Benchmarking of different software for 2D NMR spectra automatic integration for metabolomic approaches

Julien GUIBERT, Marie TREMBLAY-FRANCO, Marine P.M. Letertre, Marine Piou, Jean-Nicolas Dumez, Patrick Giraudeau, Cecile Canlet

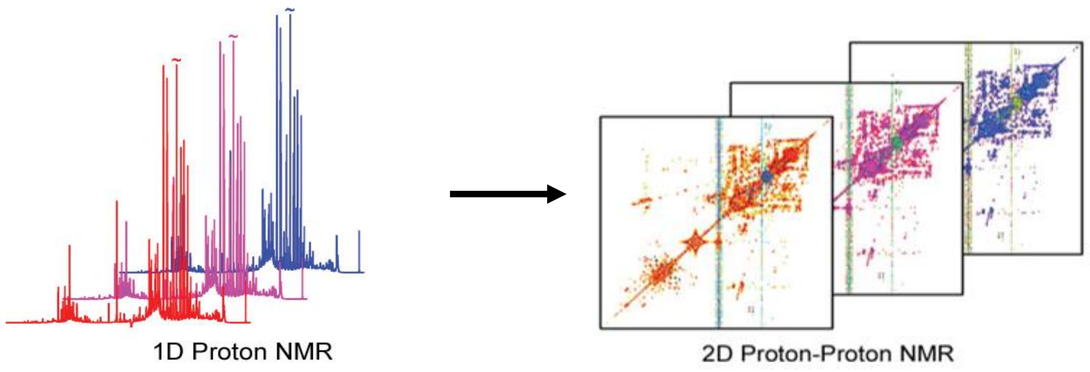
Benchmarking of different software for 2D NMR spectra automatic integration for metabolomic approaches

Julien GUIBERT[1], Marie TREMBLAY-FRANCO[1], Marine LETERTRE[2], Marine PIOU[2], Jean-Nicolas DUMEZ[2], Patrick GIRAUDEAU[2] and Cecile CANLET[1]

1 UMR1331 Toxalim, Université de Toulouse, INRAE, ENVT, INP-Purpan, UPS, Toulouse, France.
MetaToul-MetaboHUB, National infrastructure of metabolomics and fluxomics, Toulouse, France.
2 Nantes Université, CNRS, CEISAM UMR 6230, F-44000 Nantes, France.

Introduction

NMR-based metabolomic studies are mostly performed with proton 1D liquid-state NMR, using well-established protocols for (bio)fluids or extracts. Proton 1D NMR is rapid and robust but may be limited by an extensive signal overlap, which could impair the accurate identification and quantification of biomarkers.



Comparison of 1D and 2D NMR Spectroscopy for Metabolic Profiling retrieved from Que N. Van et al., 2008 [1]

To overcome this limitation, 2D NMR experiments have been shown useful to improve the resolution and metabolite identification [2]. Although optimized software already exist for 1D NMR spectra bucketing, tools for automatic integration of 2D NMR spectra show limitations (format, number of peaks detected, false positives, time consuming, etc).

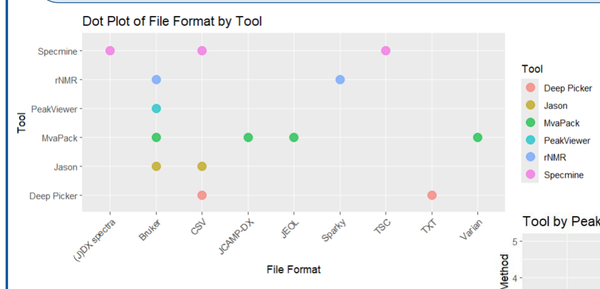
In this context, we aimed to evaluate and understand the available tools for processing 2D NMR spectra to identify the best methods for each step.

Material and methods

Different tools have been selected and tested with experimental 2D NMR COSY spectra from a mixture of 23 standards on several criteria (ease of use, format of NMR data, time, number of peaks detected, visualization, etc.):

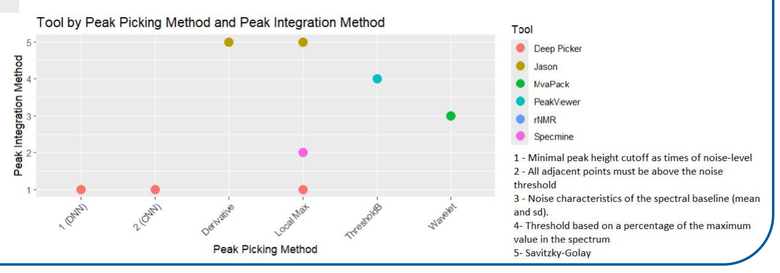
- **Deep picker** [3]: C/C++ online tool.
- **rNMR** [4]: R package.
- **Specmine** [5]: R package.
- **MVApack** [6]: GNU octave tool.
- **Jason** (commercial tool from Jeol)
- **PeakViewer**: Home-built Matlab program.

