European Network of Fourier-Transform Ion-Cyclotron-Resonance Mass Spectrometry Centers

Grant Agreement n° 731077

Deliverable D03.05 – Specific processing, analysis and mining tool for the data processing

Start date of the project: 1st January 2018
Duration: 60 months
Project Coordinator: Christian ROLANDO – CNRS-
Contact: christian.rolando@univ-lille.fr

"This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731077

Document Classification

Title	Specific processing, analysis and mining tools for the data processing
Authors	P11 CASC4DE – Camille Beluffi Marin
Work package	WP3 – Open Data and e-Infrastructure
Dissemination	PU = Public
Nature	R: Document, report
Doc ID Code	20220726_EU_FT-ICR_MS_D03.05
Keywords	Data mining, processing

Document History

Name	Date	Comment
P11 CASC4DE – Camille Beluffi Marin	2022-06-28	Upload

Document Validation

Project Coordinator	Date	E-mail
P1 CNRS – Christian Rolando	2022-07-26	christian.rolando@univ-lille.fr

Neutral Reviewer	Date	E-mail
P1 CNRS – Christian Rolando	2022-07-26	christian.rolando@univ-lille.fr

Document Abstract

The EU FT-ICR MS project is responsible for the production of a large amount of data. Some of this data is produced by external users of the platform, but each laboratory participating in the project has also been responsible for performing FTICR MS analysis of the same sample in order to evaluate the method, the repeatability within the network and the quality of the different laboratories. This so-called "round-robin" analysis generates a data set with variable parameters for the same molecule, which is ideal for the application of data mining methods.
Data mining is a method that generally allows to analyze a dataset from a more global point of view. Beyond a simple spectrum analysis, the method consists in developing an automatic processing of the whole available data in order to highlight a trend within the dataset. This can also allow the detection of systematic biases in the measurements or to generate models allowing, for example, the prediction of future behavior. This method can therefore provide answers to questions that cannot be answered by more traditional isolated analyses.

Introduction

The EU FT-ICR MS project is responsible for the production of a large amount of data. Some of this data is produced by external users of the platform, but each laboratory participating in the project has also been responsible for performing FTICR MS analysis of the same sample in order to evaluate the method, the repeatability within the network and the quality of the different laboratories. This so-called "round-robin" analysis generates a data set with variable parameters for the same molecule, which is ideal for the application of data mining methods.

In this case, the molecule analyzed by the different laboratories is glutathione (ECG) with a theoretical monoisotopic mass of 307.323 g/mol. Glutathione (Glutamate-Cysteine-Glycine) is naturally present in many plants, animal cells or fungi.

It is possible to calculate the theoretical spectrum of the molecule (Figure 1) in order to compare it to the results obtained by the different laboratories. It is studied here the positive ion of glutathione (ECG+), of monoisotopic mass 308,09 Da.

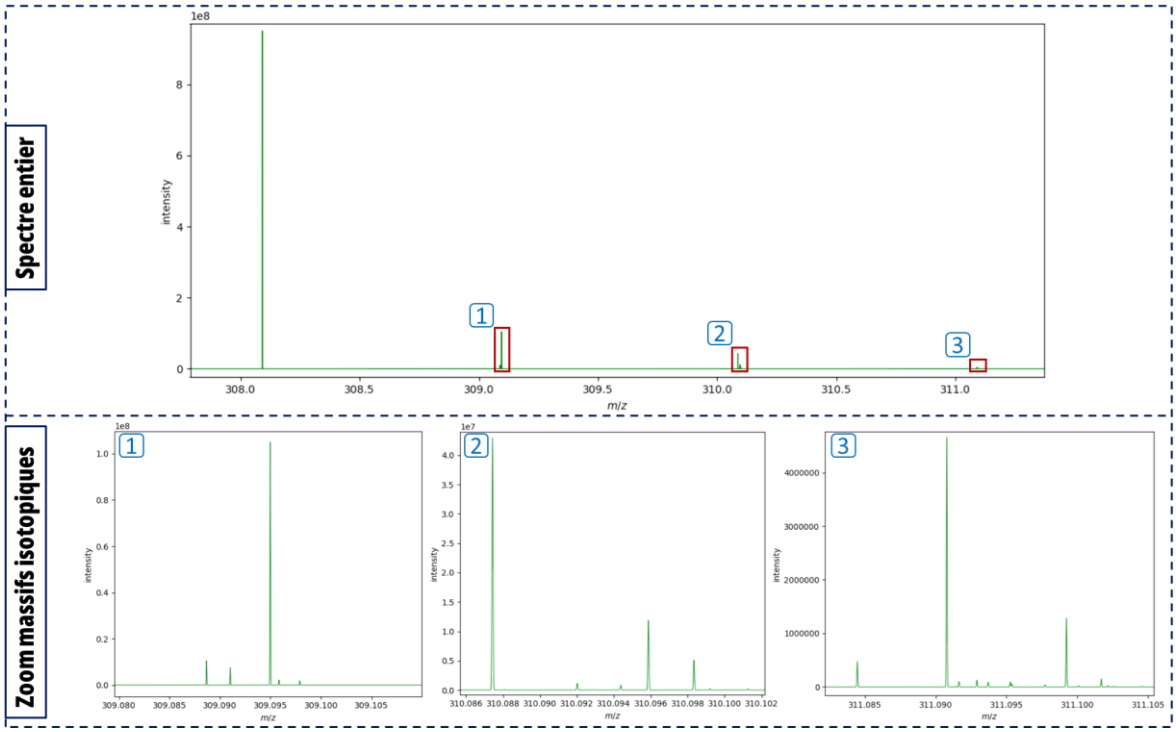

Figure 1 - Theoretical spectrum of the positive ion of Glutathione (ECG+) obtained with the NeutronStar simulation library and on the bottom-line zooms on the isotopic patterns.

