity. In particular, ultrafast 2D NMR makes it possible to considerably accelerate the acquisition times, but suffers from a sensitivity penalty that often requires signal averaging. So far, ultrafast 2D NMR has been applied to a variety of metabolic extracts (Jezequel et al. 2015, Le Guennec et al. 2012), but not yet on biofluids such as urine.

Boosting the sensitivity of such approaches may also rely on their promising combination with hyperpolarization methods (Dumez et al. 2015).

Finally, other potential further works would concern the biomarker identification issue using faster COSY experiments and the use of NUS on heteronuclear data sources, such as HSQC (Heteronuclear Single Quantum Correlation).

**Table 1** MIC results for conventional COSY NS4, COSY NS1 and COSY NUS50 spectra, according to the choice of different pre-processing workflows (GPL and Vectorization)

| Experiment | Data | ACD/Lab threshold | GPL decimals | MIC indexes (Ward / K-means) | | | | | | Columns in X |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Dunn | Davies-Bouldin | Rand | Adj-Rand | Btw (%) | Wth (%) | |
| COSY NS4 | GPL | 0.02 | 1 | 0.8088 / 0.8088 | 1.5696 / 1.5696 | **1.00 / 1.00** | **1.00 / 1.00** | 41.01 | 58.99 | 828 |
| COSY NS4 | GPL | 0.02 | 2 | 0.7917 / 0.7917 | 1.6943 / 1.6943 | **1.00 / 1.00** | **1.00 / 1.00** | 28.57 | 71.43 | 1954 |
| COSY NS4 | GPL | 0.02 | 3 | 0.8131 / 0.8366 | 1.8565 / 1.7369 | 0.5942 / 0.5797 | 0.1744 / 0.2088 | 12.12 | 87.88 | 3994 |
| COSY NS4 | GPL | 0.05 | 1 | 0.8792 / 0.8612 | 1.3635 / 1.3239 | **1.00 / 1.00** | **1.00 / 1.00** | **42.62** | **57.38** | 484 |
| COSY NS4 | GPL | 0.05 | 2 | 0.8699 / 0.8699 | 1.6074 / 1.6074 | **1.00 / 1.00** | **1.00 / 1.00** | 31.76 | 68.24 | 1126 |
| COSY NS4 | GPL | 0.05 | 3 | 0.8013 / 0.7642 | 1.8484 / 1.858 | 0.5942 / 0.6413 | 0.1743 / 0.2966 | 13.84 | 86.16 | 2332 |
| COSY NS4 | COS vectorization | | | 0.7555 / 0.758 | 0.6224 / 0.7166 | 0.9407 / **1.00** | 0.8605 / **1.00** | **91.75** | **8.25** | 6553 |
| COSY NS1 | GPL | 0.02 | 1 | 0.8592 / 0.6763 | 1.3456 / 1.1932 | **1.00 / 1.00** | **1.00 / 1.00** | **55.36** | **44.64** | 337 |
| COSY NS1 | GPL | 0.02 | 2 | 0.7543 / 0.7543 | 1.5941 / 1.5941 | **1.00 / 1.00** | **1.00 / 1.00** | 35.71 | 64.29 | 895 |
| COSY NS1 | GPL | 0.02 | 3 | 0.7332 / 0.7359 | 1.8511 / 1.8802 | 0.5889 / 0.5889 | 0.153 / 0.153 | 15.68 | 84.32 | 185 |
| COSY NS1 | GPL | 0.05 | 1 | 0.623 / 0.623 | 1.3715 / 1.3715 | **1.00 / 1.00** | **1.00 / 1.00** | **56.24** | **43.76** | 201 |
| COSY NS1 | GPL | 0.05 | 2 | **0.9902 / 0.9902** | 1.4361 / 1.4361 | 0.6996 / 0.6996 | 0.4038 / 0.4038 | 37.77 | 62.23 | 509 |
| COSY NS1 | GPL | 0.05 | 3 | 0.8248 / 0.6065 | 1.639 / 1.8308 | 0.6245 / 0.8142 | 0.2834 / 0.5869 | 16.81 | 83.19 | 106 |
| COSY NS1 | COS vectorization | | | 0.5733 / 0.6442 | 0.6226 / 0.7163 | **1.00 / 1.00** | **1.00 / 1.00** | **91.38** | **8.62** | 655 |
| COSY NUS50 | GPL | 0.02 | 1 | 0.8024 / 0.8024 | 1.3264 / 1.3264 | **1.00 / 1.00** | **1.00 / 1.00** | **56.25** | **43.75** | 348 |
| COSY NUS50 | GPL | 0.02 | 2 | 0.7531 / 0.7531 | 1.5941 / 1.5941 | **1.00 / 1.00** | **1.00 / 1.00** | 36.15 | 63.85 | 996 |
| COSY NUS50 | GPL | 0.02 | 3 | 0.7221 / 0.7221 | 1.9022 / 1.9022 | 0.8007 / 0.8007 | 0.5657 / 0.5657 | 13.25 | 86.75 | 210 |
| COSY NUS50 | GPL | 0.05 | 1 | 0.6019 / 0.5689 | 1.172 / 1.4532 | **1.00 / 1.00** | **1.00 / 1.00** | 50.07 | 49.93 | 203 |
| COSY NUS50 | GPL | 0.05 | 2 | 0.6205 / 0.8603 | 1.5559 / 1.5977 | 0.8261 / 0.7102 | 0.6102 / 0.4103 | 31.89 | 68.11 | 574 |
| COSY NUS50 | GPL | 0.05 | 3 | 0.6564 / 0.6176 | 1.5095 / 1.9596 | 0.4964 / 0.5942 | 0.093 / 0.0987 | 14.00 | 86.00 | 116 |
| COSY NUS50 | COS vectorization | | | 0.6783 / 0.6783 | 0.8518 / 0.8518 | **0.9447** / **0.9447** | 0.8693 / 0.8693 | **82.14** | **17.85** | 65536 |