The most used peak picking method is based on **local maximum (Local Max)** detected around the spectra but less known methods such as **Deep learning based (Deep & Convolutional Neural Network)**, **Threshold based (ThresholdB)** or **Wavelet detection method (Wavelet)** were also tested. Those approaches were adapted to specifically treat NMR data.

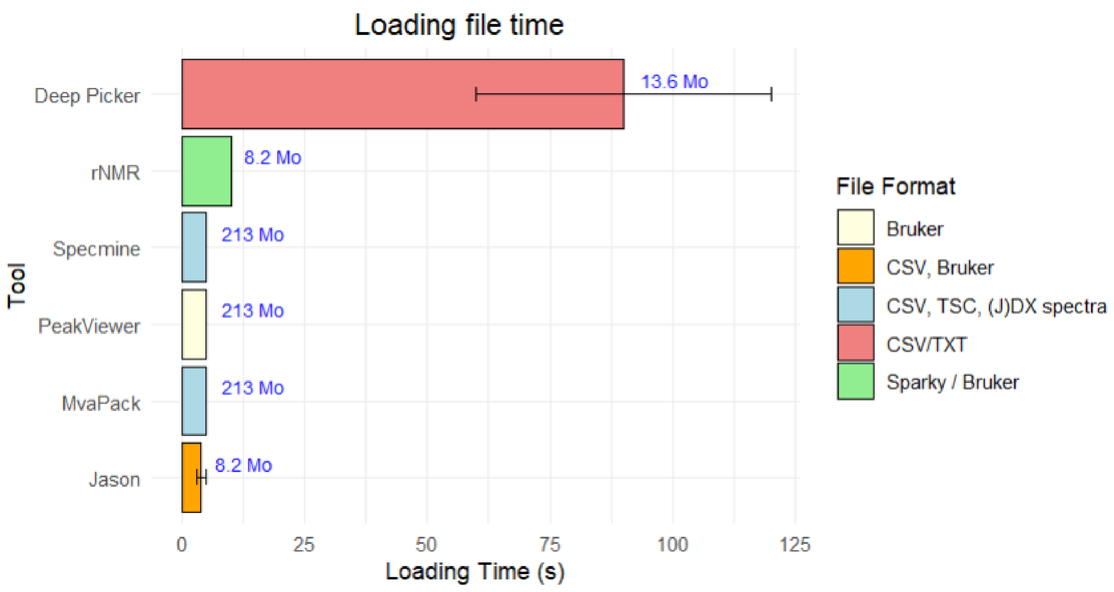


Those tools need different format of input file and are made up of different methods of peak picking and integration methods.

The results of those methods, shown on the right, will be compared using peak picking and peak integration done by an expert.