Data mining is a method that generally allows to analyze a dataset from a more global point of view. Beyond a simple spectrum analysis, the method consists in developing an automatic processing of the whole available data in order to highlight a trend within the dataset. This can also allow the detection

of systematic biases in the measurements or to generate models allowing, for example, the prediction of future behavior. This method can therefore provide answers to questions that cannot be answered by more traditional isolated analyses.

Acquisition parameters and data processing

Here, it is interesting to follow different parameters between the theoretical spectrum and the spectra acquired by the laboratories according to variables of acquisition and data processing. Among the parameters followed we can find the position and intensity of the peaks, the signal to noise ratio or the general similarities between the practical and theoretical spectra. Regarding the variables that apply to the different spectra, there are different kinds. First of all, we have the variables intrinsic to the experiment, including the magnetic field (B_0) depending on the spectrometer used and the narrowband or broadband method chosen. Then we have the acquisition variables which include the duration and size of the fid or the number of scans. Finally, the data processing variables including the chosen apodization (type, value, combinations), the zero-filling (zf) or the truncation of the fid.

The magnetic field

The magnetic field used during an acquisition on an FT-ICR mass spectrometer has an impact on the results obtained. Theoretically, the higher the magnetic field, the higher the resolution power, the higher the precision and the faster the acquisition speed (Karabacak et al. 2010).

Number of scans, size and duration of the fid

The resolution and accuracy of the spectrum obtained are also dependent on the number of scans performed and the duration of the fid. Indeed, the greater the number of scans performed, the greater the risk of losing resolution due to measurement instability, as well as increasing the signal-to-noise ratio due to the accumulation of scans. In addition, if the fid is truncated, there is a significant risk that the quality of the experiment will be affected.

Acquisition type : NarrowBand et BroadBand

The MS spectra called “narrowband” or “broadband” are derived from different types of acquisitions. Indeed, the pulse sequences used will be different and determined in order to observe different specificities (Wimperis 1994).

Broadband spectra are wide spectra, which will scan a large spectral width and obtain broader information, with the disadvantage of often having a lower resolution. Indeed, even if the resolution theoretically depends only on the duration of the fid, to have a high resolution in broadband, thus on a large spectral width, a large number of points is necessary. For practical reasons, the fid may be truncated and shorter than necessary to have a very good resolution.

A narrowband spectrum scans a smaller frequency band, depending on what we want to observe, we have information on a very precise area where we know, for example, that we want to observe a particular signal of a molecule. This method has the advantage of ease of measurement since it will contain fewer acquisition points and will generally require simpler excitation. However, the dynamic range of measurement is weaker, one will not have information on the signals out of the selected zone of measurement.

Use of apodisation functions

During the acquisition of a mass spectrum, the ions will tend to be out of phase during the experiment. This phase shift among the ions of the observed signal will cause a disturbance during the Fourier transform (FT) of the associated fid.
Indeed, the FT uses the average of the signal over the entire duration of the measurement and is therefore impacted if disturbances appear and the signal changes over time.
The application of an apodization before carrying out the FT makes it possible to eliminate the parts of the fid most sensitive to disturbances. These “perturbed” parts of the signal are generally found at the beginning of the fid when the ions have not yet fully entered into stable oscillation and at the end of the fid when the ion “packet” will start to slightly diphase, following small collisions for example.

Two types of apodization have been tested in this work, the Kaiser apodization and the sinusoidal bell apodization. These are two different apodization functions, these functions set up a window where the signal will be kept. They are generally bell-shaped functions, null outside the signal, which will have the maximum towards the middle of the interval (so here in the center of the fid) and which will become thinner on the edges. During the multiplication of these functions with the signal, the information in the center will be kept while the information on the edges of the apodization window will be eliminated.

The Kaiser apodization was therefore applied to the signal before FT. This apodization, also known as Kaiser-Bessel and developed by James Kaiser for Bell Laboratories, is a family of apodization functions

has a single parameter (which we will call β here) and is generally used in spectral analysis. As shown in Figure 2, the larger the β parameter, the thinner the bell and therefore the more the signal on the edges of the fid will be filtered out to keep only the center.