*Published in : Metabolomics (2020), vol. 16, article no. 42, pp. 1-11*
*doi.org/10.1007/s11306-020-01662-6*
*Status : Postprint (Author's version)*

## Acknowledgements

Support from the IAP Research Network P7/06 of the Belgian State (Belgian Science Policy) is gratefully acknowledged. Support from the CORSAIRE metabolomics platform (Biogenouest network) is also acknowledged. Pascal de Tullio is Research Director of the Fonds de la Recherche Scientifique (FNRS).

## Author contributions

BF, BG, PG and PT conceived and designed research. EM, JL and PT collected and supplied the data. BF analyzed data and wrote the manuscript. All authors read and approved the manuscript.

## Data availability statement

The metabolomics and metadata reported in this paper are available on demand from the Institute of Statistics, Biostatistics and Actuarial Sciences, UCLouvain, Belgium.Softwares availability statement The raw data were processed with the Bruker Topspin 3.5 software. Peak lists were extracted using ACD/Labs 12.00 (ACD/ NMR processor). The R software (http://www.R-projec t.org) environment was exclusively used for statistical purpose.

## Compliance with ethical standards

**Conflict of interest** All authors declare that they have no conflict of interest.

**Ethical Approval** All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

**Informed consent** Informed consent was obtained from all individual participants included in the study

**Research involving human participants** This study analyzes collected data which involved human participants. The studies were approved by our local Ethics Committee (CHR Citadelle, Liège, number B412201215082-1267). The samples used in this article are coming from a study approved by the CHR hospital but also for a collaboration with Pascal de Tullio from the University of Liège. All subjects gave their informed consent.

# References

Barna, J. C., & Laue, E. D. (1987). Conventional and exponential sampling for 2D NMR experiments with application to a 2D NMR spectrum of a protein. *Journal of Magnetic Resonance (1969)*, *75*(2), 384–389.

Bax, A. D., & Freeman, R. (1981). Investigation of complex networks of spin-spin coupling by two-dimensional NMR. *Journal of Magnetic Resonance (1969)*, *44*(3), 542–561.

Dalvit, C., & Bovermann, G. (1995). Pulsed field gradient one-dimensional NMR selective ROE and TOCSY experiments. *Magnetic Resonance in Chemistry*, *33*(2), 156–159.

Delikatny, E. J., Hull, W. E., & Mountford, C. E. (1991). The effect of altering time domains and window functions in two-dimensional proton COSY spectra of biological specimens. *Journal of Magnetic Resonance (1969)*, *94*(3), 563–573.

Dumez, J. N., Milani, J., Vuichoud, B., Bornet, A., Lalande-Martin, J., Tea, I., et al. (2015). Hyperpolarized NMR of plant and cancer cell extracts at natural abundance. *Analyst*, *140*(17), 5860–5863.