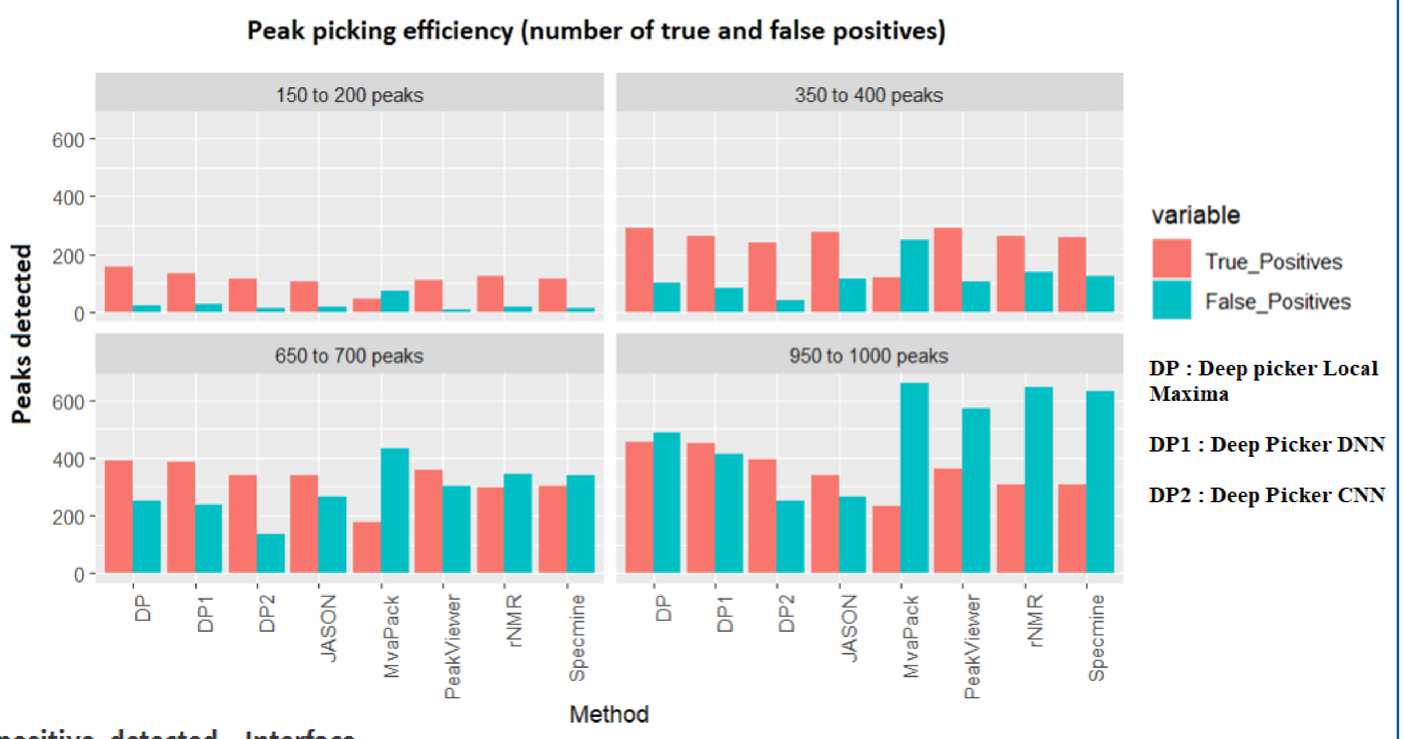
Results and discussion



The file used for the test was based on 2D COSY spectra from a single mixture and had a weight which varied depending on the input format chosen (8.2 Mo to 213 Mo).

Except for Deep picker with a loading time of 60 to 120 seconds, every other tool took around 3 to 10 second to load our input file.

However, this difference may not have been caused by their loading file function but rather because it is an online tool.



To understand the impact of the different peak picking and noise management method, the following were tested:

- 4 different noise thresholds (around 150, 400, 650 and 900 peaks detected)
- 7 different peak picking methods

While decreasing the noise threshold, we greatly increased the number of False positives peaks detected:

- For 640 peaks detected, 250 to 300 wrongfully detected peaks (almost 50%).
- Methods based on deep learning (DP1:DNN, DP2:CNN) have a better ratio of rightfully detected peaks.

Advantages and drawbacks of Peak Picking Algorithms

Tool	Language	Peak_picking_method	Loading_time	Run_time	True_positive_detected	Interface
Deep Picker	C/C++	Local Maxima	60-120s	++	+	-
Deep Picker	C/C++	DNN	60-120s	--	++	-
Deep picker	C/C++	CNN	60-120s	-	+++	-
Jason	Private Software	Derivative/Local Max	3-5s	++	+	+
PeakViewer	MatLab	Threshold Based	5s	+	+	+
rNMR	R	Local Maxima	10s	++	--	+
Specmine	R	Local Maxima	5s	++	-	-
MVAPack	GNU OCtave	Wavelet detection	5s	+	---	-

For approximately 150 to 640 peaks, run times were: Deep Picker: Local Maxima (0-30s), DNN (40-60s), CNN (20s), and around 2 to 10 second for every other tools/methods. Deep learning methods especially the CNN method, were the most accurate when we looked at True positives detected.

Local Maxima offers optimized run time but lower peak selection efficiency, **while DNN and CNN sacrifice speed for better accuracy.** Using different noise thresholds yielded similar results, except PeakViewer's Threshold-based function which excelled with a strong threshold: 93% true positives for 150 peaks, compared to 90% for CNN.

Conclusions

Throughout this process, we observed that input file format did not affect loading time. For peak picking method, the Local maxima approach was the simplest for small datasets while Deep learning and Threshold based method, especially CNN, provided more accurate results when compared to manual peak picking. Further analysis of each peak picking method with multiple spectra is needed to assess the methods usability.
We also need to determine the best way to compare peak integration methods, as different integration approaches may yield varying results.

References

[1] Que N. Van, et al., *Journal of Proteome Research* 2008 *7* (2), 630-639 DOI: 10.1021/pr700594s
[2] Marchand Jérémy, et al., « A Multidimensional 1H NMR Lipidomics Workflow to Address Chemical Food Safety Issues ».
[3] Li Da-Wei et al., « DEEP Picker Is a Deep Neural Network for Accurate Deconvolution of Complex Two-Dimensional NMR Spectra ».
[4] Lewis Ian et al.,« rNMR: Open Source Software for Identifying and Quantifying Metabolites in NMR Spectra ».
[5] Costa Christopher et al.,. « An R package for the integrated analysis of metabolomics and spectral data ».
[6] Worley, Bradley, et Robert Powers. « MVAPACK: A Complete Data Handling Package for NMR Metabolomics ».

Acknowledgments

- This research was funded by the French National Infrastructure for metabolomics and fluxomics MetaboHUB ANR-11-INBS-0010,
- This work has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement SUMMIT no. 814747).

Contact

- julienguibert@inrae.fr
- marie.tremblay-franco@inrae.fr
- cecile.canlet@inrae.fr