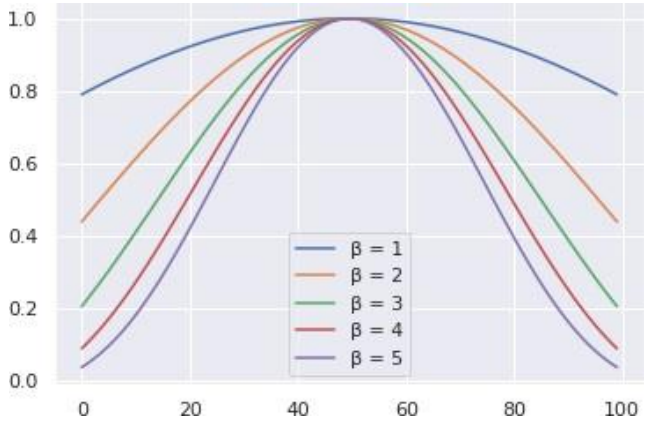

Figure 2 - Kaiser apodization function for different values of β .

A second type of apodization was evaluated, the sine-bell apodization (De Marco et Wüthrich 1976). This type of apodization was proposed in 1976 and has since been classically applied to the processing of spectra in order to improve their quality. As shown in Figure 3, we have applied here a sinusoidal bell apodization by varying the location of the maximum of the bell.

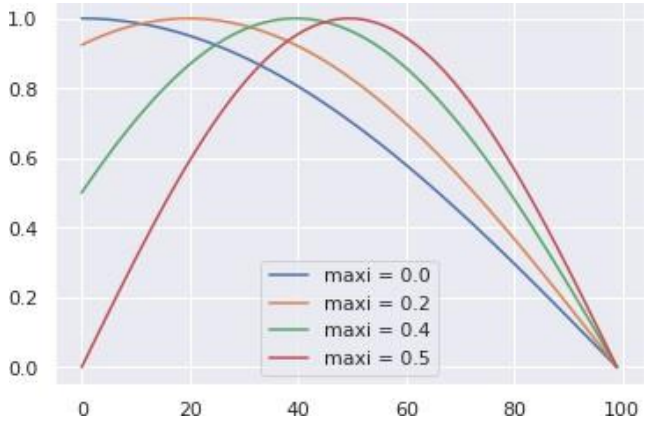

Figure 3 – Sine-bell apodization function for different values of maxi.

These two types of apodization functions are quite general and allow to represent other apodization windows depending on the parameters (for example for β = 5 we are in the presence of a Hamming apodization).

Use of Zero-Filling

The application of a Zero-Filling (or zf) process on MS data consists in adding empty points (adding zeros) at the end of the FID before performing the Fourier transform processing. As indicated, the added points are zeros, therefore of zero amplitude, which theoretically has no impact on the information of the data.

The objective of this treatment is to improve the quality of a data since the fact of artificially enlarging the FID allows to increase the number of points per ppm in the data after treatment.

Material and Methods

Datasets

The data used for this data-mining analysis are therefore, as previously indicated, those resulting from the round-robin applied to the Glutathion sample carried out within the framework of the H2020 EU-FTICR-MS project among the various laboratories involved.

In particular we have integrated here the data from 9 of the laboratories participating in the EU FT-ICR MS project:

- COBRA Laboratory - UMR6014 CNRS/URN/INSA Rouen - France
- Mass Spectrometry Laboratory - University of Liege - Belgium
- Skoltech Institute of Science and Technology (Skolkovo) - Moscow - Russia
- Institute of Microbiology - Prague - Czech Republic
- University of Rostock - Rostock - Germany
- University of Paris Sud - Orsay - France
- University of Lille - Lille - France
- University of Eastern Finland - Joensuu - Finland
- University of Lisbon (Faculty of Science) - Lisbon - Portugal

Each of the laboratories performed one or more acquisitions with specific parameters concerning for example the type of acquisition (broadband/narrowband) or the duration or the size of the fidelity, and on different sites of analysis thus providing 20 different data for the analysis.

Data Processing Algorithm

In a first step the theoretical spectrum with fine isotopic clusters was produced.

The lists of parameters to be explored were set up, as well as the list of available data sets. An iterator was then set up to generate a spectrum and perform the peak-picking associated with each possible

parameter set in order to compare them to the theoretical spectrum obtained with the `NeutronStar` library (Kreitzberg et al. 2020).

The 20 experimental data present different acquisition parameters and thus allow to explore in part the effect of the experimental parameters used. A part of the experiments was performed in NarrowBand and the other part in BroadBand. The magnetic field used during the acquisitions is included in the following interval:

$$B_0\ (T\&sla) = [7, 9, 12, 15]$$

The number of scans (NS), the duration as well as the size in points of the fid also vary between the different data in the following intervals:

$$NS = [4, 16, 24, 32, 50, 64, 100, 200, 400, 3000]$$
$$:;d\ s;z\&\ (po;nts) = [131072, 1048576, 2097152, 4194304, 8388608, 16777216]$$
$$:;d\ durat;on\ (\mathrm{sec}) = [0.42, 0.7, 0.84, 1.21, 1.4, 1.68, 1.84, 3.36, 3.91, 4.82, 8.39, 11.38]$$

In addition, the parameters related to data processing have also been varied. The intensity of the zero-filling applied to the data is in the following range (0 meaning no zero-filling is applied):

$$Z\&ro - F;ll;ng = [0, 2, 4, 6, 8]$$

Concerning the apodization, different methods were also tested. Either no apodization was applied, or one of the two apodizations presented previously was applied (Kaiser or sine bell) and finally the combination of both apodizations was also tested. Furthermore, whenever one or both apodizations were applied, the parameter associated with the apodization, Maxi or β, was varied in the following ranges:

$$\beta\ (Ka;s\&r) = [1, 2, 3, 4, 5]$$
$$Max;\ (Cloch\&\ S;nusoïdal\&) = [0.0, 0.2, 0.4, 0.5]$$

All possible combinations between experimental and processing parameters were performed, generating a total of 6000 different combinations.