Feraud, B., Govaerts, B., Verleysen, M., & De Tullio, P. (2015). Statistical treatment of 2D NMR COSY spectra in metabolomics: Data preparation, clustering-based evaluation of the Metabolomic Informative Content and comparison with 1H-NMR. *Metabolomics*, *11*(6), 1756–1768.

Feraud, B., Leenders, J., Martineau, E., Giraudeau, P., Govaerts, B., & De Tullio, P. (2019). Two data pre-processing workflows to facilitate the discovery of biomarkers by 2D NMR metabolomics. *Metabolomics*, *15*, 63. https://doi.org/10.1007/s11306-019-1524-3.

Frydman, L., Scherf, T., & Lupulescu, A. (2002). The acquisition of multidimensional NMR spectra within a single scan. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(25), 15858–15862.

Hoch, J. C., Maciejewski, M. W., Mobli, M., Schuyler, A. D., & Stern, A. S. (2014). Nonuniform sampling and maximum entropy reconstruction in multidimensional NMR. *Accounts of Chemical Research*, *47*(2), 708–717.

Hyberts, S. G., Takeuchi, K., & Wagner, G. (2010). Poisson-gap sampling and forward maximum entropy reconstruction for enhancing the resolution and sensitivity of protein NMR data. *Journal of the American Chemical Society*, *132*(7), 2145–2147.

Jezequel, T., Deborde, C., Maucourt, M., Zhendre, V., Moing, A., & Giraudeau, P. (2015). Absolute quantification of metabolites in tomato fruit extracts by fast 2D NMR. *Metabolomics*, *11*(5), 1231–1242.

Kazimierczuk, K., & Orekhov, V. Y. (2011). Accelerated NMR spectroscopy by using compressed sensing. *Angewandte Chemie*, *123*(24), 5670–5673.

Le Guennec, A., Dumez, J. N., Giraudeau, P., & Caldarelli, S. (2015). Resolution-enhanced 2D NMR of complex mixtures by nonuniform sampling. *Magnetic Resonance in Chemistry*, *53*(11), 913–920.

Le Guennec, A., Tea, I., Antheaume, I., Martineau, E., Charrier, B., Pathan, M., et al. (2012). Fast determination of absolute metabolite concentrations by spatially encoded 2D NMR: Application to breast cancer cell extracts. *Analytical Chemistry*, *84*(24), 10831–10837.

Liland, K. H. (2011). Multivariate methods in metabolomics, from preprocessing to dimension reduction and statistical analysis. *TrAC Trends in Analytical Chemistry*, *30*(6), 827–841.

Maciejewski, M. W., Mobli, M., Schuyler, A. D., Stern, A. S., & Hoch, J. C. (2011). Data sampling in multidimensional NMR: Fundamentals and strategies. In M. Billeter & V. Orekhov (Eds.), *Novel sampling approaches in higher dimensional NMR* (pp. 49–77). Berlin: Springer.

Marchand, J., Martineau, E., Guitton, Y., Dervilly-Pinel, G., & Giraudeau, P. (2017). Multidimensional NMR approaches towards highly resolved, sensitive and high-throughput quantitative metabolomics. *Current Opinion in Biotechnology*, *43*, 49–55.

Marchand, J., Martineau, E., Guitton, Y., Le Bizec, B., Dervilly-Pinel, G., & Giraudeau, P. (2018). A multidimensional 1H-NMR lipidomics workflow to address chemical food safety issues. *Metabolomics*, *14*(5), 60.

Martin M., Legat B., Leenders J., Vanwinsberghe J., Rousseau R., et. al., (2017) PepsNMR for the 1H-NMR metabolomic data preprocessing. ISBA Discussion Paper, 2017/22, http://hdl.handl e.net/2078.1/18715 9.

Robson, S., Arthanari, H., Hyberts, S. G., & Wagner, G. (2019). Nonuniform sampling for NMR spectroscopy. *Methods in Enzymology*, *614*, 263–291.

Rouger, L., Gouilleux, B., & Giraudeau, P. (2017). Fast n-dimensional data acquisition methods. In G. E. Tranter & D. W. Koppenaal (Eds.), *Encyclopedia of spectroscopy and spectrometry* (3rd ed., pp. 588–596). Oxford: Academic Press.

Xi, Y., de Ropp, J. S., Viant, M. R., Woodruff, D. L., & Yu, P. (2006). Automated screening for metabolites in complex mixtures using 2D COSY NMR spectroscopy. *Metabolomics*, *2*(4), 221–233.