For the sake of readability and as an illustration for the figures, the 20 file names have been replaced by the name of the laboratory followed by the data number if there are several for the same laboratory.

For the development of the program, the python language and various associated libraries were used, in particular the `SPIKE` libraries for the treatment of the mass spectrometry data, and `NeutronStar`, program allowing to finely simulate the isotopic masses.
The `Pandas` library was used for the management of the data after the processing, and the `Seaborn` library is used to produce the analysis graphs.

Results

Once the data set has been processed by varying the parameters presented above, the results are stored in csv format and interpreted using the `pandas` library.

For the representation of the data mining results, we use “violin” graphs.
A violin graph is a hybrid graph between a box plot and a kernel density plot. It is used to visualize the distribution of numerical data. Unlike a box plot, which only shows the statistics, "violin" plots represent the statistics and the entire distribution of the data, so they are more detailed than a classical box plot (Figure 4).

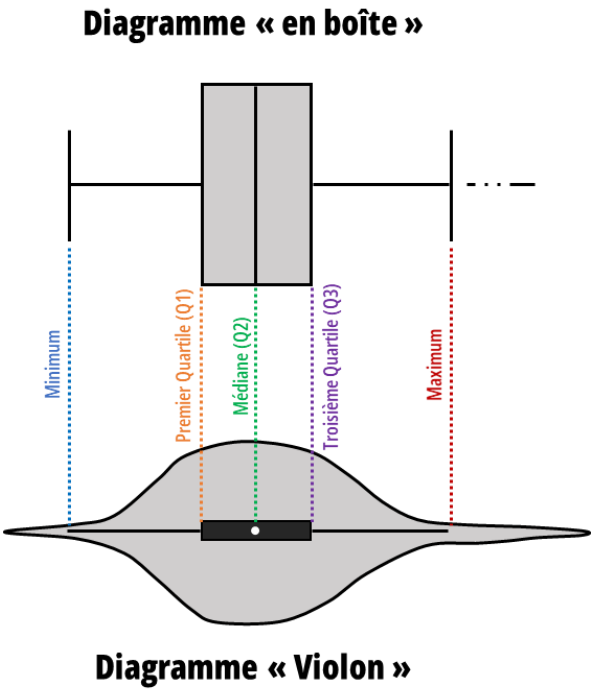

Figure 4 - Diagram of the reading of a violin graph in comparison with a box plot. On each of the graphs the median and quartile information is available as well as the "minimum" and "maximum". Outliers are also visible outside the "minimum" and "maximum" points. Generally a value is considered an outlier when it is outside 1.5 times the interquartile range [Q1 – 1.5x(Q3-Q1) ; Q3 + 1.5x(Q3-Q1)].

The effects of different data acquisition and processing parameters were observed on the signal-to-noise ratio (SNR) of the acquired data as well as on the cosine similarities between a theoretical spectrum and the experimental spectrum.

The signal to noise ratio (SNR) allows to compare the level of the desired signal to the level of the background noise present on the measurement. By measuring the SNR, we describe the difference in power between the signal and the noise. The SNR can therefore be interpreted as an approximation of the quality of a measurement or spectrum, in fact, the lower the noise compared to the desired signal, the better the quality of measurement.

The cosine similarity allows to describe the similarity between two elements by calculating the cosine of the angle between two given vectors. It is a method commonly used in mass spectrometry for spectrum comparison, especially in the framework of molecular networks (GNPS).

Here we have calculated the cosine similarities between the spectra processed with different parameters and the theoretical spectrum presented in Figure 1. In order to apply the calculation, a preprocessing of the data is applied upstream. Indeed, the spectra must be transformed into vectors. Each of the two vectors, theoretical and experimental, are calculated to make the same length, keeping information only between a minimum and a maximum value of m/z and only on the areas where information is observed on the theoretical spectrum, the rest of the information being set to 0. The spectrum is scanned in steps of 0.0005 Da to maintain sufficient accuracy and the data present in each interval is transposed as a point in the final vector.

Acquisition parameters

First, the variations of the signal-to-noise ratio as well as of the cosine similarity between the experimental spectra and the theoretical spectrum were observed as a function of the acquisition parameters, which are thus dependent and chosen at the time of the experiment.
All elements have been gathered on Figure 5 and Figure 6, allowing to identify the elements that will impact or not the experimental spectra.

Within the framework of this study and with the data at our disposal, it is observed through Figure 5 that the choice of acquisition parameters has an impact on the data obtained when a simple Fourier transform treatment is applied.

It should be noted that these elements are obtained from the original data, without increasing the database since no variations due to post-acquisition processing parameters are taken into account. Some information seems to be “missing” on the graphs of the figure because there was not enough information to generate a violin graph.

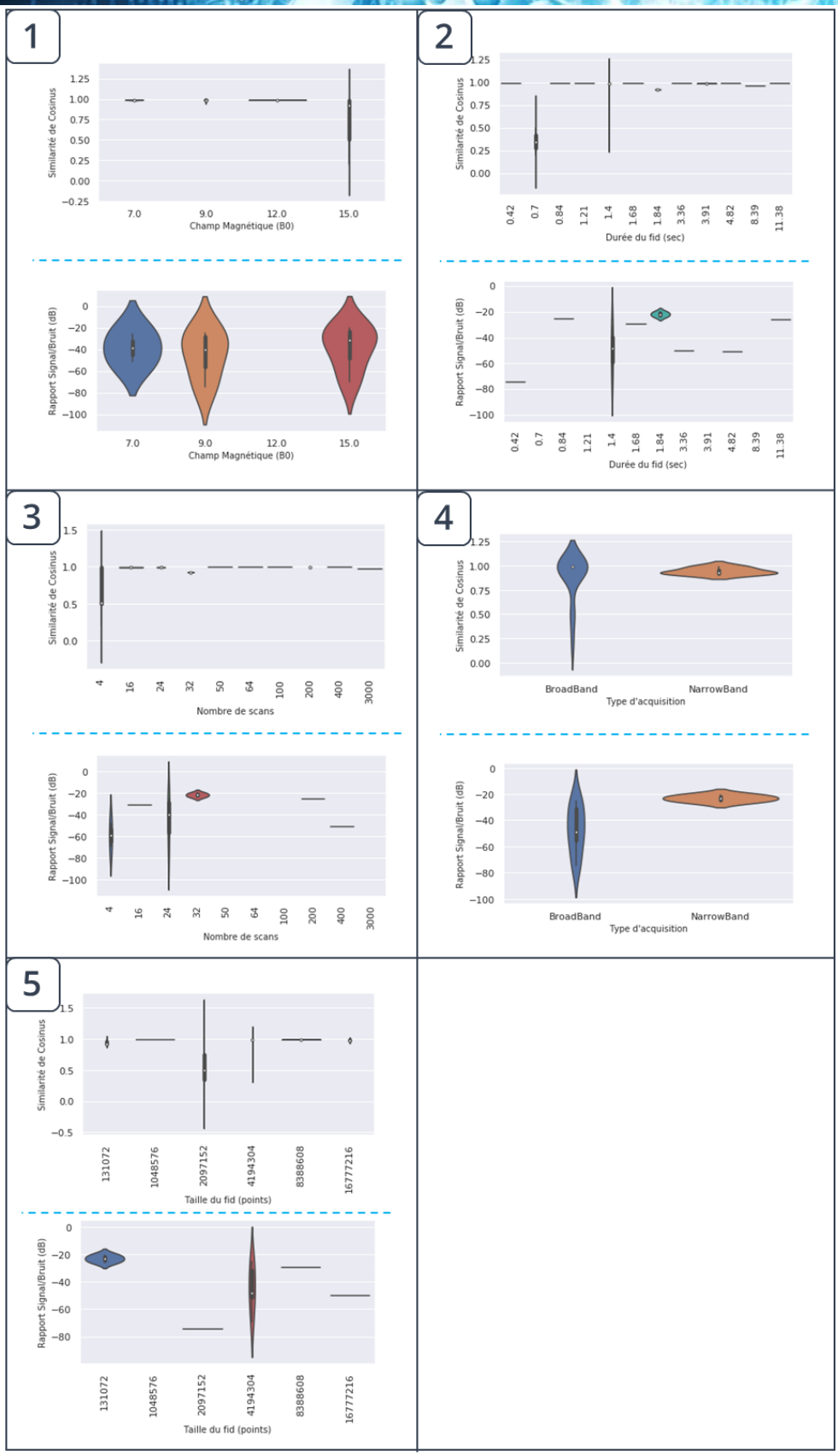

Figure 5 - Effects of different acquisition-associated parameters on cosine similarity and signal-to-noise ratio (SNR) without any post-acquisition data processing other than a Fourier transform. 1. The magnetic field B0, 2. the fid duration (in seconds), 3. the number of scans (NS), 4. the type of acquisition (Narrow and Broad Band) and 5. The size of the fid (in points).

In particular, we observe that NarrowBand acquisitions allow a better signal-to-noise ratio as well as a greater proximity to the theoretical spectrum. These data also show that a fidelity time or a number of scans that are too small induce higher noise levels.

On the other hand, Figure 6 regrouping the effects of the acquisition parameters after a Fourier transform treatment with apodization and zero-filling, shows that the different experimental parameters explored have no significant influence on the two variables once a post-acquisition treatment is applied to the data.
Indeed, none of the variations in the values of the magnetic field, the duration and size of the fid, the number of scans or the type of acquisition seem to decrease or improve the quality of the different acquisitions, neither in terms of noise nor in terms of cosine similarity.

The process of processing the data set thus allows smoothing of differences that could be related to the acquisition. This smoothing is probably also partly induced by the increase of the available data base through the number of possibilities tested for the different processing parameters.

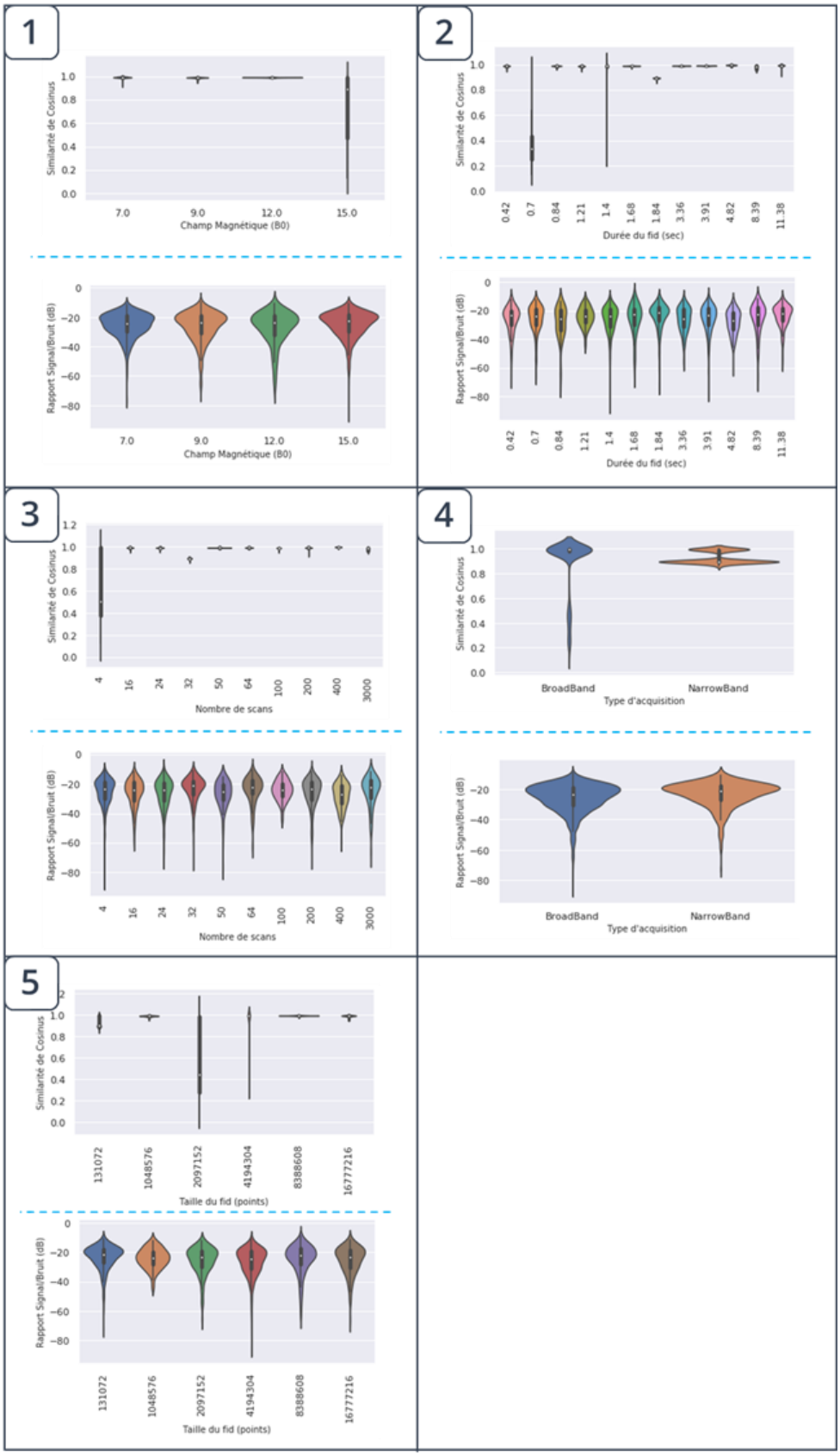

Figure 6 - Effects of different acquisition-associated parameters on cosine similarity and SNR after data processing including zero-filling and apodization. 1. The magnetic field B0, 2. the fid duration (in seconds), 3. the number of scans (NS), 4. the type of acquisition (Narrow and Broad Band) and 5. The size of the fid (in points).

Data Processing parameters

The parameters used for the treatment of the whole dataset were then also evaluated.

Figure 7 first shows the observed effect of applying or not applying an apodization or a combination of apodizations on the experimental data, still using cosine similarity and signal-to-noise ratio as evaluation criteria.

The first panel allows us to observe the effect of using an apodization when processing the dataset in a general way.
It is thus possible to see that the signal to noise ratio is not really impacted by this treatment.
The cosine similarity between the experimental spectra and the theoretical one is improved by the apodization. Indeed, we can see the average getting closer to a cosine similarity equal to 1 and the data being more tightly packed around it. The apodization therefore brings an improvement of the quality of the data sets. The type of apodization (Kaiser or sine-bell) does not seem to make a real difference, although the combination of the two apodizations seems to accentuate the effect of tightening the data around “1” and thus be the best treatment to apply.

The second panel aims at identifying the ideal parameter specific to each of the two apodizations presented previously.
On the left, we observe the effect of the different values of the "Maxi" associated with the sine bell apodization. It seems that the impact of the variation of this value is quite limited and the quality of the data after processing is not much affected. Nevertheless, with a value of Maxi=0.5, the data are tighter around the mean in terms of signal to noise ratio and the mean value of cosine similarity is very close to 1. This value can therefore be chosen for the sine bell apodization when applying the combination of apodizations for the optimized processing of the data set.
On the right is the effect of different values of β associated with Kaiser apodization. It is visible here with the increase of β an improvement both on the signal to noise ratio and on the cosine similarity, with a plateau after β=3, which can thus be the value chosen for the optimal processing of the whole dataset.

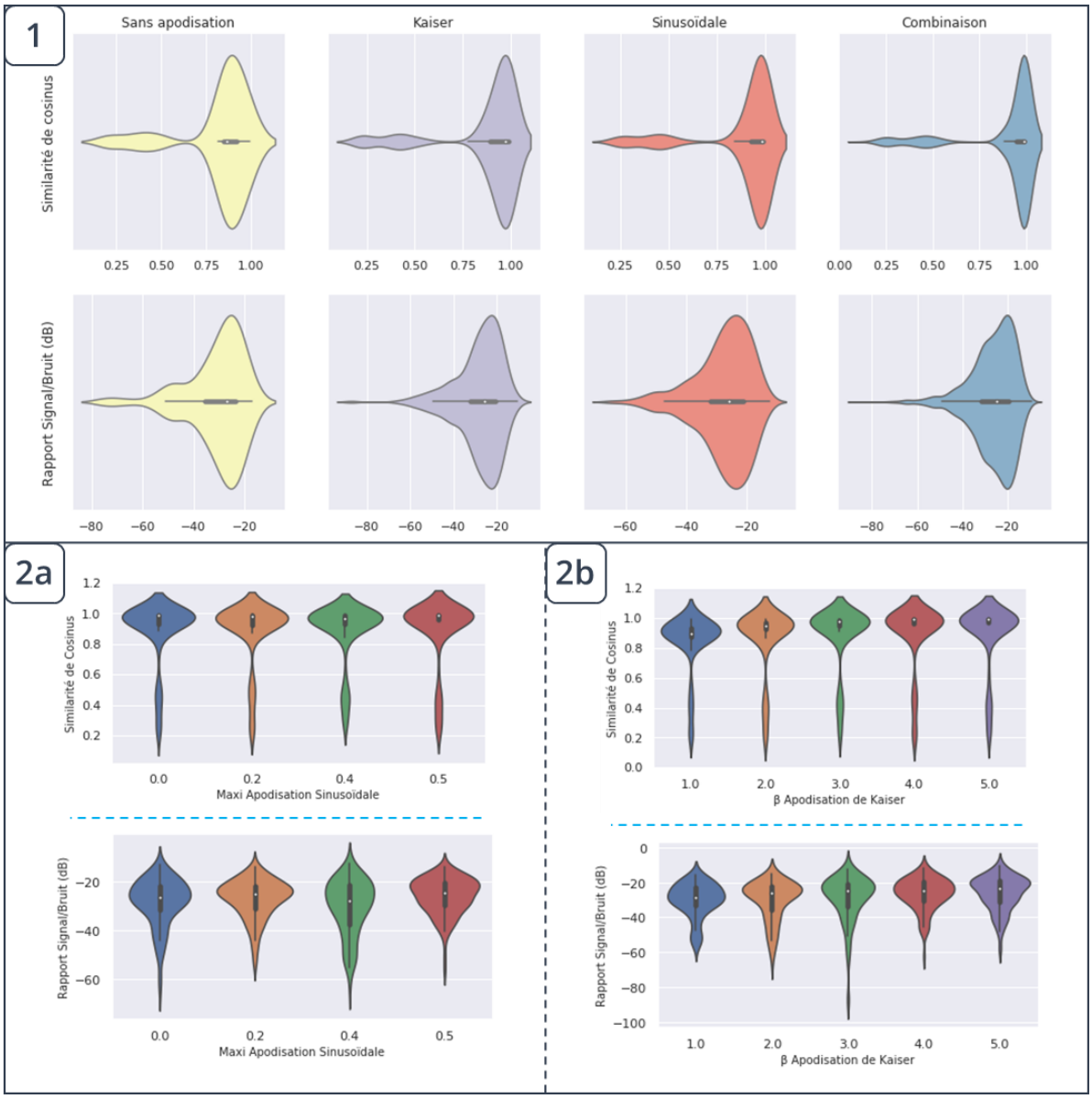

Figure 7 – 1. Effect of applying or not applying an apodization (Kaiser, sine bell, or a combination of both) on the cosine similarity between the data and a theoretical spectrum and the signal-to-noise ratio (SNR). 2. impact of varying the Maxi and β parameters associated with (a) sine-bell (b) Kaiser apodizations.

Finally, the impact of applying zero-filling during data processing was also analyzed. Figure 1 shows the results obtained on the cosine similarity as well as on the signal to noise ratio when applying different degrees of zero-filling, without any other associated data processing (no apodization in particular).

Thus, zero-filling slightly impacts the cosine similarity between the theoretical and experimental spectra. Indeed, the more intense the zero-filling applied, the more the data move away from the theory since the cosine similarity deviates from 1

Nevertheless, the more the zero-filling is important the less the noise in the spectrum is intense. This phenomenon of noise reduction seems to diminish with a zero-filling beyond 4.

It is therefore possible to use a zero-filling set to 2 or 4 which seem to be optimal values for the whole data processing, in order to decrease the noise without impacting too strongly the cosine similarity.

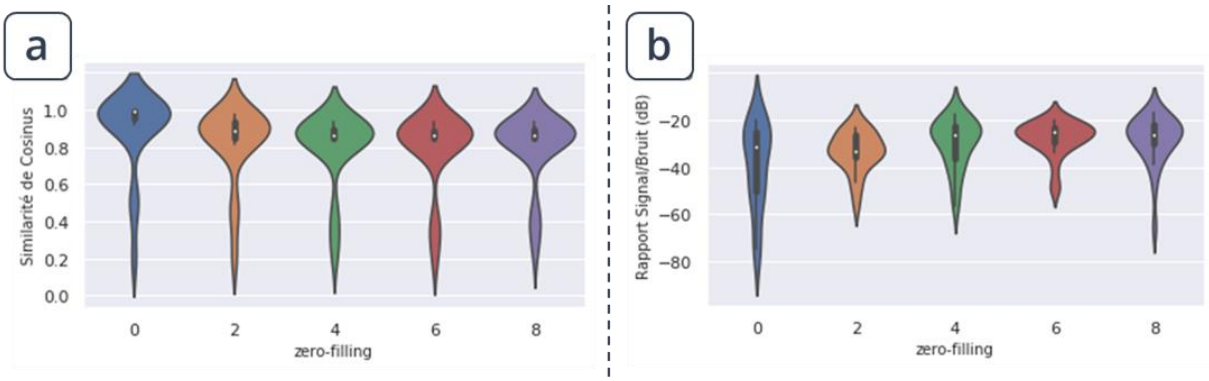

Figure 1 - Effect of applying zero-filling alone on (a) cosine similarity and (b) SNR.

Conclusions & Perspectives

This work, carried out on data from different laboratories, performed by different experimenters on different equipment, has therefore allowed the implementation of a common analysis of all available data. This analysis includes the development of a post-acquisition treatment as well as generalized and global comparisons of the data between them and with what can be expected theoretically.

The evaluation of the reproducibility of the analysis could be very extensive here, since experimental measurements are performed in different locations, with different operators and measurement systems as well as in different temporalities on a similar glutathione sample between the different laboratories.

In this study, many parameters involved during the acquisition or during the subsequent processing were varied and their impact was evaluated.
It was shown that the experimental parameters have a relatively large impact on the quality of the data for the two variables, cosine similarity and signal to noise ratio, used in the evaluation.
On the other hand, after processing by Fourier transform, apodization and zero-filling, the impact of the acquisition parameters on the cosine similarity or the signal-to-noise ratio is strongly reduced for the final spectrum. The differences that can be observed are indeed smoothed by the different post-acquisition treatments.

Furthermore, the influence of the data processing parameters on the two evaluation variables could be observed and this allows to set optimal parameters for the global processing of the dataset. Therefore, it was possible to conclude that it is preferable, in the context of this study and the associated data, to apply a combination of apodization with a Maxi=0.5 for the sinusoidal bell apodization and a β=3 for the Kaiser apodization. It was also chosen to apply a zero-filling.

This treatment allows us to obtain the most qualitative data possible by grouping all the data from the different participating laboratories.

A particular attention must however be paid to the fact that for some of the tested parameters few experiments were available, so biases of analysis are likely to be induced and present in this study.

To go further and now that the post-acquisition processing parameters have been determined, more methods of comparing the spectra to analyze their quality and their proximity to the theoretically expected spectrum could be implemented. This is a step that will be more efficiently implemented through collaboration with the different laboratories that will be able to identify the important points to observe during the comparison, including among others the accuracy of the position of the peaks as well as their intensity. There are many avenues of development because the field is very active around the methods of comparison of mass spectra with, in particular, the growing development of molecular networks for which this step of comparison is the starting point.

References

De Marco, Antonio, et Kurt Wüthrich. 1976. « Digital Filtering with a Sinusoidal Window Function: An Alternative Technique for Resolution Enhancement in FT NMR ». *Journal of Magnetic Resonance (1969)* 24 (2): 201-4. https://doi.org/10.1016/0022-2364(76)90028-7.

Karabacak, N. Murat, Michael L. Easterling, Nathalie Y. R. Agar, et Jeffrey N. Agar. 2010. « Transformative Effects of Higher Magnetic Field in Fourier Transform Ion Cyclotron Resonance Mass Spectrometry ». *Journal of the American Society for Mass Spectrometry* 21 (7): 1218-22. https://doi.org/10.1016/j.jasms.2010.03.033.

Kreitzberg, Patrick, Jake Pennington, Kyle Lucke, et Oliver Serang. 2020. « Fast Exact Computation of the k Most Abundant Isotope Peaks with Layer-Ordered Heaps ». *Analytical Chemistry* 92 (15): 10613-19. https://doi.org/10.1021/acs.analchem.0c01670.

Wimperis, S. 1994. « Broadband, Narrowband, and Passband Composite Pulses for Use in Advanced NMR Experiments ». *Journal of Magnetic Resonance, Series A* 109 (2): 221-31. https://doi.org/10.1006/jmra.1994.1159.